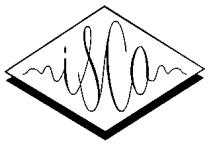


EUROSPEECH 2001
SCANDINAVIA

7'TH EUROPEAN CONFERENCE ON SPEECH COMMUNICATION AND TECHNOLOGY

SEPTEMBER 3 - 7, 2001
AALBORG CONGRESS AND CULTURE CENTRE
AALBORG - DENMARK

BOOK OF ABSTRACTS

EDITED BY
PAUL DALSGAARD, BØRGE LINDBERG, HENRIK BENNER, ZHENG-HUA TAN
CENTER FOR PERSONKOMMUNIKATION
AALBORG UNIVERSITY, DENMARK

ORGANISED BY

CENTER FOR PERSONKOMMUNIKATION, AALBORG UNIVERSITY, DENMARK

IN COLLABORATION WITH

DEPARTMENT OF TELECOMMUNICATIONS, NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY, NORWAY
LABORATORY OF ACOUSTICS AND AUDIO SIGNAL PROCESSING, HELSINKI UNIVERSITY OF TECHNOLOGY, FINLAND
CENTRE FOR SPEECH TECHNOLOGY, KTH, SWEDEN

ISBN 87-90834-10-0
ISSN 1018-4074
ISSN 0908-1224

Each abstract in this book is a direct reproduction of the text-file that was electronically forwarded to the Conference Committee at the time of submitting the corresponding paper.
To limit the size of the "Book of Abstracts" the formatting has been changed such that each abstract appears as one long text.
The authors are fully responsible for the content of their abstract.

The "Book of Abstracts" is printed in Denmark by Kommunik Grafiske Løsninger A/S, Aalborg.

Message from the ISCA President

On behalf of the International Speech Communication Association (ISCA), I would like to welcome you to the lovely town of Aalborg and to the 7th European Conference on Speech Communication and Technology - EUROSPEECH 2001-SCANDINAVIA.

As you are well aware, the EUROSPEECH series of conferences has now established itself as one of the major landmarks in the annals of speech communication, science and technology. From the first gathering in Paris in 1989, each subsequent conference - Genova'91, Berlin'93, Madrid'95, Rhodes'97 and Budapest'99 - has imprinted it's own style as well as adding new ideas to the overall format. This year is no exception, and we should thank our Scandinavian colleagues for embarking on a number of new ideas - such as full-paper submission - many of which I am sure will be carried forward to future ISCA conferences. As always, please let the Association know if you have any particular likes or dislikes, and we will do our best to evolve the conference to meet the needs of our membership.

Turning to ISCA, you may recall that the ESCA General Assembly in 1999 voted unanimously in favour of internationalising the Association - and the International Speech Communication Association was born. Elections were held and, by early 2000, the ISCA Board had expanded to include representatives from Japan, Australia and the USA. ISCA's new Board immediately established a positive and cooperative working relationship, and a number of new initiatives have been started that are aimed specifically at realising the aspirations of our international membership. For example, the Board is developing a strategy for giving increased support to local, regional and global activities, and relationships are being negotiated with other national and international associations with a view to providing mutual benefits for both members of ISCA and these other organisations. For information about these initiatives, and the wide range of other ISCA activities, please check out the ISCA website at http://www.isca-speech.org/.

More recently, you may have heard that the Permanent Council for the International Conferences on Spoken Language Processing (ICSLP) agreed unanimously at ICSLP'2000 in Beijing to enter into a formal relationship with ISCA that will allow future ICSLP conferences to receive the same financial and organisational support as EUROSPEECH. Bringing ICSLP under the umbrella of ISCA will greatly enhance the continuity and stability of ICSLP, as well as providing ISCA members with the benefits that derive from an annual conference. Whilst this means that both EUROSPEECH and ICSLP can be viewed as 'INTERSPEECH' events, these arrangements do not imply any convergence in terms of scientific or disciplinary balance. On the contrary, it might now be possible to enhance the character of the different conferences in ways which we have yet to imagine. Again, we welcome new ideas that could be implemented in future conferences.

Finally, returning to EUROSPEECH 2001 - SCANDINAVIA, on your behalf I would like to offer a formal vote of thanks to the Scandinavian team responsible for seeing this project through to completion. As you can imagine, planning and organising a major international event such as this is no small undertaking. Many people have dedicated huge quantities of their personal time and effort towards creating the event we are now enjoying. Whilst it is not the Scandinavian way to single out individuals, we nevertheless should offer a huge thank you to Paul Dalsgaard and his wonderful team for an excellent job well done. We also need to offer thanks to all those members of the international community who provided input to the reviewing process, to the organisers of the special sessions, and to the people who have mounted satellite workshops in support of the main conference. Last, but by no means least, I would like to thank all of you for supporting and attending this conference - thanks. I look forward to seeing you again at the next ISCA event.

I wish you all a successful conference, an enjoyable time in Aalborg and a memorable visit to this lovely part of Scandinavia.

De bedste ønsker

Prof. Dr. Roger K. Moore
ISCA President

Message from the Conference Chair

On behalf of the Conference Committee it is a great pleasure for me to welcome you to the European Conference on Speech Communication and Technology - Eurospeech 2001 - Scandinavia.

Eurospeech 2001 - Scandinavia is the seventh biennial conference of ISCA - the International Speech Communication Association. The previous conferences were held in Paris, Genoa, Berlin, Madrid, Rhodes and Budapest. Following the move from ESCA to ISCA and the co-ordination with ICSLP we are happy to be able to declare Eurospeech "An InterSpeech Event" for the first time, in the spirit of co-operation between all scientists around the world who are actively involved in speech and spoken language research and development. The conference is organized through the combined efforts of a group of speech and language scientists from four Nordic countries: Denmark, Finland, Norway and Sweden.

We are very pleased to see the great interest in this year's Eurospeech. Most of you did probably not know of Aalborg a year ago - or know of it only as the name of an aquavit. Aalborg is by far the smallest town with a speech research group that has offered to organize Eurospeech. We hope that you will enjoy the cozy atmosphere of Aalborg, where there will be ample opportunities to meet colleagues by chance outside the conference.
Moving north, combined with the move from abstract to full paper submission could have been a risky move. We are pleased to see that we received as many as 902 contributions, which have been thoroughly reviewed by at least two reviewers each. With a reasonable work load on our reviewers, a maximum of 10 papers each, it is easy arithmetic that we needed to engage a huge number of reviewers, actually about 250 persons, who's devoted effort we thankfully acknowledge. The review committee's impression is that the full paper scheme substantially improved the quality and we are confident that the present programme with 672 papers (3 Keynotes, 47 ESE and 622 regular papers) in parallel oral and poster presentations will reflect that.

For the present Eurospeech conference, ISCA and the Conference Committee have introduced some innovations to the technical programme. Beside the more traditional sessions with oral and poster presentations, Eurospeech 2001 - Scandinavia is offering a number of Eurospeech Special Events (ESE's). Each ESE is focussed on a specially selected theme within speech or spoken language science or technology. The Eurospeech Special Events will run in parallel with the regular Conference sessions.
In addition we offer a mini-poster Open Forum in which all registered participants are invited to exhibit a 'mini-poster' on any topic relating to the conference.

Each oral presentation is given 20 minutes in total for its presentation and discussion. We urge the presenters to use the extra five minutes, compared to last Eurospeech, for discussion of the presentation.
The generous support of our sponsors has made it possible for us to include services normally not included in the basic conference registration fee. Buffet lunches, Welcome Reception and the conference 'gala' with entertainment and food on Thursday are included for all participants. We hope that this will provide many excellent occasions for participants to meet old and new colleagues and to make new friends. In the same vein we have experimentally provided the possibility to add a link to a photo to the registration. At the time this goes to the printer we still don't know if enough participants used this option to justify a publication of a Eurospeech photo booklet that could make it easier for you to connect names and faces.

While in Scandinavia we hope that you will venture going further north and explore and enjoy the varied nature and other attractions of the Nordic countries. You won't experience the midnight sun this time, but maybe some northern lights. We are looking forward to hosting you in Aalborg, and I hope that we all together will make Eurospeech 2001 - Scandinavia a worthwhile and memorable event.

Paul Dalsgaard
Chair of Eurospeech 2001 - Scandinavia

Organising Committees of *Eurospeech 2001 - Scandinavia*

Chair
Paul Dalsgaard (CPK)
Center for PersonKommunikation (CPK)
Aalborg University
Denmark

Co-Chair
Björn Granström
Centre for Speech Technology (CTT)
KTH
Sweden

Conference Committee
Gösta Bruce
Department of Linguistics and Phonetics
Lund University
Sweden

Rolf Carlson
Centre for Speech Technology (CTT)
KTH
Sweden

Wim van Dommelen
Department of Linguistics
Norwegian University of Science and Technology
Norway

Matti Karjalainen
Laboratory of Acoustics & Audio Signal Processing
Helsinki University of Technology
Finland

Unto K Laine
Laboratory of Acoustics & Audio Signal Processing
Helsinki University of Technology
Finland

Børge Lindberg
Center for PersonKommunikation
Aalborg University
Denmark

Torbjörn Svendsen
Department of Telecommunications
Norwegian University of Science and Technology
Norway

Local Organising Committee
Henrik Benner
CPK, Aalborg University

Paul Dalsgaard

Børge Lindberg

Web-Master
Henrik Benner
CPK, Aalborg University

Scientific Committee of *Eurospeech 2001 - Scandinavia*

A

Acero, Alex, Microsoft Research, USA
Adda, Gilles, LIMSI, France
Ahrenberg, Lars, IDA, Linköping, Sweden
Ainsworth, William, Mackay Inst. of Comm.& Neuroscience, UK
Akagi, Masato, Japan Advanced Institute of Sc. and Tech., Japan
Alku, Paavo, HUT, Finland
Alter, Kai, Max-Planck-Institute of Cognitive Neuroscience, Germany
Andre, Elisabeth, DFKI, Germany
Andre-Obrecht, Regine, IRIT-UPS, France
Araki, Masahiro, Kyoto Institute of Technology, Japan
Aubert, Xavier, Philips Research Laboratories Aachen, Germany

B

Badin, Pierre, ICP, Grenoble, France
Bailly, Gerard, INPG Grenoble, France
Bard, Ellen Gurman, University of Edinburgh, United Kingdom
Barry, Bill, Universität des Saarlandes, Germany
Bechet, Frederic, LIA Universite D'Avognon, France
Beckman , Mary, Ohio State Univ., Columbus, USA
Beddor, Patrice Speeter, Univ. of Michigan, USA
Bell-Berti, Fredericka, St. John's Univ., Jamaica
Bernsen, Niels Ole, University of Southern Denmark, Denmark
Berthommier, Frédéric, ICP, France
Bimbot, Frederic, IRISA, Rennes, France
Black, Alan, CMU, USA
Blauert, Jens, Bochum University, Germany
Blomberg, Mats, KTH, Stockholm, Sweden
Bloothooft, Gerrit, University of Utrecht, The Netherlands
Boda, Peter, Nokia Research Center, Finland
Boe, Louis-Jean, Univ. Stendhal, Grenoble, France
Bonastre, Jean-Francois, LIA-Universite d'Avignon, France
Bonneau, Anne, Loria, France
Bourlard, Hervé, IDIAP, Switzerland
Boves, Louis, Katholieke Univ. Nijmegen, The Netherlands
Breen, Andrew, UEA, United Kingdom
Brennan, Susan, State Univ. NY, USA
Bridle, John, Phonetic Systems UK Ltd, United Kingdom
Broad, David, Santa Barbara, USA
Broersen, Piet M T, Delft, The Netherlands
Bruce, Gøsta, Dept. of Linguistics and Phonetics, Sweden
Brøndsted, Tom, CPK, Aalborg University, Denmark
Burnett, Ian, University of Wollongong, Australia

C

Cahn, Janet, Motorola Inc., Human Interface Laboratory, USA
Campbell, Nick, ATR, Japan
Carey, Michael, Ensigma Technologies, United Kingdom
Carlson, Rolf, KTH, Stockholm, Sweden
Carre, Rene, ENST, Paris, France
Chien, Jen-Tzung, Dept. of Comp. Sci. and Inf. Eng., Tainan, Taiwan
Chollet, Gerard, ENST, Paris, France
Chu-Carroll, Jennifer, IBM T.J. Watson Research Center, USA
Clements, Mark A., Georgia Tech, USA
Cutler, Anne, MPI for Psycholing., Nijmegen, The Netherlands

D

D'Alessandro, Christophe, LIMSI, France
Dalsgaard, Paul, CPK, Aalborg University, Denmark
Damper, Robert I., University of Southampton, United Kingdom
De, Aloknath, Hughes (HSS), India
de Cheveigne, Alain, IRCAM, France
De Mori, Renato, Univ. of Avignon, France
Delcloque, Philippe, University of Abertay Dundee, Scotland, UK
den Os, Els, KPN Leidschendam, The Netherlands
Diehl, Randy, Uni. Of Texas, USA
Dobler, Stefan, Ericsson Research, Germany
Draxler, Christoph, LMU, München, Germany
Drygajlo, Andrzej, EPFL, Switzerland
Dutoit, Thierry, Faculte Polytechnique de Mons, Belgium

E

Elenius, Kjell, KTH, Sweden
Engstrand, Olle, Dept. of Linguistics, Stockholm University, Sweden
Eriksson, Anders, Stockholm University, Sweden
Eskenazi, Maxine, CMU, USA
Euler, Stephan, Bosch, Germany

F

Fakotakis, Nikos, Univ. of Patras, Greece
Falcone, Mauro, FUB, Rome, Italy
Farnetani, Edda, CNR Padova, Italy
Faundez, Marcos, Escola Universitaria Politecnica de Mataro, Spain
Fellbaum, Klaus, Brandenburg University of Technology, Germany
Fujisaki, Hiroya, Sience University of Tokyo, Japan
Furui, Sadaoki, Tokyo Inst. of Technology, Japan

G

Gauvain, Jean-Luc, LIMSI, France
Glass, Jim, MIT, Boston, USA
Gong, Yifan, Texas Instruments, USA
Gordos, Geza, Budapest Univ. of Techn. and Economics, Hungary
Granström, Björn, CTT/TMH/KTH, Stockholm, Sweden
Green, Phil, University of Sheffield, United Kingdom

H

Haavisto, Petri, Nokia, Finland
Hansen, John H.L., Univ. of Colorado at Boulder, USA
Harborg, Erik, SINTEF Telecom and Informatics, Norway
Hawkins, Sarah, University of Cambridge, United Kingdom
Hazan, Valerie, University College London, United Kingdom
Hedelin, Per, Chalmers, Sweden
Heeman, Peter, OGI, Oregon, USA
Hermansky, Hynek, Oregon Grad. Inst. of Science and Tech., USA
Hernando, Javier, Universitat Politecnica de Catalunya, Spain
Hess, Wolfgang, IKP Universität Bonn, Germany
Heute, Ulrich, University of Kiel, Germany
Hindle, Donald, AnswerLogic, UK
Hirose, Keikichi, Uni. of Tokyo, Japan
Hirsch, Hans-Guenter, University of Applied Sci.Niederrhein, Germany
Hirschberg, Julia, AT&T, Florham Park, USA
Hirschman, Lynette, MITRE, USA
Hirst, Daniel, CNRS, Universite de Provence, France
Holmes, Wendy, 20/20 Speech, United Kingdom
Holter, Trym, Motorola Australian Research Centre, Australia
House, David, KTH, Sweden
Hunt, Melvyn, Phonetic Systems UK Ltd, United Kingdom
Härmä, Aki, Agere Systems, Media Signal Processing Research, USA
Höge, Harald, Siemens, Germany

I

Iivonen, Antti, Univ. of Helsinki, Finland
Ingram, John, University of Queensland, Australia

J

Jensen, Søren Holdt, Aalborg University, Denmark
Johansen, Finn Tore, Telenor R&D, Norway
Johnsen, Magne H., NTNU, Norway
Johnson, Keith, Ohio State Univ., Columbus, USA
Junqua, Jean-Claude, Panasonic, USA
Jönsson, Arne, IDA, Linköping, Sweden

K

Kacic, Zdravko, University of Maribor, Slovenia
Kakehi, Kazuhiko, Nagoya University, Japan
Kamm, Candace, AT&T, USA
Karjalainen, Matti, Helsinki University of Technology, Finland
Karlsson, Fred, Dept. of General Ling., University of Helsinki, Finland
Kato, Hiroaki, ATR, Japan
Keller, Eric, Univ. de Lausanne, Switzerland
Kellner, Andreas, Philips, Germany
Kitawaki, Nobuhiko, University of Tsukuba, Japan
Kleijn, Bastiaan, KTH, Stockholm, Sweden
Kokkinakis, George, Univ. of Patras, Greece
Kondoz, Ahmet, Univ. of Surrey, United Kingdom
Krauwer, Steven, Utrecht University/ELSNET, The Netherlands
Kroon, Peter, Agere Systems
Kubin, Gernot, Graz University of Technology, Austria
Kuhn, Roland, Panasonic, USA
Kurimo, Mikko, Helsinki University of Technology, Finland

L

Lacerda, Francisco, Stockholm University, Sweden
Lahiri, Aditi, Univ. of Konstanz, Germany
Laine, Unto K., Helsinki University of Technology, Finland
Lamel, Lori, LIMSI, France
Laurila, Kari, Nokia Research, Finland
Lee, Chin Hui, Bell Labs, Lucent Tech., USA
Lee, Lin-Shan, Taiwan University, Taiwan
Lewin, Ian, Netdecisions, USA
Linares, Georges, LIA, France
Lindberg, Børge, Aalborg University, Denmark
Lindblom, Björn, Stockholm University, Sweden
Litman, Diane, AT&T Labs - Research, USA
Ljolje, Andrej, AT&T Labs, USA
Lofqvist, Anders, Lund University Hospital, Sweden

M

Maddieson, Ian, University of California, Berkeley, USA
Makhoul, John, BBN Technologies, USA
Mangold, Helmut, DaimlerChrysler, Germany
Martens, Jean-Pierre, Univ. Gent, Belgium
Mason, John, The University of Wales Swansea, UK
Massaro, Dominic, Univ. California, USA
McCree, Alan, TI, USA
McDonough, John, Karlsruhe University, Germany
McTear, Mike, Belfast, Ireland
Millar, Bruce, Australian National University, Australia
Moore, Johanna, HCRC, Edinburgh, United Kingdom
Moore, Roger, 20/20 Speech, United Kingdom
Moreno, Pedro J., Cambridge Res. Lab, Compaq Comp. Corp., UK
Morgan, Nelson, ICSI, USA
Moriya, Takehiro, NTT, Japan
Möbius, Bernd, Univ. Stuttgart, Germany

N

Nakamura, Satoshi, ATR Spoken Lang.Trans. Res. Lab., Japan
Ney, Hermann, Aachen University of Technology, Germany
Noeth, Elmar, Universitât Erlangen-Nürnberg, Germany
Nouza, Jan, TU Liberec, Czech Republic
Németh, Géza, Dept. of Telecom. & Telematics, Hungary

O

O'Shaughnessy, Douglas, INRS-Télécommunications, Canada
Ohala, John, Univ. of California, Berkeley, USA
Olaszy, Gabor, Kempelen Farkas Speech Research Lab., Budapest
Olsen, Jesper, Vox Generation Ltd., United Kingdom
Omologo, Maurizio, ITC-IRST, Italy
Ortega-Garcia, Javier, Universidad Politécnica de Madrid, Spain
Ostendorf, Mari, University of Washington, USA
Oviatt, Sharon, OGI, Oregon, USA

P

Paliwal, Kuldip, Griffith University, Australia
Pallett, David, NIST, USA
Parthasarathy, Partha, AT&T Labs, USA
Pearce, David, Motorola Labs., USA
Perkell, Joseph, MIT, Cambridge, Massachusetts, USA
Perrier, Pascal, ICP - INPG & Université Stendhal, France
Picheny, Michael, IBM, USA
Pieraccini, Roberto, SpeechWorks International, USA
Pols, Louis C.W., Univ. of Amsterdam, The Netherlands

R

Rahim, Mazin, AT&T Labs, USA
Rajasekaran, Raja, Texas Instruments Incorporated, USA
Recasens, Daniel, Univ. Autònoma de Barcelona, Spain
Reithinger, Norbert, DFKI GmbH, Saarbruecken, Germany
Renals, Steve, Sheffield Univ, United Kingdom
Reynolds, Doug, MIT, USA
Rigoll, Gerhard, Univ. Duisburg, Germany
Robinson, Tony, SoftSound Limited, United Kingdom
Rose, Richard, AT&T Labs - Research, USA
Rosenberg, Aaron, AT&T Labs-Research, USA
Rudnicky, Alexander, CMU, USA
Ruske, Guenther, TU Munich, Germany
Russell, Martin, Univ. Of Birmingham, United Kingdom

S

Sagayama, Shigeki, University of Tokyo, Japan
Salmela, Petri, Tampere University of Technology, Finland
Sams, Mikko, Helsinki Univ. of Technology, Finland
Schiel, Florian, LMU, München, Germany
Schwartz, Jean-Luc, ICP, France
Serniclaes, Willy, Lab. de Statistique Médicale., Belgium
Shockey, Linda, Univ. of Reading, United Kingdom
Shriberg, Elizabeth, SRI, Menlo Park, USA
Sidner, Candy, Mitsubishi Electric Research Labs, Japan
Siohan, Olivier, Bell Labs, USA
Smaili, Kamel, LORIA, France
Sondhi, M. Mohan, Bell Labs, Lucent Technologies, USA
Soong, Frank, Bell Labs, USA
Sproat, Richard, AT&T Labs - Research, USA
Steeneken, Herman, TNO, The Netherlands
Stern, Richard, CMU, USA
Strange, Winifred, City Univ. of New York Grad. School, USA
Strik, Helmer, KU Nijmegen, The Netherlands
Svendsen, Torbjørn, NTNU, Norway
Swerts, Marc, Eindhoven Univ. of Technology, The Netherlands
Syrdal, Ann, AT&T Research Labs, USA

T

Talkin, David, Rhetorical Systems, Ltd., USA
ten Bosch, Louis, Lernout & Hauspie, Belgium
Terken, Jacques, Eindhoven Univ. of Technology, The Netherlands
Trancoso, Isabel, INESC, Lisboa, Portugal
Traum, David, USC Institute for Creative Technology, USA
Traunmüller, Hartmut, Univ. of Stockholm, Sweden

U

Uhlir, Jan, CVUT, Czech Republic

V

Van Coile, Bert, Lernout & Hauspie, Belgium
Van Compernolle, Dirk, KU Leuven, Belgium
van den Heuvel, Henk, University of Nijmegen, The Netherlands
van Dommelen, Wim, Department of Linguistics, NTNU, Norway
van Heuven, Vincent, Universiteit Leiden, The Netherlands
van Santen, Jan, OGI, USA
Vary, Peter, Aachen, Germany

Vicsi, Klara, Budapest Univ. of Technology and Economics, Hungary
Vidal, Enrique , Univ. Politecnica de Valencia, Spain
Vieira de Sá, Luis, Univ. Coimbra, Portugal
Viikki, Olli, Nokia Research, Finland
Vonwiller, Julie, Appen Pty Limited, Australia

W

Wagner, Michael, University of Canberra, Australia
Walker, Marilyn, AT&T Labs Research, USA
Wellekens, Christian, Eurecom, France
Whiteside, Sandra, University of Sheffield, United Kingdom
Wichmann, Anne, Univ. of Central Lancashire, UK
Woodland, Phil, Cambridge Univ., UK

Y

Yoo, Chang D., KAIST, Korea

Z

Zhao, Yunxin, University of Missouri, USA

Eurospeech 2001 - Scandinavia Special Events (ESE)

Eurospeech 2001 - Scandinavia offers a number of Eurospeech Special Events (ESE's). Each ESE is focussing on a specially selected theme within speech or spoken language science or technology. Some of the ESE's include poster presentations and a panel discussion. The posters will be on display in the ESE-session room (Europa Hall).

The Eurospeech Special Events run in parallel to the regular Conference sessions. The ESE's are:

1 ***What do Industry and Universities Expect from Each Other ? Panel Discussion***
Organised by Matti Karjalainen and Unto K Laine, Laboratory of Acoustics & Audio Signal Processing, Helsinki University of Technology, Finland

2 ***Noise Robust Recognition***
Organised by David Pearce, Motorola Labs.,USA, Børge Lindberg, CPK, Denmark, Hans-Guenter Hirsch, University of Applied Sciences Niederrhein, Germany, Raja Rajasekaran, Texas Instruments Incorporated, USA, Hynek Hermansky, Oregon Graduate Institute of Science and Technology, USA and Petri Haavisto, Nokia Research, Finland

3 ***IMAGINATION2001 - A contest for young researchers***
Organised by Gerrit Bloothoft, Utrecht University, The Netherlands and Els den Os, KPN Research, Leidschendam, The Netherlands
The winner of the contest will win 5000 Euro contributed by Hewlett-Packard European Research Labs. At Eurospeech 2001 - Scandinavia a jury will judge the ideas of feasibility, creativeness and originality. The jury consists of Sadaoki Furui, Julia Hirschberg, Joseph Mariani, Hans Kamperman and Roger Tucker. The winner will be announced during the closing ceremony of the conference.

4 ***SIGshow***
Organised by Gerrit Bloothoft, Utrecht University, The Netherlands

5 ***Existing and Future Corpora - Acoustic, Linguistic and Multi-modal Requirements***
Organised by Christoph Draxler and Florian Schiel, University of Munich, Germany

6 ***Education Arena***
Organised by Mark Tatham, University of Essex, United Kingdom, Anders Eriksson, Stockholm University, Sweden and Martin Cooke, University of Sheffield, United Kingdom

7 ***Integration of Phonetic Knowledge in Speech Technology***
Organised by William Barry, Institut für Phonetik und Phonologie, Saarbrücken, Germany and Wim van Dommelen, Department of Linguistics, Norwegian University of Science and Technology, University of Technology, Norway

The Conference Committee highly appreciate the efforts and work done by the organising researchers and is grateful to the ESE-contributors.

Grants and Paper Awards at Eurospeech 2001 - Scandinavia

GRANTS

The ISCA grant fund aims at supporting the Speech Community to participate in conferences on speech. All grants for participation in Eurospeech 2001 - Scandinavia is administered by the ISCA Board with the focus on assuring a broad representation of the scientific topics of the conference and that the finally chosen candidates broadly represent the speech community internationally. Priority in the selection process were given to applicants who have submitted a paper, whether it is accepted or not.
The Conference Committee has given 20 grants to students and young researchers.

ISCA AWARDS

The ISCA Board offers three awards at Eurospeech 2001 - Scandinavia

a) one scientific paper 'best paper award'
b) one student paper 'best student paper award'
c) one 'best Open Forum award' for the most inventive mini-poster.

ELRA AWARD

The European Language Resource Association (ELRA) offers one award for best (student) paper addressing Language Resources. This award is a best student paper that addresses Language Resources and report on work/project conducted by students.
The author of the selected paper will have his/her travel costs reimbursed (up to 500 Euro's).

"COCOSDA" BEST PAPER AWARD

The International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques for Speech (COCOSDA) offers a Best Paper on Speech Technology Evaluation at Eurospeech 2001 - Scandinavia.
The award includes a certificate plus the reimbursement of travel cost up to 500 Euro's.

All grants and awards are handled by a committee established by the ISCA Board. The name and award of each award winner will be presented at the closing ceremony.

Sponsors of Eurospeech 2001 - Scandinavia

The Conference Committee acknowledges with gratitude the generous sponsoring and support of the *Eurospeech 2001 - Scandinavia* conference from the following companies and institutions:

Aalborg City Council, Aalborg, Denmark

Aalborg University, Aalborg, Denmark

Det Obelske Familiefond, Aalborg, Denmark

ELRA/ELDA, France

ELSNET

Ericsson AB, Sweden

Finnish Tourist Office, Helsinki, Finland

Nokia, Helsinki, Finland

Nordisk Språkteknologi, Voss, Norway

Sail Port Northern Europe, Voss, Norway

Siemens Mobile Phones, Aalborg, Denmark

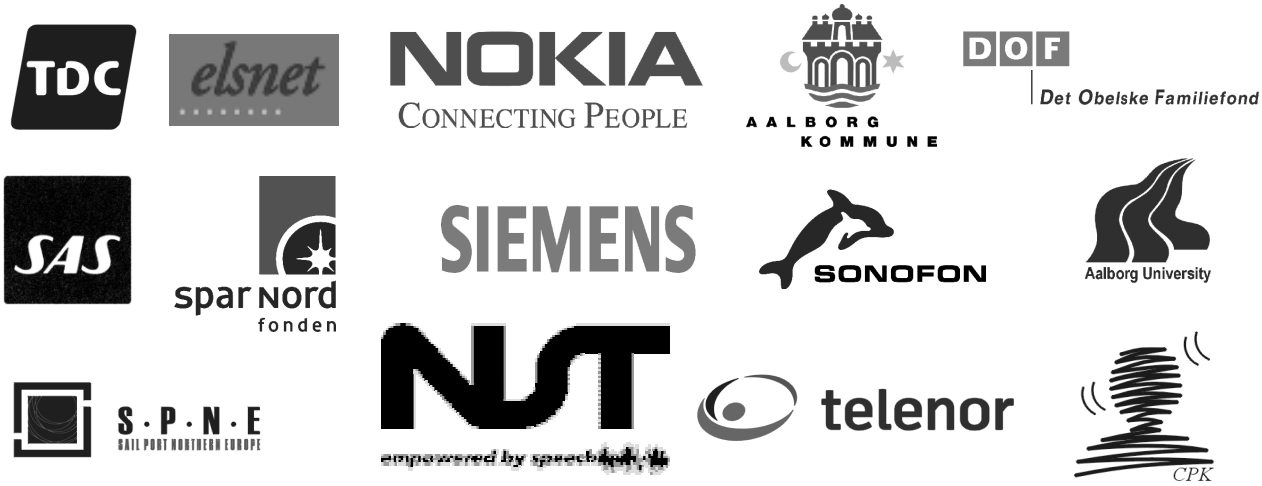
Sonofon A/S, Aalborg, Denmark

Spar Nord Fonden A/S, Aalborg, Denmark

Tele Denmark Communications, Copenhagen, Denmark

Telenor, Oslo, Norway

Scandinavian Airline Systems, Copenhagen, Denmark



S · P · N · E
SAIL PORT NORTHERN EUROPE

S.AI.L Port Northern Europe
SPNE

SPNE promotes and develops commercial initiatives in information and communication technology, with a primary focus on Speech, Artificial Intelligence, and Language technology, enhancing business creation, technological breakthrough and employment.
SPNE delivers business development services to start-ups companies, growing and other high-tech initiatives, building on networks of expertise, technology and finance providers. SPNE is a facilitator and provider of an entrepreneurial environment for innovation, technology transfer, application development and commercialisation. S.AI.L is speech artificial intelligence and language technology, - the future of ICT.

SPNE's deliver the 5 core services of:

- Infrastructure,
- Technology,
- Finance,
- Business consultancy
- Global networking,

Services tailored to the needs and allowing for the develop of

- Start-up companies
- Industry Spin-offs
- Projects and Partnerships
- University Spin-offs
- Growing company services

.

For more information, check out www.sail.no or give us a ring and ask us how to give your company or concept a voice in the market

Business Development for Human Communication Technology

NST (Nordisk Språkteknologi) is the leader in the language technology market in the Nordic countries. Language technology enables machines to talk to people and people to talk to machines without any technical barriers. It facilitates dialogue between humans and computers. This technology is already widely used and there is explosive growth in the field internationally.

NST was established in Voss in 1997 by Arne Gilbakken and Rune Relling, who are both still active in the business. The company now has 82 employees. The bulk of operations, such as the development of core technologies and systems and product development are carried out in Voss, whereas sales activities and marketing management are based in Copenhagen, where a sales office was opened in June 2001.

NST's main activity is to develop, market and sell language technology solutions, services and products. The company's primary focus is on speech synthesis and voice recognition. Speech synthesis means that a computer program reads text with a near-human voice, and voice recognition makes it possible for a computer to understand natural speech. NST is concentrating on products in the areas of "telecommunications and the mobile Internet", "medical dictation" and "assistive technologies" for the Scandinavian market.

In addition, NST develops core language technology software. Core technology products under development include SDKs (Software Development Kits) and licensing programs for end products and solutions.

NST's own, commercial products include NST VoiceCard, a voice server that converts text messages sent from the Internet or mobile phones into talking messages. Using NST VoiceCard, NST developed the mobile application, Vera Vox, which reads SMS text messages out to the recipient. This service has been a great success and will shortly be launched on the international market.

NST has also developed NST SmallTalk, a computer-aid program for the visually impaired that reads text from the screen. The company is also developing a series of writing and reading tools for people with special needs.

NST has entered an international strategic alliance with the software, computer equipment and consultancy giant, IBM. This partnership means that NST is IBM's first appointed developer and distributor of language technology products in the Scandinavian market. It also gives NST access to IBM's global laboratories, research and expertise. IBM has in turn license its dictation technology to NST. The company will further develop, market and sell these products under their own trademark.

Exhibitors at Eurospeech 2001 - Scandinavia (Known at time of printing)

Academic Press
http://www.academicpress.com/csl
Contact: Joan Dargan

Carstens Medizinelektronik
http://www.articulograph.de
Contact: Brigitta Carstens

ELSNET
http://www.elsnet.org
Contact: Brigitte Burger

Human Quality (HuQ) Speech Technologies
http://www.huq.nl
Contact: Tjeerd Andringa

ICSLP-2002 - Interspeech-2002
http://www.icslp.org
Contact: John H. L. Hansen

InSTIL - ISCA SIG
http://dbs.tay.ac.uk/instil
Contact: Philippe Delcloque

ISCA
http://www.isca-speech.org
Contact: Emmanuelle Petchot-Gardia

Kluwer Academic Publishers
http://www.wkap.com
Contact: Vincent van Leest

NST - Nordisk Språkteknologi as
http://www.nst.as
Contact: Margunn Instefjord

rhetorical
http://www.rhetorical.com
Contact: Mark Durrant

S.AI.L Port Northern Europe
http://www.sail.no
Contact: Lene Rutle

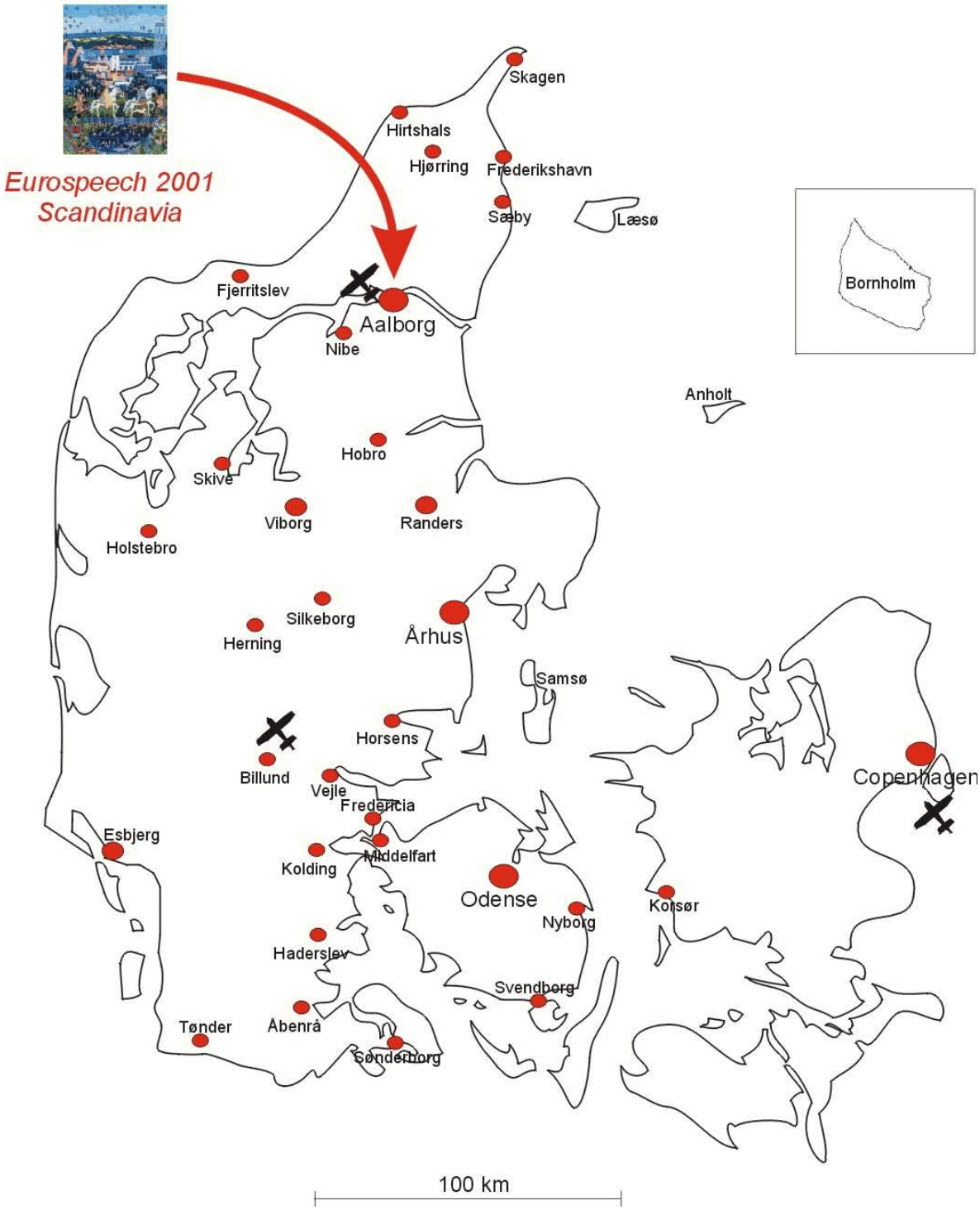
SmartKom
http://www.smartkom.org/
Contact: Anselm Blocher

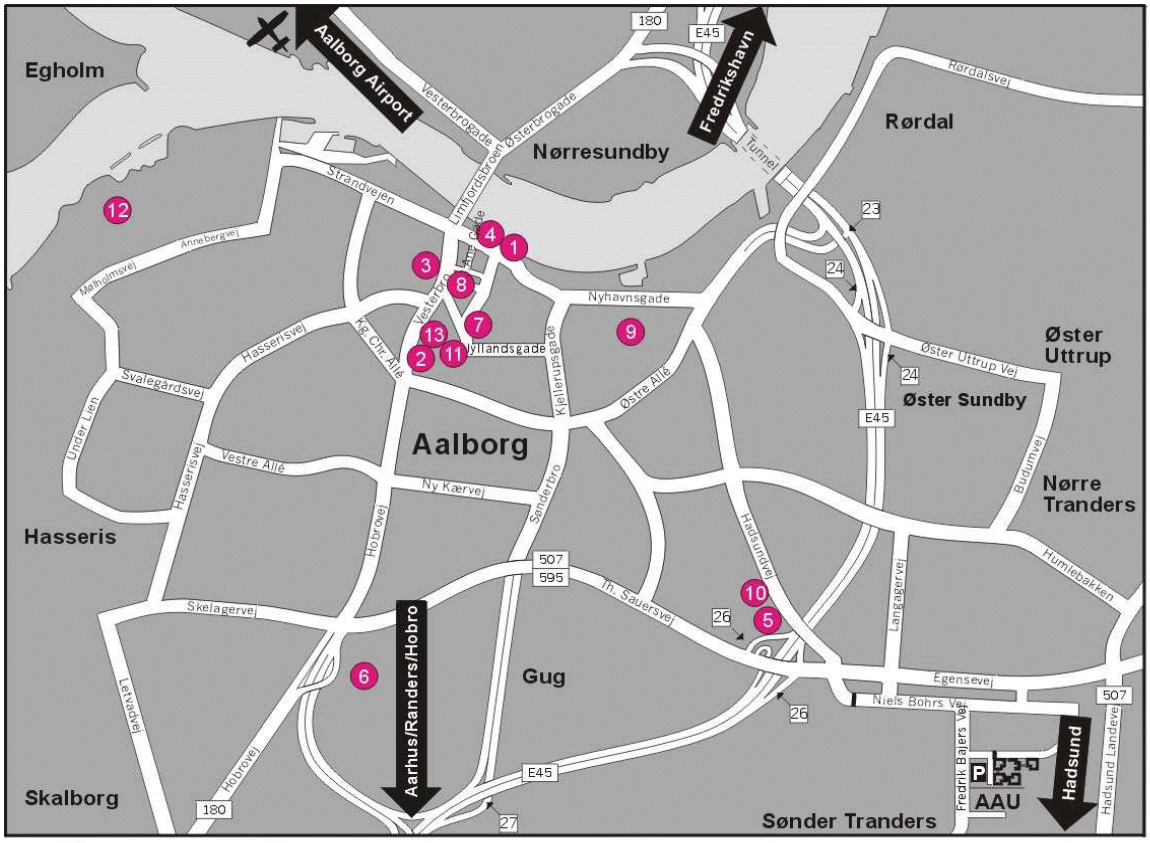
Speech-Ware
http://www.speech-ware.com
Contact: Per Bang

SYMPALOG AG
http://www.sympalog.de
Contact: Manuela Boros

Aalborg Centerboghandel
http://www.centerboghandel.auc.dk
Contact: Keld Juhl Jensen

Denmark

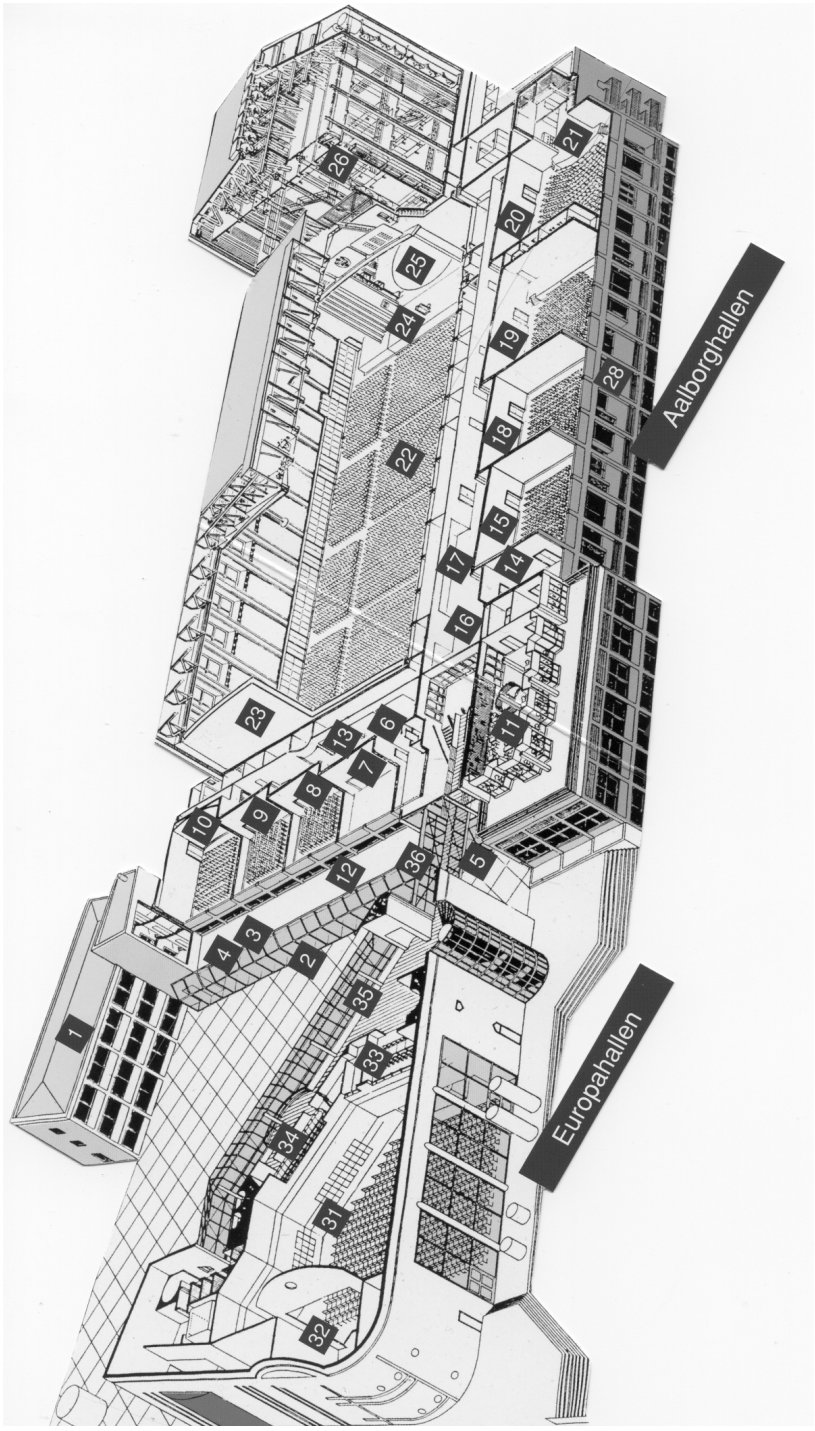

Eurospeech 2001
Scandinavia
Skagen
Hirtshals
Hjørring
Frederikshavn
Sæby
Læsø
Bornholm
Fjerritslev
Aalborg
Nibe
Anholt
Hobro
Skive
Viborg
Randers
Holstebro
Silkeborg
Herning
Århus
Samsø
Horsens
Billund
Vejle
Fredericia
Copenhagen
Esbjerg
Middelfart
Kolding
Odense
Nyborg
Korsør
Haderslev
Svendborg
Tønder
Åbenrå
Sønderborg
100 km



1. First Slotshotellet
2. Hotel Hvide Hus
3. Helnan Phønix Hotel
4. Radisson SAS Limfjord Hotel
5. Scandic Hotel Aalborg
6. Quality Hotel Scheelsminde
7. Park Hotel
8. Hotel Chagall
9. Aalborg Sømandshjem
10. Kommunedata UKC
11. Railway Station (Banegården)
12. Danhostel Aalborg Vandrerhjem
13. Aalborg Congress & Culture Centre

Eurospeech 2001 - Scandinavia

Aalborg Congress and Culture Centre - Room allocation



Aalborghallen

GROUND FLOOR

2: MAIN ENTRANCE
3: INFORMATION DESK & EUROSPEECH 2001 SECRETARIAT
22: POSTER AREAS (WESTERN PART)
22: OPENING CEREMONY & THEATRE (EASTERN PART)
28: FOYER & LUNCH AREA

FIRST FLOOR

10: INTERNET SERVICE ROOM
12: EXHIBITION & CREDIT CARD MACHINE
18: MUSIKSALEN (Music Hall)
19: RADIOSALEN (Radio Hall)
20: THE LITTLE THEATRE

Europahallen

GROUND FLOOR

31: EUROPA HALL (ESE's, ISCA GENERAL ASSEMBLY, CLOSING CEREMONY)
35: OPEN FORUM POSTERS

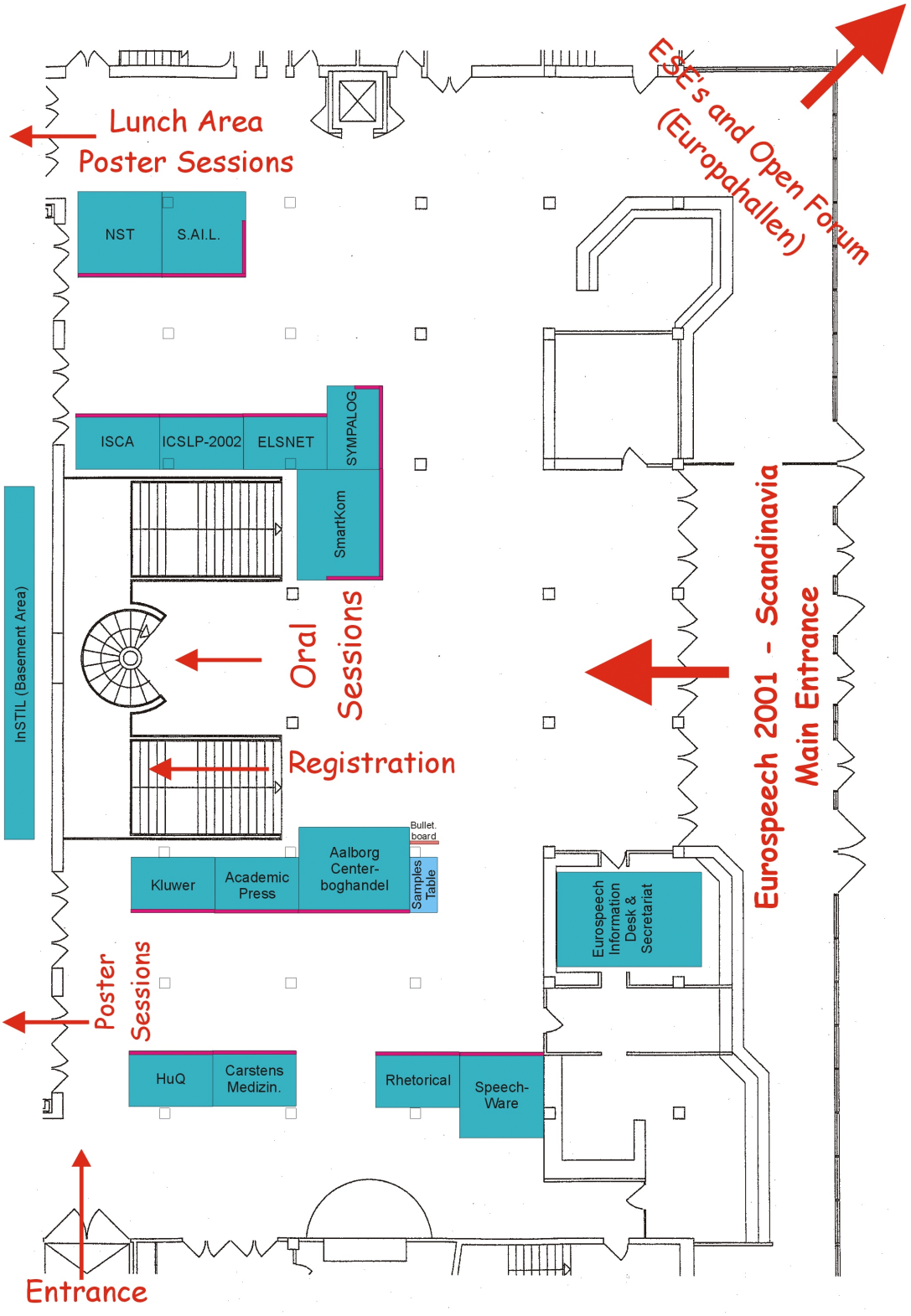
FIRST FLOOR

33: REHEARSAL ROOM EQUIPPED WITH PC, VIDEO BEAMER & OVERHEAD PROJECTOR
36: CONNECTION BETWEEN AALBORGHALLEN & EUROPAHALLEN

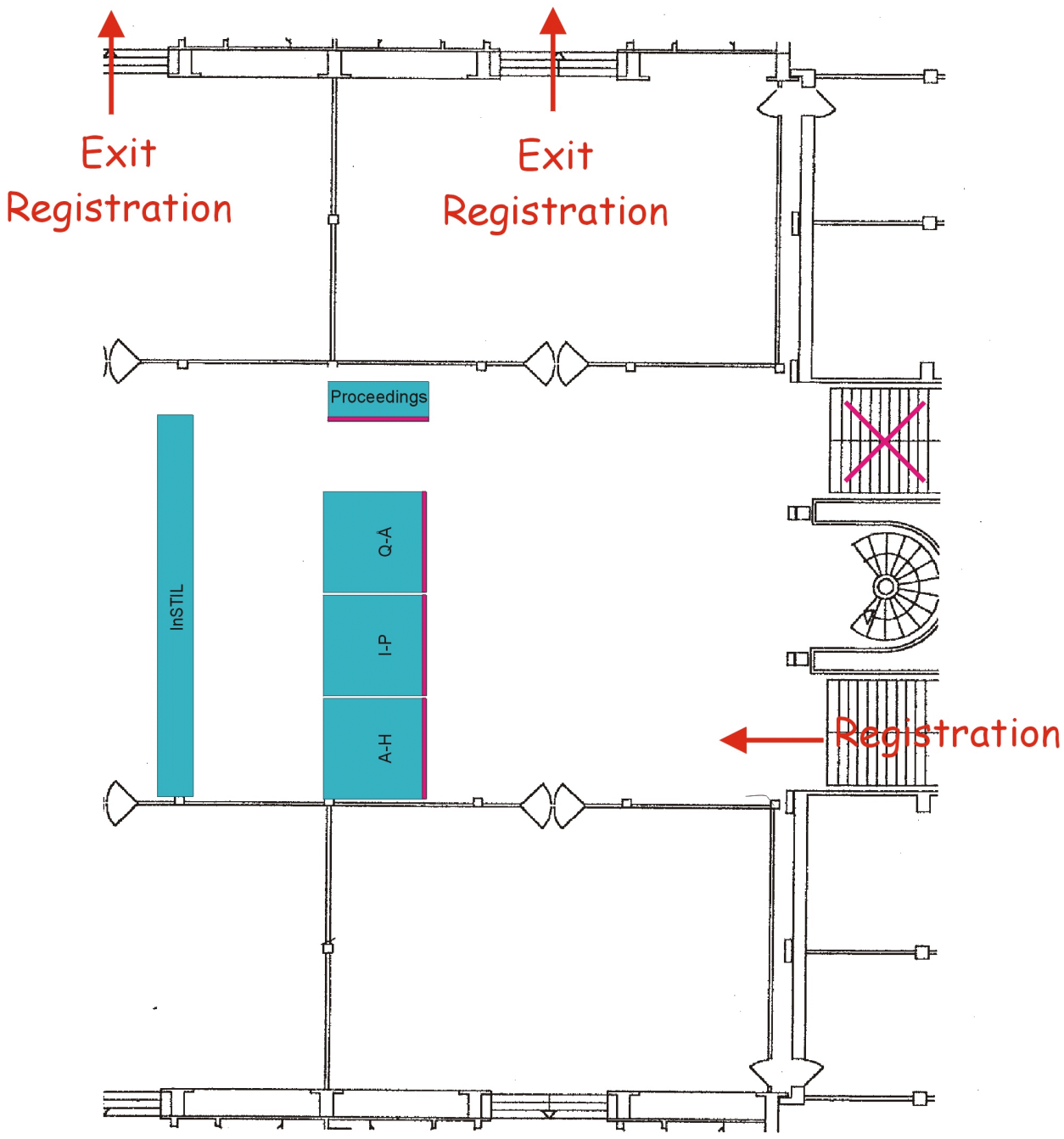
BASEMENT AREA

REGISTRATION, INSTIL EXHIBITION, CLOAK ROOM & TOILETS

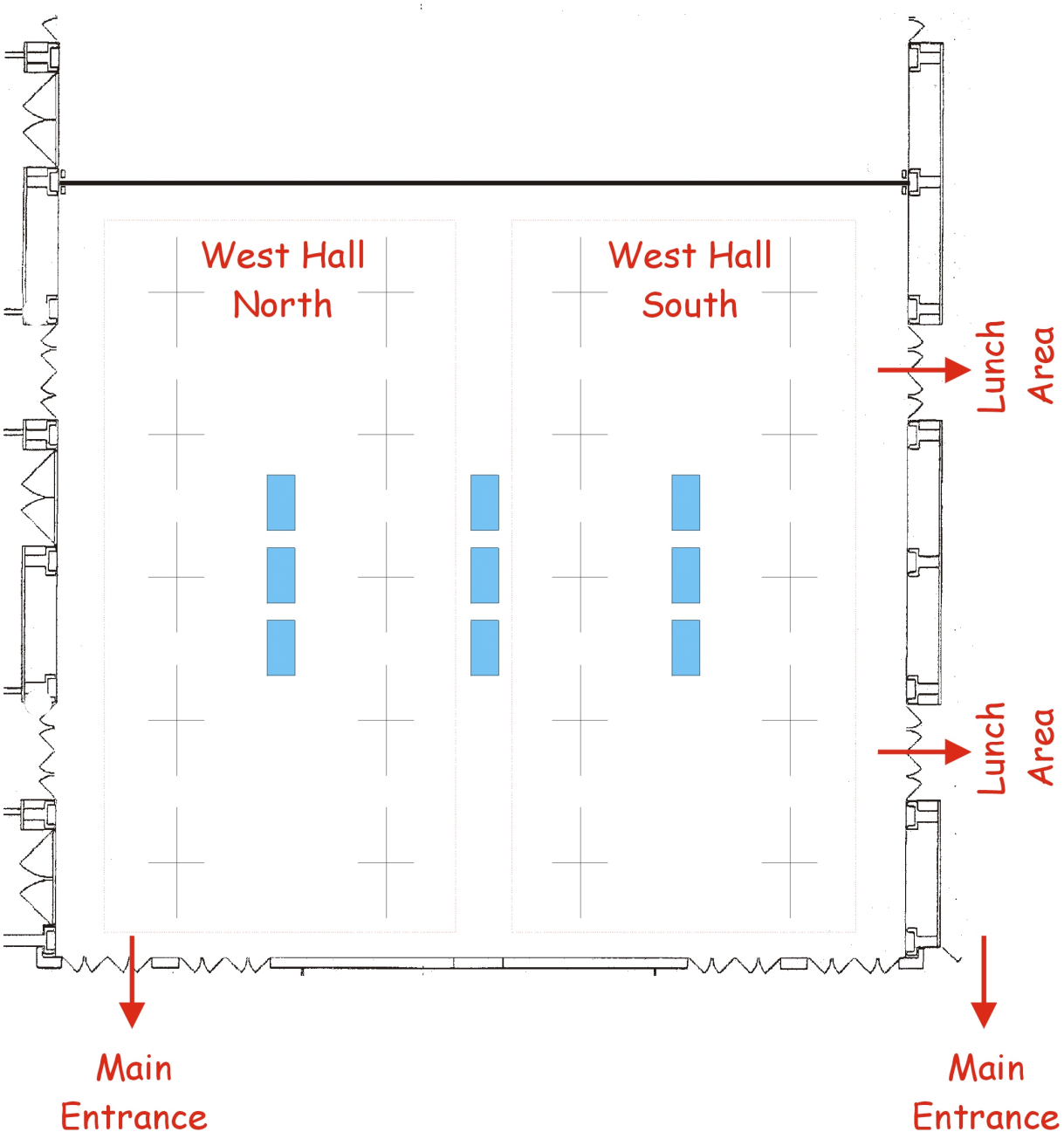
Oral session rooms are each equipped with a PC computer with CD-ROM reader, running Office 2000, and connected to a Video Beamer and to the audio system of the room. Also an overhead projector is available. Test of your presentation may be conducted in ROOM 33 in Europahallen.


Lunch Area
Poster Sessions
ESE's and Open Forum
(Europahallen)
NST
S.A.I.L.
ISCA
ICSLP-2002
ELSNET
SYMPALOG
SmartKom
InSTIL (Basement Area)
Oral Sessions
Registration
Eurospeech 2001 - Scandinavia
Main Entrance
Bullet. board
Kluwer
Academic Press
Aalborg Center-boghandel
Samples Table
Eurospeech Information Desk & Secretariat
Poster Sessions
HuQ
Carstens Medizin.
Rhetorical
Speech-Ware
Entrance

Basement Area - Eurospeech 2001



Poster Session Area - Eurospeech 2001



ICSLP - 2002

7th International Conference on Spoken Language Processing

INTERSPEECH 2002

http://www.icslp2002.org/

September 17 – 20, 2002 Denver, Colorado

General Chair
John H.L. Hansen
CSLR, Univ. Colorado
jhlh@cslr.colorado.edu

Technical Program
Bryan L. Pellom (chair)
bp@cslr.colorado.edu
John H.L. Hansen
Wayne Ward

Finance
Cynthia Ocken
ocken@colorado.edu

Local Arrangements
Centennial Conferences
centennial@orci.com

Corporate Sponsorship
Ronald Cole
cole@cslr.colorado.edu

Publicity
Dan Jurafsky
jurafsky@colorado.edu

Exhibits
Guojun Zhou
Guojun.zhou@intel.com

Tutorials
Shrikanth Narayanan
shri@sipi.usc.edu

Publications
Kathryn Arehart
arehart@spot.colorado.edu

Registration
Wayne Ward
whw@cslr.colorado.edu

Social Activities
Terry Durham
durham@cslr.colorado.edu

Web Master
Jariya Tuantranont
tuantraj@cslr.colorado.edu

Call for Papers

The 7th International Conference on Spoken Language Processing (ICSLP-2002) will be held in Denver, Colorado at the Adams Mark Hotel during Sept. 17 - 20, 2002. ICSLP will be the worlds largest and most comprehensive INTERDISCIPLINARY technical conference focused on speech processing and language technology. For the first time, ICSLP-2002 will also feature a full day of tutorials presented by world-class experts in their fields. ICSLP-2002 is being organized by educators, researchers, and scientists from Center for Spoken Language Research (CSLR), Dept. of Speech, Language and Hearing Sciences (Univ. of Colorado Boulder), and W.G. Voice Research Center. Significant support also comes from additional University and Industry affiliates. Topics of interest for Paper Submission include:

- Phonetics & Phonology
- Prosody & Emotion
- Speaker Assessment
- Discourse and Dialogue
- Speech Modeling
- Speech Physiology
- Audiology & Hearing Aids
- Speech Perception
- Speech Pathology Processing
- Speech Enhancement
- Spoken Dialog Systems
- Speech Generation & Synthesis
- Speech Recognition & Understanding
- Speech Coding and Transmission
- Speech Resources & Standards
- Speaker Recognition & Verification
- Language Identification
- Multi-Lingual Issues in SLP
- Multi-Model SLP (Speech, Facial, Gesture)
- Others

Denver is close to some of the most beautiful scenery in the United States, with the Colorado Rocky Mountains as a backdrop. Many top resorts are close by (Vail, Aspen, Breckenridge, Keystone) offering excellent opportunities for satellite workshops. Colorado Springs is a short drive away, and features the Garden of the Gods and the U.S. Olympic Training Center. Downtown Denver is clean, safe, and home to many excellent restaurants and shopping along its walking street just outside the Adams Mark Hotel. Other attractions include the U.S. Mint, Denver Center for the Performing Arts, and sports teams including the Denver Broncos, Colorado Rockies, and the NHL Stanley Cup Champion Avalanche.

Prospective authors are invited to submit full-length, four-page papers for presentation in any of the areas listed above. Papers will be submitted electronically via WWW with electronic reviewer feedback provided for all submissions. We also encourage the submission of proposals for *tutorials* and *sessions on special topics*. Check the conference website www.icslp2002.org for up-to-date information.

Important Dates:

Proposals for tutorials & special sessions deadline	December 1, 2001
Four-Page, Full Paper Submission deadline	March 30, 2002
Notification of paper acceptance	June 10, 2002
Early registration (one author per paper)	before June 25, 2002
Advanced registration	before August 1, 2002
ICSLP-2002:	September 17-20, 2002

8th European Conference on Speech Communication and Technology

INTERSPEECH 2003

http://www.eurospeech2003.org

September 1– 4, 2003 Geneva, Switzerland

Call for Papers

As the largest interdisciplinary conference on speech processing and language technology, the 8th European Conference on Speech Communication and Technology will be held in the truly international city of Geneva, Switzerland, at the International Conference Centre Geneva (www.cicg.ch) during Sept. 1- 4, 2003. EUROSPEECH-2003 is being organized by scientists and educators from IDIAP (www.idiap.ch, Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny), in close collaboration with the Swiss Federal Institute of Technology Lausanne (EPFL) and University of Geneva. Topics of interest for paper submission include:

- Phonetics & Phonology (incl. Prosody)
- Speech Production
- Speech Perception
- Acquisition, Learning and Pathology of Spoken Language
- Discourse and Dialogue
- Signal Analysis, Processing and Feature Estimation
- Speech Coding and Transmission
- Speech Recognition and Understanding (incl. Language Modeling)
- Speech Generation and Synthesis
- Spoken Dialogue System
- Multimodality
- Resources, Assessment and Standards
- Other Applications of Spoken Language Processing
- Others

Geneva is set in one of the most enchanting country sides, by one of Europe's largest lakes and surrounded by majestic mountains. Geneva is also the ideal gateway for discovering the beautiful scenery of the surrounding lake and mountain region (including Mont-Blanc). Located in the very heart of Europe, Geneva has a long tradition as a venue for major meetings and is a host city to many international organizations. It has been a cultural centre for many centuries and home to many creative spirits in the fields of science and art. Many top resorts (including Verbier and Cran-Montana in the Swiss Valais) are also close by, offering excellent opportunities for satellite workshops.

Prospective authors will be invited to submit full-length, four-page papers for presentation in any of the areas listed above. Papers will be submitted electronically via WWW with electronic reviewer feedback provided for all submissions. Submission of proposals for *tutorials* and *sessions on special topics* will also be encouraged.

General Chair
Hervé Bourlard
IDIAP/EPFL

Vice-Chair
Maghi King
ISSCO, Univ. of Geneva

Scientific Secretaries
Martin Rajman, EPFL
Beat Pfister, ETHZ

Administrative & financial matters
Jean-Pierre Rausis

Publications (proceedings)
Jean-Cédric Chappelier
EPFL

Web Master
Sandra Manzi
ISSCO, Univ. of Geneva

Sponsorship & Industrial Contacts
Murat Kunt, EPFL
Jean-Pierre Rausis

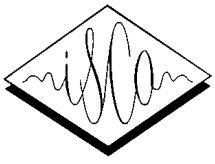
Public & Institutional relations
Bernard Levrat
Univ. of Geneva

Social events & on-site organisation
Maghi King
ISSCO, Univ. of Geneva

Exhibition
Hervé Bourlard
IDIAP/EPFL

Conference Management
SYMPORG SA
7, Avenue Krieg
CH-1208 Geneva
Switzerland
Tel (+4122) 839 84 84
Fax (+4122) 839 84 85
http://www.symporg.ch

Conference Information
http://www.eurospeech2003.org
organisers@eurospeech2003.org



IDIAP
MARTIGNY - VALAIS - SUISSE

GENERAL INSTRUCTIONS FOR SPEAKERS

The REHEARSAL room "Columbine" is located on the first floor of Europahallen. The room is at your disposal for testing your presentation on a PC with similar setup to the ones in the oral session rooms.
If your presentation is in electronic form then please, bring it on a CD-ROM (preferable) or floppy disk.
Please bring a set of overheads containing your presentation in case there are inconsistencies between the system on which you originally designed your presentation and the PC system available in the session room.

INSTRUCTIONS FOR ORAL PRESENTERS

1 You should be in the allocated oral session room at least 10 minutes in advance of the session, in order to introduce yourself to the chairperson
2 If you give a PC-based presentation or demonstration, please contact the technical assistant of the respective session room well in advance to load your file onto the local PC. The specifications of the PC and installed software are given below. If you do need to use your laptop, please refer to the specifications of the AV equipment given below. If you have special needs, contact the Local Organisation in the Internet Service Room as soon as possible
3 All oral presentations are assigned 20 minutes in total for their presentation. You have 15 minutes to present your paper. The remaining 5 are reserved for the introduction by the chairperson and questions/answers upon completion of your presentation
4 Please make sure that you do not exceed the above time limit. Your chairperson has been instructed to be very strict on timing and ensure that the session ends on time as we have to enable synchronisation across sessions.
5 If you have a demo included in your oral presentation, please be aware that this is part of the 20 minutes time-slot of your presentation.

INSTRUCTIONS FOR POSTER PRESENTERS

1 Poster sessions are being held at the Congress Hall - West closest to the main entrance of the Congress Centre.
2 At the entrances of the Poster room, you will find a map giving an overview of poster locations and a scheme that identifies your poster by a location number
3 For posters that include a demo - please notify the organisers in advance - a table for your equipment will be available next to the poster board.
4 Posters scheduled for one of the morning sessions all have to be 'boarded' before 09.00. They will be on display during the full morning, and shall be removed at lunch time. Posters for one of the afternoon sessions must be 'boarded' before 14.00 - Friday before 13.30!
5 You are requested to be present at your poster at the respective session during the entire time slot
6 For any assistance you might need, please contact the student assistants who will be in the room before and during the session.

INSTRUCTIONS FOR OPEN FORUM PRESENTERS

For the first time at a Eurospeech conference there will be an opportunity for ALL registered participants to exhibit an A3 sized mini-poster on ANY topic relating to the conference. These may be used to provide last minute results, to raise important discussion points or simply to advertise the skills of the author to prospective employers.
Please follow the instructions below:

1 Your poster must be an A3-sized mini-poster
2 Name, photo, affiliation and contact details of the presenter must be displayed together with the mini-poster. You should indicate your time schedule for being present at your mini-poster
3 Layout and all other content is up to the individual, and people are encouraged to be imaginative.
4 The mini-posters will be exhibited for the full duration of the conference
5 A prize will be awarded by ISCA for the most inventive mini-poster.

SPECIFICATIONS OF PRESENTATION PC'S AND AV EQUIPMENT

PC: Pentium/Celeron 600 MHz or more; Video resolution: 1024 X 768 - 24 bit colour; 16 bit Stereo Audio; Internet Web acces; Microsoft Windows2000; Microsoft Office 2000 (including PowerPoint 2000); Internet Explorer / Netscape Navigator and Adobe Acrobat Reader. Please, avoid using special fonts etc.

AV: In oral session rooms
Video Projector resolution: 1024 x 768; Video Connector: Standard VGA 15 pin; Audio Connector. Please, check in advance how to switch the video output to the external connector to your laptop.

The presentation should not depend on internet access, due to access problems that might be without reach of the Local Organisers.

EUROSPEECH 2001 - SCANDINAVIA

Hour	Sunday 02.09					
16.00 - 20.00	Registration and Information - Room: Basement Area					
	Monday 03.09					
08.30 - 10.00	Registration and Information - Room: Basement Area					
10.00 - 12.30	Opening Ceremony - ISCA Medallist and Keynote Speeches Room: **Concert- and Congress Hall** A21					
12.30 - 14.00	Lunch - South Foyer					
Room	***Europa Hall Oral/ Poster***	***"Musiksalen" Oral***	***"Radiosalen" Oral***	***"The Little Theatre" Oral***	***West Hall - North Poster***	***West Hall - South Poster***
14.00 - 15.20	ESE1 - What do Industry and Universities Expect from Each Other?: Panel discussion A31	Linguistic Modelling: Language Model Compression A32	Speech Production: Voice Source A33	Speech Recognition and Under-standing: Pronunciation and Subword Units - I A34	Phonetics and Phonology: Prosody and Others A35	Speech Perception: First and Second Language Learning Speech Perception: Miscellaneous - I A36a/A36b
15.20 - 15.50	Coffee/Tea Break					
15.50 - 17.30	ESE2 - Noise Robust Recognition: Front-end and Compensation Algorithms A41	Linguistic Modelling: Language Model Adaptation A42	Speech Production: Articulation A43	Speech Recognition and Understanding: Topic Detection and Information Retrieval A44	Phonetics and Phonology: Segmentals and Synthesis A45	Speech Perception: Miscellaneous - II A46
17.30 - 18.30	Break					
18.30 - 20.00	Welcome Reception - Room: South Foyer					

Hour	Tuesday 04.09					
09.00 - 10.40	ESE2 - Noise Robust Recognition: Front-end Algorithms B11	Linguistic Modelling: Semantic Modelling B12	Speech Perception: Recognition and Intelligibility B13	Speech Recognition and Understanding: LVCSR - I B14	Speech Synthesis: Systems and Prosody B15	Speech Recognition and Understanding: Articulatory and Perceptual Approaches to ASR B16
10.40 - 11.10	Coffee/Tea Break					
11.10 - 12.30	ESE2 - Noise Robust Recognition: Robust systems - What helps ? B21	Phonetics and Phonology: Segmentals B22	Speech Production: Prosody - I B23	Speech Recognition and Understanding: Acoustic Modelling - I B24	Linguistic Modelling: Language Models - I B25	Speaker Recognition: Identification, Verification and Tracking. Speech Recognition and Understanding: Language Identification B26
12.30 - 14.00	Lunch - South Foyer					
14.00 - 15.20	ESE3 - Imagination 2001 B31	Phonetics and Phonology: Prominence and Timing B32	Speech Synthesis: Concatenation - I B33	Speech Recognition and Understanding: LVCSR - II B34	Speech Recognition and Understanding: Noise Robustness - I B35	Speech Production: Prosody - II B36
15.20 - 15.50	Coffee/Tea Break					
15.50 - 17.30	ESE3 - Imagination 2001 - Continued B41	Speech Synthesis: Concatenation - II B42	Signal Analysis: Microphone Arrays & Source Localisation B43	Speech Recognition and Understanding: Audio-Visual Processing B44	Linguistic Modelling: Language Models - II B45	Speech Recognition and Understanding: Noise Robustness - II B46
17.30 - 19.00	Break					
19.00 -	ISCA General Assembly **Room: Europa Hall**					

Hour	Wednesday 05.09					
09.00 - 10.40	ESE4 - SIGshow C11	Speech Synthesis: Prosody C12	Applications: Multimodal Applications C13	Speech Recognition and Understanding: Speaker Adaptation C14	Speech Recognition and Understanding: Adaptation C15	Dialogue Systems: Project Descriptions - I C16
10.40 - 11.10	Coffee/Tea Break					
11.10 - 12.30	ESE4 - SIGshow - Continued C21	Dialogue Systems: Resources C22	Speaker Recognition: Features and Transforms C23	Speech Perception: Prosody C24	Speech Recognition and Understanding: Pronunciation and Subword Units - II C25	Speech Production: Miscellaneous C26
12.30 - 14.00	Lunch - South Foyer					
14.00 -	Free Afternoon and Evening Groups may address the Organisers for Reservation of a Meeting Room					

Hour	Thursday 06.09					
09.00 - 10.40	ESE5 - Existing and Future Corpora: Next Generation Speech Resources D11	Dialogue Systems: Project Descriptions - II D12	Signal Analysis: Speech Processing in Car Environments D13	Speech Recognition and Understanding: Finite State Transducers for ASR D14	Speech Recognition and Understanding: Acoustic Modelling - II D15	Resources, Assessment and Standards: Assessment Tools & Methodology D16
10.40 - 11.10	Coffee/Tea Break					
11.10 - 12.30	ESE5 - Existing and Future Corpora: Automated Analysis of Speech Resources D21	Dialogue Systems: Dialogue Systems and Generation D22	Speaker Recognition: Alternative Trends in Verification - I D23	Speech Recognition and Understanding: Speech Understanding D24	Speech Recognition and Understanding: Algorithms and Architectures D25	Signal Analysis: Speech Enhancement and Noise Processing D26
12.30 - 14.00	Lunch - South Foyer					
14.00 - 15.20	ESE6 - Education Arena D31	Speech Synthesis: Grapheme-to-Phoneme Conversion D32	Signal Analysis: Speech Enhancement D33	Speech Recognition and Understanding: Discriminative Training D34	Speech Coding: Advances in Speech Coding D35	Resources, Assessment and Standards: Corpora D36
15.20 - 15.50	Coffee/Tea Break					
15.50 - 17.30	ESE6 - Education Arena - Continued D41	Resources, Assessment and Standards: Assessment Methodology D42	Speech Recognition and Understanding: Confidence Measures D43	Speech Recognition and Understanding: Language Modelling D44	Dialogue Systems: Techniques and Strategies D45	Speech Synthesis: Miscellaneous D46
17.30 - 19.30	Break					
19.30 -	Songs from Musicals 20.00, Room: **Concert- and Congress Hall** Open Buffet 20.45, Room: **South Foyer**					

Hour	Friday 07.09					
09.00 - 10.40	ESE7 - Integration of Phonetic Knowledge in Speech Technology: Experiments and Experiences E11	Speech Coding: Wideband Speech Coding E12	Dialogue Systems: Techniques and Strategies E13	Speech Recognition and Understanding: Robust ASR E14	Applications: Miscellaneous Applications E15	Signal Analysis: Pitch and Speech Analysis E16
10.40 - 11.10	Coffee/Tea Break					
11.10 - 12.30	ESE7 - Integration of Phonetic Knowledge in Speech Technology: Is Phonetic Knowledge any use? Panel discussion E21	Speech Coding: Speech Transmission Systems E22	Speaker Recognition: Alternative Trends in Verification - II E23	Speech Recognition and Understanding: Rhythm and Timing in ASR E24	Speech Recognition and Understanding: Confidence Measures and OOV E25	Signal Analysis: Source Localisation and Beam Forming E26
12.30 - 13.30	Lunch - South Foyer					
13.30 - 14.50	Signal Analysis: Speech Features and Modelling E31	Speech Recognition and Understanding: Kids, Toys and Emotions E32	Applications: Media Applications E33	Speech Recognition and Understanding: Distributed Speech Recognition E34	Speech Recognition and Understanding: Prosody and Cross-Language in ASR E35	Education: Education and Training Speaker Recognition: Features and Robustness E36a/E36b
14.50 - 15.00	Break					
15.00 - 16.00	Closing Ceremony **Room: Europa Hall**					

13 August 2001

Eurospeech 2001 - Scandinavia

Authors Index

A

Abdel-Kader N S . . . . . B36 . . 951
Abdou S . . . . . D24 . 1783
Abutalebi H R . . . . . D26 . 1855
Acero A . . . . . D26 . 1903
Acero A . . . . . A41 . . 217
Acero A . . . . . B35 . . 901
Achariyakulporn V . . . . . B26 . . 775
Adank P . . . . . B13 . . 481
Adda G . . . . . A42 . . 255
Adda-Decker M . . . . . A42 . . 255
Afify M . . . . . C16 . 1323
Afify M . . . . . E14 . 2355
Afify M . . . . . B14 . . 495
Afify M . . . . . B21 . . 633
Afify M . . . . . B34 . . 851
Ahkuputra V . . . . . E35 . 2753
Ahn D-H . . . . . D25 . 1821
Ahn S . . . . . A35 . . 107
Ainsworth W . . . . . A46 . . 371
Ainsworth W A . . . . . B13 . . 477
Akagi M . . . . . E16 . 2439
Akahane-Yamada R . . . . . A36a . 145
Akiba T . . . . . B25 . . 705
Albino N . . . . . E32 . 2679
Alexander G . . . . . D12 . 1559
Alku P . . . . . B36 . . 919
Allen J . . . . . C16 . 1323
Alm N . . . . . E15 . 2401
Alm N . . . . . E15 . 2405
Alonso L . . . . . D23 . 1753
Altosaar T . . . . . D11 . 1537
Álvarez Salgado J F . . . . . E15 . 2393
Álvarez-Marquina A . . . . . E26 . 2615
Álvarez-Marquina A . . . . . E26 . 2619
Alwan A . . . . . A36b . 175
Alwan A . . . . . A36b . 179
Alwan A . . . . . A41 . . 185
Alwan A . . . . . E34 . 2703
Amador-Hernandez M . . . . . B15 . . 513
Amaral R . . . . . E33 . 2689
Ambikairajah E . . . . . A46 . . 411
Amir N . . . . . A35 . . 127
Ammicht E . . . . . D45 . 2217
Andersen E . . . . . E32 . 2675
Andersen O . . . . . E11 . 2289
Andersen O . . . . . B42 . . 975
Ando A . . . . . B14 . . 495
Ando A . . . . . B25 . . 709
Ando A . . . . . B34 . . 859
Andorno M . . . . . E13 . 2325
Andrassy B . . . . . A41 . . 193
André-Obrecht R . . . . . C11 . 1145
Andreasen P N . . . . . E15 . 2401
Andrews W D . . . . . E23 . 2517
Angkititrakul P . . . . . D13 . 1573
Angkititrakul P . . . . . D36 . 2023
Antonio B . . . . . E32 . 2679
Applebaum T H . . . . . B15 . . 537
Arai T . . . . . A36a . 149
Arai T . . . . . E26 . 2639
Arai T E36a 2791
Arai T . . . . . A46 . . 391
Arai T . . . . . B13 . . 473
Araki K . . . . . A46 . . 375
Araki K . . . . . B36 . . 935
Araki M . . . . . C16 . 1283
Araki M . . . . . D22 . 1743
Araki M . . . . . D45 . 2185
Araki S . . . . . E26 . 2595
Araki S . . . . . E26 . 2599
Arcienega M E36b 2821
Ariki Y . . . . . C23 . 1381
Ariki Y . . . . . D26 . 1879
Ariki Y . . . . . E25 . 2577
Ariyaeeinia A M . . . . . B26 . . 795
Arnott J L . . . . . E15 . 2405
Arunachalam S . . . . . E32 . 2675
Asano F . . . . . B43 . 1013
Ashby S . . . . . A45 . . 321
Asoh H . . . . . B43 . 1013
Asunción M . . . . . E32 . 2679
Attias H . . . . . D26 . 1903
Attwater D . . . . . C16 . 1323
Aubert X . . . . . B14 . . 499
Auer E . . . . . A36b . 179
Avanzini F . . . . . A33 . . . 51
Avesani C . . . . . B15 . . 509
Aylett M P . . . . . C26 . 1491
Azzini I . . . . . C16 . 1327

B

Baba A . . . . . C14 . 1219
Baba A . . . . . D15 . 1657
Baba A . . . . . B35 . . 869
Babic R . . . . . A41 . . 229
Bacchiani M . . . . . C16 . 1299
Bacchiani M . . . . . E15 . 2373
Backfried G . . . . . D36 . 2039
Bae K S . . . . . D35 . 2005
Bahoura M . . . . . D33 . 1937
Bakcsi Z . . . . . D36 . 2059
Baloyi N T E36b 2833
Bannert R . . . . . B36 . . 923
Barbosa P . . . . . B36 . . 967
Barkat M . . . . . B45 . 1065
Barker J P . . . . . A41 . . 213
Baron D . . . . . C22 . 1359
Barros M J . . . . . D16 . 1707
Barry W . . . . . B42 . . 975
Batliner A . . . . . E11 . 2285
Batliner A . . . . . E35 . 2781
Batusek R . . . . . D42 . 2099
Baudry M . . . . . E31 . 2661
Bauer J G . . . . . C25 . 1417
Bauer J G . . . . . D15 . 1633
Baumann S . . . . . B15 . . 557
Bazzi I . . . . . A34 . . . 61
Beaugeant C . . . . . A41 . . 193
Bechet F . . . . . B25 . . 725
Beckham J L . . . . . C22 . 1363
Behne D M . . . . . A36a . 133
Bel B . . . . . C11 . 1144
Bellegarda J . . . . . C12 . 1167
Bellegarda J . . . . . B12 . . 455
Bellot O . . . . . C15 . 1245
Benarousse L . . . . . B26 . . 799
Benest I . . . . . D22 . 1739
Bengio S . . . . . B16 . . 607
Bengler K . . . . . E15 . 2421
Benitez C . . . . . B11 . . 429
Benitez M C . . . . . A41 . . 221
Benítez M C . . . . . D15 . 1625
Berckmans D . . . . . E16 . 2435
Berger J . . . . . D16 . 1687
Beringer N . . . . . D45 . 2197
Bernard A . . . . . E34 . 2703
Bernsen N O . . . . . C13 . 1189
Bernstein J E36a 2803
Bernstein L . . . . . A36b . 179
Berthommier F . . . . . B44 . 1023
Bertoldi N . . . . . A42 . . 239
Besacier L . . . . . C16 . 1291
Besacier L . . . . . D36 . 2043
Beskow J . . . . . A46 . . 387
Bessette B . . . . . D35 . 1997
Bessette B . . . . . E12 . 2303
Beyerlein P . . . . . C25 . 1457
Beyerlein P . . . . . B14 . . 499
Bigbee T . . . . . D11 . 1533
Bimbot F . . . . . C11 . 1145
Bimbot F . . . . . C21 . 1344
Bimbot F . . . . . E11 . 2293
Bimbot F . . . . . E23 . 2513
Binnenpoorte D . . . . . D16 . 1679
Binnenpoorte D . . . . . D36 . 2051
Bisani M . . . . . C25 . 1429
Blache P . . . . . B45 . 1061
Black A . . . . . A36b . 171
Black A W . . . . . D32 . 1919
Blanchon H . . . . . C16 . 1291
Blocher A . . . . . D12 . 1547
Bloothooft G . . . . . C11 . 1144
Bloothooft G . . . . . A33 . . . 39
Blouet R . . . . . E23 . 2513
Bobrovsky B-Z . . . . . D15 . 1637
Bode M . . . . . D26 . 1867
Boë L-J . . . . . A36b . 163
Boë L-J . . . . . A36b . 167
Boeffard O . . . . . D16 . 1711
Boeffard O . . . . . B33 . . 829
Boeffard O . . . . . B42 . . 983
Bohus D . . . . . D43 . 2121
Bonastre J-F . . . . . C11 . 1145
Bonastre J-F . . . . . C15 . 1245
Bonastre J-F . . . . . C21 . 1344
Bonneau A . . . . . A45 . . 313
Bonneau A . . . . . B13 . . 469
Botha E C . . . . . C15 . 1249
Botha E . . . . . D36 . 2055
Botinis A . . . . . B23 . . 657
Botinis A . . . . . B36 . . 923
Boucher J-M . . . . . E12 . 2311
Boudy J . . . . . D36 . 2059
Boula de Mareüil P . . . . . D32 . 1923
Boulianne G . . . . . C14 . 1207
Boulianne G . . . . . D14 . 1595
Boulis C . . . . . E34 . 2715
Bourlard H . . . . . A41 . . 225
Bourlard H . . . . . E35 . 2765
Bourlard H . . . . . B16 . . 587
Bourlard H . . . . . B16 . . 607
Bouwman G . . . . . E25 . 2585

Boves L ..... A41 .. 201
Boves L ..... E15 . 2381
Boves L ..... E25 . 2585
Boves L ..... B11 .. 433
Boves L ..... B35 .. 865
Bracy A ..... D12 . 1559
Braga D ..... D16 . 1707
Bravetti P ..... A32 ... 21
Brennan S E ..... A21 ... 11
Breuer S ..... B15 .. 521
Broeder D ..... D11 . 1529
Brøndsted T ..... B45 . 1089
Brøndsted T ..... E35 . 2777
Brugman H ..... D11 . 1529
Brugnara F ..... D15 . 1641
Brungart D S ..... B26 .. 747
Buchner H ..... B43 . 1001
Buckow J ..... E35 . 2781
Bulyko I ..... B42 .. 987
Burger S ..... D36 . 2043
Burget L ..... B11 .. 429
Burnett I S ..... D35 . 1989
Burnett I S ..... E12 . 2315
Burnham D ..... A46 .. 395
Butzberger J ..... B24 .. 679
Byrd D ..... E32 . 2675
Byrne W ..... C14 . 1203
Byrne W ..... E25 . 2569
Byrne W ..... B14 .. 487
Byrne W ..... A34 ... 57

C

Cai Q ..... D45 . 2169
Callan A ..... A36a . 145
Callan D ..... A36a . 145
Campbell J P ..... E23 . 2517
Campbell N ..... C11 . 1144
Campbell N ..... C11 . 1149
Campbell N ..... D11 . 1525
Campbell N ..... E15 . 2409
Campbell N ..... A45 .. 337
Campbell N ..... A45 .. 361
Carayannis G ..... D36 . 2075
Carayannis G ..... B33 .. 837
Cardenal-Lopez A ..... E15 . 2369
Carpenter P ..... D43 . 2121
Carson-Berndsen J ..... B45 . 1073
Carson-Berndsen J ..... E11 . 2281
Carson-Berndsen J ..... A45 .. 321
Carter D ..... B12 .. 443
Casacuberta F ..... E11 . 2297
Casacuberta F ..... E15 . 2385
Caseiro D A ..... D44 . 2131
Caspers J ..... C24 . 1395
Castelli E ..... E15 . 2417
Cernocky J ..... D36 . 2059
Cervera T ..... A46 .. 371
Chalamandaris A ..... B33 .. 837
Chan C C ..... D12 . 1551
Chan S F ..... D12 . 1551
Chan Y C ..... B46 . 1095
Chang E ..... C15 . 1261
Chang E ..... C23 . 1377
Chang E ..... D34 . 1951
Chang E ..... D45 . 2169
Chang E ..... E35 . 2741
Chang E ..... E35 . 2769
Chang E E36a 2799
Chang S ..... D21 . 1725
Chang S ..... D21 . 1729
Chang S-C ..... C16 . 1307
Chang W-W ..... B26 .. 767
Chang W-W ..... B26 .. 771
Chao H ..... D34 . 1951
Charlet D ..... D43 . 2113
Charnvivit P ..... E35 . 2753
Chaudhari U V ..... C23 . 1389
Chazan D ..... D15 . 1637
Chazan D ..... D25 . 1789
Chazan D ..... E16 . 2427
Cheetham B M G ..... D35 . 1985
Chen A ..... C24 . 1403
Chen B ..... B45 . 1045
Chen B ..... A44 .. 299
Chen B ..... B11 .. 429
Chen F ..... E35 . 2737
Chen F ..... E35 . 2769
Chen J ..... A41 .. 233
Chen J ..... B16 .. 571
Chen K ..... D12 . 1551
Chen L ..... A42 .. 255
Chen M ..... A36b . 175
Chen S-H ..... E35 . 2773
Chen T ..... C23 . 1377
Chen W ..... C12 . 1159
Chen Y-C ..... B45 . 1081
Cheng Y M ..... B11 .. 425
Cheng Y-W ..... D26 . 1895
Chengalvarayan R ..... E35 . 2733
Chengalvarayan R ..... B35 .. 897
Chi H S ..... D12 . 1551
Chien J-T ..... B46 . 1131
Chien S-C ..... C16 . 1307
Chin-Hui L ..... E25 . 2573
Ching P C ..... D12 . 1551
Choi S-W ..... B33 .. 841
Choi W N ..... D12 . 1551
Choi Y-S ..... E16 . 2447
Chollet G ..... D23 . 1761
Chotimongkol A ..... D25 . 1829
Chou F-C ..... C25 . 1445
Choukri K ..... A31 ... 17
Christensen H ..... E11 . 2289
Chu M ..... D42 . 2087
Chu M ..... B36 .. 927
Chu Y C ..... B26 .. 767
Chun H J ..... E15 . 2397
Chung H ..... B32 .. 815
Chung J-H E36b 2829
Chung M ..... D25 . 1821
Ciocca V ..... A46 .. 395
Clement C J ..... C26 . 1471
Cohen A ..... D25 . 1793
Cohen G ..... E16 . 2427
Cohen I ..... D33 . 1933
Colás J ..... C16 . 1279
Colás J ..... D45 . 2165
Cole D ..... D36 . 2031
Cole R ..... D36 . 2023
Coletti P ..... D36 . 2043
Colotte V ..... B13 .. 469
Cook N ..... D22 . 1739
Cooke M ..... A41 .. 213
Cooke M E36a 2795
Córdoba R ..... C16 . 1279
Córdoba R ..... A45 .. 357
Cosi P ..... B44 . 1035
Cosi P ..... B15 .. 509
Coulston R ..... A45 .. 365
Couvreur C ..... E26 . 2635
Couvreur L ..... E26 . 2635
Cowie R ..... A35 ... 87
Cox S ..... E13 . 2337
Cranen B ..... B35 .. 865
Cremelie N ..... C25 . 1421
Crépy H ..... A32 ... 21
Cronk A ..... D36 . 2031
Crouzet O ..... B13 .. 477
Csatári F E36a 2807
Cucchiarini C ..... D16 . 1679
Cucchiarini C ..... D42 . 2091
Cui X ..... A41 .. 185
Cuperman V ..... D35 . 2013
Cuperman V ..... E16 . 2479
Cutler A ..... B13 .. 465
Cyphers D S ..... C16 . 1331
Czepulonis M ..... C26 . 1519
Czigler P E ..... A36a . 133
d'Alessandro C ..... A33 ... 47

D

D'Imperio M ..... A35 ... 99
Dahmen J ..... D25 . 1825
Damnati G ..... B35 .. 885
Daoudi K ..... D15 . 1669
Dashtseren E ..... A46 .. 415
Daubias P ..... B44 . 1031
de Cheveigné A ..... E16 . 2451
de la Torre A ..... E22 . 2503
de la Torre Á ..... D15 . 1625
de Wet F ..... A41 .. 201
de Wet F ..... B11 .. 433
de Wet F ..... B35 .. 865
de Veth J ..... A41 .. 201
de Jong J H A L E36a 2803
de Veth J ..... B11 .. 433
de Veth J ..... B35 .. 865
De Schutter G ..... A35 ... 75
De Mori R ..... B25 .. 725
Degerstedt L ..... D45 . 2193
Delcloque P ..... C21 . 1349
Deleglise P ..... B44 . 1031
Deligne S ..... D25 . 1833
Delvaux V ..... B22 .. 647
Delvaux V ..... B22 .. 651
Demolin D ..... B22 .. 651
Demuynck K ..... D15 . 1621
Deng L ..... D26 . 1903
Deng L ..... A41 .. 217
Deng L ..... B16 .. 603
Deng L ..... B35 .. 901
Dermatas E ..... B43 . 1009
Dermatas E ..... E26 . 2631
Dermatas E ..... B43 .. 997
Deviren M ..... D15 . 1669
Dharanipragada S ..... E14 . 2359
Di Fabbrizio G ..... C16 . 1339
Di Fabbrizio G ..... C22 . 1363
Díaz Martín J C ..... E15 . 2393
Diaz-de-Maria F ..... B46 . 1103
Digalakis V ..... B33 .. 833
Dines J D S ..... D46 . 2239
Dobler S ..... D25 . 1837
Dobler S A31 7
Docio-Fernandez L ..... E15 . 2369
Doddington G ..... E23 . 2521
Dogil G ..... B23 .. 665
Donovan R E ..... A45 .. 329

Douglas-Cowie E ..... A35 ... 87
Doval B ..... A33 ... 47
Dowding J ..... B25 .. 729
Draxler C ..... D11 . 1524
Draxler C ..... E15 . 2421
Droppo J ..... A41 .. 217
Drygajlo A ..... D26 . 1887
Drygajlo A E36a 2787
Drygajlo A E36b 2821
du Preez J A ..... A44 .. 283
Duchateau J ..... D15 . 1621
Dumouchel P ..... C14 . 1207
Dumouchel P ..... D14 . 1595
Dupont S ..... B11 .. 429
Durston P ..... C16 . 1323
Dybkjær L ..... C13 . 1189
Dybkjær L ..... C21 . 1345
Dye R ..... E15 . 2405

E

Ealey D ..... B11 .. 425
Ealey D ..... B11 .. 437
Edmondson W ..... B16 .. 595
Eichner M ..... C25 . 1433
Eide E ..... D15 . 1613
Eide E ..... D25 . 1833
El-Ramly S H ..... B36 .. 951
Elgendy A M ..... A43 .. 269
Ellis D P W ..... A41 .. 189
Ellouze N ..... E16 . 2471
Elordieta G ..... A35 .. 115
Emori T ..... D15 . 1649
Engwall O ..... C26 . 1475
Engwall O ..... A43 .. 261
Erdogan H ..... B14 .. 503
Eriksson A ..... C11 . 1144
Eriksson A ..... A46 .. 399
Eskenazi M ..... A36b . 171
Espada Bueno P ..... E15 . 2393
Esteve Y ..... B25 .. 725
Estienne C ..... E25 . 2561
Etemoglu C O ..... D35 . 2013
Eto M ..... D46 . 2255
Euler S ..... C11 . 1145
Euler S ..... B26 .. 783
Evans N W D ..... B35 .. 893
Ezzaidi H E36b 2825

F

Fábián T ..... E24 . 2535
Faizakov A ..... D25 . 1793
Fakotakis N ..... D26 . 1899
Fakotakis N ..... E13 . 2329
Fakotakis N ..... A42 .. 247
Fakotakis N ..... E16 . 2475
Falavigna D ..... C16 . 1327
Faltlhauser R ..... B26 .. 751
Fant G ..... B23 .. 657
Farinas J ..... E24 . 2539
Farrell M ..... C16 . 1323
Farrugia M ..... E12 . 2307
Faulkner D ..... B15 .. 513
Faundez-Zanuy M ..... D35 . 1977
Federico M ..... A42 .. 239
Fegyó T ..... C25 . 1465
Feijóo S ..... A36b . 155
Fék M ..... E12 . 2311
Feng Y ..... B36 .. 927
Ferencz A ..... B33 .. 841
Fernández S ..... A36b . 155
Ferreiros J ..... C16 . 1279
Ferreiros J ..... E25 . 2553
Ferrer L ..... E25 . 2561
Finan R ..... E13 . 2341
Fine S ..... D23 . 1757
Fink G A ..... E24 . 2527
Fink G A ..... B16 .. 615
Fischer A ..... B21 .. 625
Fissore L ..... E13 . 2325
Fitt S ..... B45 . 1069
Fitt S ..... D46 . 2235
Floricic F ..... D32 . 1923
Flueckiger M ..... D12 . 1563
Foldvik A K ..... A35 .. 111
Fortuna J ..... B26 .. 795
Fosler-Lussier E ..... C16 . 1323
Fosler-Lussier E ..... D22 . 1735
Fosler-Lussier E ..... D45 . 2217
Fosler-Lussier E ..... B12 .. 447
Fotinea S-E ..... D36 . 2075
Fougeron C ..... B22 .. 639
Founda M ..... B33 .. 837
Fouquet Y ..... C16 . 1291
Fourakis M ..... B36 .. 923
Franco H ..... B24 .. 679
Francois H ..... B33 .. 829
Frankel J ..... B16 .. 599
Fränti P ..... E26 . 2627
Franz M ..... A44 .. 287
Frauenfelder U H ..... B22 .. 639
Freitas D ..... D16 . 1707
Frey B J ..... B35 .. 901
Frid J ..... B36 .. 915
Fuchs S ..... C26 . 1487
Fujie S ..... D45 . 2173
Fujimoto M ..... D26 . 1879
Fujisaki H ..... B23 .. 661
Fuliang W ..... B45 . 1077
Funaki K ..... E31 . 2649
Fung P ..... C25 . 1425
Fung P ..... A34 ... 57
Fung T Y ..... D12 . 1551
Fung W-N ..... A45 .. 325
Furui S ..... D24 . 1771
Furui S ..... B14 .. 491

G

Gajic B ..... B16 .. 591
Gales M J F ..... B24 .. 675
Gallant S ..... D36 . 2023
Gallardo-Antolin A ..... B46 . 1103
Galley M ..... D22 . 1735
Galunov V ..... D36 . 2059
Gao Y ..... B14 .. 503
Garcia Molina G E36a 2787
Garcia Lecumberri M L E36a 2795
García Zapata J L ..... E15 . 2393
Garcia-Mateo C ..... E15 . 2369
Garudadri H ..... B11 .. 429
Gauvain J-L ..... C15 . 1241
Gauvain J-L ..... A42 .. 255
Geoffrois E ..... B26 .. 799
Georgila K ..... A42 .. 247
Gibbon D ..... D36 . 2063
Gibbon D ..... A35 ... 83
Gibbon D ..... A35 ... 95
Gick B ..... A43 .. 273
Gielen S ..... A35 ... 87
Gillis S ..... A35 ... 75
Gilloire A ..... D26 . 1871
Glass J ..... C16 . 1331
Glass J ..... C16 . 1335
Glass J ..... C25 . 1437
Glass J ..... A34 ... 61
Goddijn S ..... D16 . 1679
Goel V ..... E25 . 2569
Goel V ..... B14 .. 503
Goldman J-P ..... B22 .. 639
Gómez Vilda P ..... E15 . 2393
Gómez-Vilda P ..... E26 . 2615
Gómez-Vilda P ..... E26 . 2619
Gonon G ..... E31 . 2661
Gonzalez-Rodriguez J ..... E26 . 2591
Gopinath R ..... D23 . 1757
Gopinath R ..... D25 . 1833
Gordos G ..... C25 . 1465
Gordos G E36a 2807
Gordos G ..... B15 .. 517
Gori M ..... B35 .. 889
Gorin A L ..... D15 . 1645
Goronzy S ..... A45 .. 309
Gorrell G ..... D24 . 1779
Goto M ..... B43 . 1013
Gould D ..... E32 . 2675
Gowdy J N ..... C13 . 1181
Gowdy J ..... B35 .. 873
Graff D ..... D36 . 2067
Graham R ..... D42 . 2083
Grainger B J ..... A32 ... 21
Gransden I ..... B12 .. 443
Granser T ..... A45 .. 317
Granström B ..... A46 .. 387
Gravier G ..... E11 . 2293
Green P ..... A41 .. 213
Greenberg S ..... D21 . 1725
Greenberg S ..... D21 . 1729
Greenberg S ..... E21 . 2485
Greenberg S A31 3
Greenberg S ..... B13 .. 473
Greenberg S ..... A35 ... 79
Gretter R ..... C16 . 1327
Gretter R ..... B15 .. 509
Grigat R -R ..... E16 . 2431
Grue J ..... A43 .. 265
Gu L ..... B16 .. 583
Gu Y ..... D23 . 1765
Gu Y ..... E25 . 2565
Guilbaud J-P ..... C16 . 1291
Gunawardana A ..... C14 . 1203
Guo B ..... D45 . 2169
Gurbuz S ..... C13 . 1181
Gurbuz S ..... B35 .. 873
Gussenhoven C ..... C24 . 1403
Gustafson K ..... B15 .. 565
Gustafson-Capkova S ..... B36 .. 931
Gut U ..... A35 ... 95
Gutiérrez J M ..... D45 . 2165
Gutiérrez-Arriola J M ..... A45 .. 357

H

Haase M ..... D45 . 2157
Hacioglu K ..... D24 . 1775
Hacioglu K ..... D45 . 2209
Hagen A ..... A41 .. 225
Hagen A ..... B16 .. 587
Hajic J ..... B14 .. 487

Hakulinen J . . . . . D45 . 2189
Hanna P . . . . . B46 . 1111
Hansen J . . . . . C14 . 1215
Hansen J H L . . . . . D13 . 1573
Hansen J H L . . . . . D36 . 2023
Hansen J H L . . . . . A41 . . 209
Hansen J . . . . . B24 . . 687
Hansen J H L . . . . . B35 . . 905
Harding S . . . . . A36b . 159
Harju M . . . . . E35 . 2729
Harper G . . . . . E15 . 2405
Harper L . . . . . D11 . 1533
Harris M . . . . . B14 . . 499
Haton J-P . . . . . A32 . . . 29
Hawkins S . . . . . A46 . . 407
Hazen T J . . . . . C16 . 1331
Hazen T J . . . . . D14 . 1591
He X . . . . . C25 . 1461
Hecht R . . . . . D36 . 2039
Heckmann M . . . . . B44 . 1023
Heeman P A . . . . . D36 . 2031
Heggtveit P O . . . . . C12 . 1163
Heikkinen A . . . . . D35 . 1965
Heikkinen A . . . . . D35 . 1969
Helen M . . . . . B45 . 1077
Hellwig K . . . . . D25 . 1837
Helme S . . . . . C16 . 1291
Helmuth E . . . . . B44 . 1027
Henrich N . . . . . A33 . . . 47
Henrichsen P J . . . . . B15 . . 553
Herbordt W . . . . . B43 . 1001
Hermansky H . . . . . B11 . . 429
Hermansson H A31 7
Hernández Gómez L A . . . . . D16 . 1695
Hernando J . . . . . B26 . . 779
Herzog M . . . . . A32 . . . 21
Hess W . . . . . C11 . 1149
Hessen, van A . . . . . B45 . 1085
Hetherington I L . . . . . D14 . 1591
Hetherington I L . . . . . D14 . 1599
Heute U . . . . . D16 . 1699
Hickey M . . . . . C16 . 1295
Hidai K-I . . . . . C13 . 1193
Hieronymus J . . . . . D44 . 2143
Higuchi N . . . . . B46 . 1099
Hilger F . . . . . B46 . 1135
Hindle D . . . . . C16 . 1299
Hindle D . . . . . E14 . 2351
Hirano I . . . . . A34 . . . 65
Hirose K . . . . . C16 . 1315
Hirose K . . . . . D46 . 2231
Hirose K . . . . . D46 . 2255
Hirose K . . . . . E16 . 2455
Hiroshige M . . . . . A46 . . 375
Hiroshige M . . . . . B36 . . 935
Hirsch H-G . . . . . D25 . 1837
Hirsch H-G . . . . . A41 . . 184
Hirschberg J . . . . . C12 . 1175
Hirschberg J . . . . . C16 . 1299
Hirschberg J . . . . . E15 . 2373
Hirschberg J . . . . . A45 . . 333
Hirst D . . . . . B45 . 1061
Hirst D . . . . . C11 . 1144
Hitchcock L . . . . . A35 . . . 79
Ho M-S . . . . . D35 . 1985
Hockey B A . . . . . B25 . . 729
Hoffmann R . . . . . C25 . 1433
Hoffmann R . . . . . B15 . . 549
Hogan L A . . . . . B36 . . 955
Holmes W H . . . . . A46 . . 411
Homma S . . . . . B34 . . 859
Honda M . . . . . C26 . 1479
Hone K . . . . . D42 . 2083
Hoory R . . . . . E16 . 2427
Hori C . . . . . D24 . 1771
Hori C . . . . . B14 . . 491
Hori T . . . . . D25 . 1809
Horibe Y . . . . . B34 . . 855
Horiuchi Y . . . . . E33 . 2697
Horne M . . . . . A35 . . 119
Horvat B . . . . . B46 . 1123
Horvat B . . . . . B46 . 1127
Horvat B . . . . . A41 . . 197
Horvat B . . . . . E35 . 2725
Hoshino H . . . . . D42 . 2095
House D . . . . . A46 . . 387
House D . . . . . B15 . . 565
Hualde J I . . . . . A35 . . 115
Huang C . . . . . C23 . 1377
Huang C E36a 2799
Huang C-S . . . . . B26 . . 767
Huang C-S . . . . . B26 . . 771
Huang J . . . . . A44 . . 291
Huang Y . . . . . D45 . 2153
Huang Y . . . . . D45 . 2161
Huber R . . . . . E35 . 2781
Huckvale M . . . . . B32 . . 815
Huerta J . . . . . B21 . . 629
Hung J-W . . . . . D34 . 1959
Huseby M . . . . . A43 . . 265
Hutchinson B . . . . . D16 . 1683
Hwang T-H . . . . . B35 . . 877

I

Ibrahim A . . . . . E33 . 2685
Ibrahim O A G . . . . . B36 . . 951
Ichikawa A . . . . . E33 . 2697
Iida A . . . . . E15 . 2409
Iivonen A K . . . . . A35 . . 103
Imai T . . . . . B25 . . 709
Imai T . . . . . B34 . . 859
Imperl B . . . . . E35 . 2725
Ircing P . . . . . D36 . 2067
Ircing P . . . . . B14 . . 487
Iseli M . . . . . A41 . . 185
Isenhour P . . . . . C16 . 1299
Isenhour P . . . . . E15 . 2373
Ishi C T . . . . . E16 . 2455
Ishida T . . . . . B15 . . 533
Ishimoto Y . . . . . E16 . 2439
Iskra A . . . . . E35 . 2777
Isogai S . . . . . A32 . . . 25
Istrate D . . . . . E15 . 2417
Itakura F . . . . . B46 . 1115
Itakura F . . . . . D36 . 2027
Itoh N . . . . . B25 . . 713
Itoh Y . . . . . D25 . 1805
Itou K . . . . . B43 . 1013
Itou K . . . . . B25 . . 705
Iwabuchi M . . . . . E15 . 2401
Iyengar G . . . . . B44 . 1027

J

Jacob B . . . . . E11 . 2293
Jain P . . . . . B11 . . 429
Jan E-E . . . . . B21 . . 629
Jancovic P . . . . . B46 . 1111
Jancovic P . . . . . B16 . . 579
Jansche M . . . . . B12 . . 459
Janse E . . . . . C24 . 1407
Jarc B . . . . . A41 . . 229
Jarvinen K . . . . . D35 . 1997
Järvinen K . . . . . E12 . 2303
Jasiuk M . . . . . D26 . 1859
Jelinek F . . . . . B14 . . 487
Jelinek M . . . . . D35 . 1997
Jelinek M . . . . . E12 . 2303
Jensen C . . . . . A45 . . 341
Jensen K J . . . . . C25 . 1441
Jia L . . . . . C15 . 1225
Jiang H . . . . . E25 . 2573
Jiang H . . . . . B14 . . 495
Jiang H . . . . . B21 . . 633
Jiang H . . . . . B34 . . 851
Jiang J . . . . . A36b . 179
Jin C . . . . . D43 . 2121
Jin L . . . . . D44 . 2139
Jitapunkul S . . . . . E35 . 2753
Jitsuhiro T . . . . . B25 . . 697
Johansen F T . . . . . E35 . 2725
Johansson J . . . . . E33 . 2685
Johansson V . . . . . A35 . . 119
Johnsen M H . . . . . C15 . 1273
Jokisch O . . . . . B36 . . 947
Jong, de F . . . . . B45 . 1085
Jönsson A . . . . . D45 . 2193
José B. M . . . . . E32 . 2679
Joue G . . . . . B45 . 1073
Joue G . . . . . A45 . . 321
Jouvet D . . . . . A41 . . 201
Jouvet D . . . . . D43 . 2113
Jouvet D . . . . . B11 . . 433
Jung H-K E36b 2829
Jung J . . . . . E15 . 2397
Jung S K . . . . . E22 . 2499
Jung S-K . . . . . D35 . 2017
Junqua J-C . . . . . E14 . 2347
Junqua J-C . . . . . B24 . . 683

K

Kabal P . . . . . D35 . 1993
Kaburagi T . . . . . C26 . 1479
Kacic Z . . . . . B46 . 1123
Kacic Z . . . . . B46 . 1127
Kacic Z . . . . . C26 . 1507
Kacic Z . . . . . A41 . . 197
Kacic Z . . . . . D46 . 2251
Kacic Z . . . . . A42 . . 243
Kacic Z . . . . . E35 . 2725
Kacic Z E36a 2807
Kajarekar S . . . . . B11 . . 429
Kakehi K . . . . . E15 . 2413
Kamm C . . . . . C16 . 1339
Kandel S . . . . . A36b . 163
Kanedera N . . . . . E26 . 2639
Kanevsky D . . . . . D25 . 1833
Kang H-G . . . . . B11 . . 421
Karjalainen M . . . . . D11 . 1537
Karjalainen M . . . . . D46 . 2271
Karjalainen M . . . . . A33 . . . 51
Karlinski D . . . . . A35 . . 127
Karnebäck S . . . . . D26 . 1891
Kashima H . . . . . C16 . 1319
Kassaei M . . . . . B15 . . 513
Kasuriya S . . . . . B26 . . 775
Kasuya H . . . . . D46 . 2267
Kasuya H . . . . . B15 . . 545

Katagiri S . . . . . B34 . . 847
Kato K . . . . . E24 . 2531
Kato T . . . . . B46 . 1099
Katsurada K . . . . . C13 . 1197
Katugampala N . . . . . D35 . 1981
Kawaguchi N . . . . . D36 . 2027
Kawahara H . . . . . E16 . 2451
Kawahara H . . . . . E16 . 2459
Kawahara T . . . . . C16 . 1319
Kawahara T . . . . . D16 . 1691
Kawahara T . . . . . E24 . 2531
Kawamura T . . . . . B46 . 1115
Kawamura T . . . . . E26 . 2603
Kelleher H . . . . . B11 . . 425
Kelleher H . . . . . B11 . . 437
Kellermann W . . . . . B43 . 1001
Kenny P . . . . . C14 . 1207
Kerret O . . . . . A35 . . 127
Keshet J . . . . . D15 . 1637
Kessens J M . . . . . D21 . 1721
Kessens J . . . . . D42 . 2091
Keysers D . . . . . D25 . 1825
Kharroubi J . . . . . D23 . 1761
Khudanpur S . . . . . B14 . . 487
Khudanpur S . . . . . B25 . . 717
Kienappel A K . . . . . D32 . 1911
Kim D-S . . . . . D33 . 1929
Kim H K . . . . . E14 . 2351
Kim H K . . . . . B11 . . 421
Kim H-G . . . . . D26 . 1867
Kim J . . . . . B44 . 1019
Kim J . . . . . D33 . 1929
Kim J-H . . . . . E35 . 2757
Kim K-T . . . . . D35 . 2017
Kim M Y . . . . . D33 . 1929
Kim M H . . . . . E15 . 2397
Kim S . . . . . D46 . 2231
Kim T . . . . . A35 . . 107
Kim W . . . . . A35 . . 107
Kim W . . . . . B25 . . 717
Kim Y . . . . . C14 . 1211
Kim Y-J E36b 2829
King S . . . . . B16 . . 599
Kinnunen T . . . . . E26 . 2627
Kipp M . . . . . C22 . 1367
Kirchhoff K . . . . . B26 . . 803
Kiriyama S . . . . . C16 . 1315
Kiss G . . . . . B15 . . 517
Kitamura T . . . . . D46 . 2263
Kitamura T E36b 2837
Kitamura T . . . . . A46 . . 415
Kitano H . . . . . C13 . 1193
Kitano H . . . . . E26 . 2643
Kitazawa S . . . . . A46 . . 415
Klabbers E . . . . . B15 . . 521
Klakow D . . . . . A34 . . . 69
Kleijn W B . . . . . D35 . 1973
Kloots H . . . . . A35 . . . 75
Kneissler J . . . . . A34 . . . 69
Kneser R . . . . . D32 . 1911
Knight S . . . . . D24 . 1779
Ko H . . . . . A35 . . 107
Kobayashi A . . . . . B25 . . 709
Kobayashi A . . . . . B34 . . 859
Kobayashi S . . . . . C13 . 1197
Kobayashi S E36a 2815
Kobayashi T . . . . . D45 . 2173
Kobayashi T . . . . . D46 . 2263
Kobayashi T . . . . . A45 . . 345
Kobayashi T . . . . . B26 . . 759
Kochanina J . . . . . D36 . 2059
Kochanski G P . . . . . B23 . . 669
Kochanski G . . . . . B36 . . 911
Kodama Y . . . . . E25 . 2549
Koechling C . . . . . B26 . . 779
Koeling R . . . . . D24 . 1779
Kohler M A . . . . . E23 . 2517
Kohlrausch A . . . . . C26 . 1503
Kokkinakis G . . . . . B43 . 1009
Kokkinakis G . . . . . D26 . 1899
Kokkinakis G . . . . . E13 . 2329
Kokkinakis G . . . . . A42 . . 247
Kokkinakis G . . . . . E26 . 2631
Kokkinakis G . . . . . B43 . . 997
Komatani K . . . . . C16 . 1319
Komatsu K . . . . . D44 . 2127
Komatsu M . . . . . A36a . 149
Komatsu M . . . . . A46 . . 391
Kondoz A . . . . . D35 . 1981
Kondoz A M . . . . . E12 . 2307
Koo M-W . . . . . B33 . . 841
Koopmans-van Beinum F J . . C26 . 1471
Korkmazskiy F . . . . . B21 . . 633
Korthauer A . . . . . D13 . 1581
Kotnik B . . . . . B46 . 1123
Kotnik B . . . . . A41 . . 197
Koumpis K . . . . . E15 . 2377
Kouroupetroglou G . . . . . D46 . 2247
Koutras A . . . . . B43 . 1009
Koutras A . . . . . B43 . . 997
Krahmer E . . . . . A46 . . 383
Krauss J . . . . . B46 . 1107
Krauss J . . . . . D26 . 1883
Krbec P . . . . . B14 . . 487
Kriechbaum W . . . . . D45 . 2157
Kristjansson T . . . . . B35 . . 901
Kroschel K . . . . . B44 . 1023
Kruckenberg A . . . . . B23 . . 657
Kruta J . . . . . D36 . 2047
Kryze D . . . . . E14 . 2347
Kuhn R . . . . . B24 . . 683
Kui X . . . . . B45 . 1077
Külekcý M O . . . . . B45 . 1053
Kumar S . . . . . E25 . 2569
Kummert F . . . . . B16 . . 615
Kuo H-K J . . . . . C16 . 1323
Kuo H-K J . . . . . A44 . . 295
Kurimo M . . . . . B25 . . 737
Kuroda Y . . . . . D44 . 2127
Kuroiwa S . . . . . B46 . 1099
Kushner B . . . . . D26 . 1859
Kushner W . . . . . B11 . . 425
Kyung Y . . . . . E15 . 2397

L

Lachiri Z . . . . . E16 . 2471
Lacroix A . . . . . E16 . 2467
Laface P . . . . . E13 . 2325
Lagus K . . . . . B25 . . 737
Lahdili H . . . . . E12 . 2319
Lamel L . . . . . C15 . 1241
Lamel L . . . . . A42 . . 255
Langlois T . . . . . E33 . 2689
Lanzola G . . . . . C16 . 1327
Laprie Y . . . . . A43 . . 277
Laprie Y . . . . . A45 . . 313
Laprie Y . . . . . B13 . . 469
Latsch V . . . . . D16 . 1707
Lau S-L . . . . . A45 . . 325
Lau W . . . . . B42 . . 991
Law K M . . . . . B42 . . 991
Lee A . . . . . C14 . 1219
Lee A . . . . . D15 . 1657
Lee A . . . . . D16 . 1691
Lee A . . . . . D44 . 2127
Lee A . . . . . B35 . . 869
Lee C-H . . . . . C16 . 1323
Lee C-H . . . . . D16 . 1715
Lee C-H . . . . . D45 . 2213
Lee C-H . . . . . E14 . 2355
Lee C-H . . . . . A44 . . 295
Lee C-H . . . . . B12 . . 447
Lee C-H . . . . . B14 . . 495
Lee C-H . . . . . B21 . . 633
Lee J . . . . . B44 . 1019
Lee J . . . . . D26 . 1875
Lee K Y . . . . . D26 . 1875
Lee K-T . . . . . C25 . 1413
Lee L-S . . . . . B45 . 1045
Lee L-S . . . . . C15 . 1269
Lee L-S . . . . . C23 . 1385
Lee L-S . . . . . C25 . 1445
Lee L-S . . . . . D26 . 1895
Lee L-S . . . . . D34 . 1959
Lee L-S . . . . . A44 . . 299
Lee M . . . . . D46 . 2227
Lee R . . . . . E25 . 2545
Lee S . . . . . C16 . 1339
Lee S W . . . . . D35 . 2005
Lee S-M . . . . . C15 . 1269
Lee T . . . . . D12 . 1551
Lee T . . . . . B42 . . 991
Lee W-S . . . . . B22 . . 643
Lee Y . . . . . D26 . 1875
Lee Y . . . . . D46 . 2231
Lefebvre R . . . . . D35 . 1997
Lefebvre R . . . . . E12 . 2303
Lefebvre R . . . . . E12 . 2319
Lefevre F . . . . . C15 . 1241
Lemon O . . . . . D12 . 1559
Lenzo K . . . . . C12 . 1167
Leppänen J . . . . . E35 . 2729
Levin E . . . . . C16 . 1339
Levit M . . . . . D15 . 1645
Lewin I . . . . . D24 . 1779
Lewin I . . . . . E13 . 2333
Lewis E . . . . . D16 . 1703
Li F . . . . . D45 . 2153
Li J . . . . . C12 . 1159
Li J . . . . . D15 . 1617
Li Q . . . . . E32 . 2671
Li Q . . . . . B14 . . 495
Li Q . . . . . B16 . . 619
Li Q . . . . . B21 . . 633
Li S . . . . . C23 . 1377
Li X . . . . . D12 . 1551
Li Y . . . . . B14 . . 503
Li Z E36b 2841
Lieb M . . . . . B21 . . 625
Liljencrants J . . . . . B23 . . 657
Lin F . . . . . C12 . 1159
Lin L L . . . . . A46 . . 411
Lin Y-C . . . . . B45 . 1049
Lin Y-C . . . . . B45 . 1081
Linares G . . . . . C15 . 1245
Lindberg B . . . . . E11 . 2289
Litichever Z . . . . . D25 . 1789
Liu F . . . . . B14 . . 495
Liu F . . . . . B34 . . 851

Liu X . . . . . C15 . 1237
Liu Y . . . . . C25 . 1425
Livescu K . . . . . C25 . 1437
Lleida E . . . . . D13 . 1585
Llitjós A F . . . . . D32 . 1919
Lo W-K . . . . . C16 . 1303
Loehr D . . . . . D11 . 1533
Logan B . . . . . D43 . 2109
Loots S . . . . . D36 . 2039
López-Cózar R . . . . . B25 . . 741
Lotter T . . . . . E26 . 2623
Louloudis D . . . . . E13 . 2329
Lourens T . . . . . E26 . 2643
Louw P H . . . . . D36 . 2055
Lowry D . . . . . B15 . . 513
Lucey S . . . . . C13 . 1185
Lucke H . . . . . E32 . 2667
Luckin M . . . . . B15 . . 513
Ludwig T . . . . . D16 . 1699
Luetgert S . . . . . E16 . 2431
Luk P C . . . . . B45 . 1077
Lukasiak J . . . . . D35 . 1989
Lun D . . . . . B16 . . 611
Lundberg J . . . . . E33 . 2685
Luo C . . . . . D15 . 1617

M

Ma C . . . . . C25 . 1453
Ma J . . . . . B16 . . 603
Macherey K . . . . . D45 . 2205
Macherey W . . . . . D25 . 1825
Macho D . . . . . A41 . . 205
Macho D . . . . . B11 . . 425
Macías-Guarasa J . . . . . C16 . 1279
Macías-Guarasa J . . . . . E25 . 2553
Macon M . . . . . B15 . . 509
Maddieson I . . . . . B32 . . 823
Madievski A . . . . . E25 . 2545
Magimai Doss M . . . . . E35 . 2765
Magno-Caldognetto E . . . . . B44 . 1035
Magrin-Chagnolleau I . . . . . C11 . 1145
Mahé G . . . . . D26 . 1871
Maidment J E36a 2795
Maison B . . . . . D25 . 1833
Majewski W . . . . . D36 . 2059
Mak B . . . . . C15 . 1253
Mak B . . . . . B16 . . 575
Makino S . . . . . E26 . 2595
Makino S . . . . . E26 . 2599
Maltese G . . . . . A32 . . . 21
Mana F . . . . . D32 . 1915
Maneenoi E . . . . . E35 . 2753
Mao J . . . . . D45 . 2169
Martin A . . . . . B26 . . 787
Martin A . . . . . B35 . . 885
Martin R . . . . . E26 . 2623
Martín P . . . . . E25 . 2553
Martínez-Olalla R . . . . . E26 . 2615
Martínez-Olalla R . . . . . E26 . 2619
Masaki S . . . . . A36a . 145
Masgrau E . . . . . D13 . 1585
Mashao D J E36b 2833
Mashimo M . . . . . A45 . . 361
Mason J . . . . . C11 . 1145
Mason J S . . . . . E23 . 2509
Mason J . . . . . E33 . 2693
Mason J S . . . . . B35 . . 893
Massaro D W . . . . . C11 . 1153
Massimino P . . . . . D32 . 1915
Masuda-Katsuse I . . . . . B46 . 1119
Masuko T . . . . . D46 . 2263
Masuko T . . . . . A45 . . 345
Masuko T . . . . . B26 . . 759
Matassoni M . . . . . D13 . 1569
Matousek J . . . . . D36 . 2047
Matousek J . . . . . D36 . 2067
Matousek V . . . . . D45 . 2201
Matrouf D . . . . . C15 . 1245
Matsubara S . . . . . D36 . 2027
Matsui A . . . . . B25 . . 709
Matsui T . . . . . D15 . 1653
Matsumoto H . . . . . B35 . . 881
Matsunaga S . . . . . D25 . 1809
Matsunami K . . . . . C14 . 1219
Matsusaka Y . . . . . D45 . 2173
Mauuary L . . . . . A41 . . 201
Mauuary L . . . . . B11 . . 433
Mauuary L . . . . . B35 . . 885
Mayfield Tomokiyo L . . . . . C25 . 1449
Mazenot S . . . . . C16 . 1291
McCarley J S . . . . . A44 . . 287
McDermott E . . . . . B34 . . 847
McDonough J . . . . . E15 . 2389
McKeown K . . . . . A45 . . 333
Meermeier R . . . . . E34 . 2711
Megyesi B . . . . . B36 . . 931
Meinedo H . . . . . E33 . 2689
Ménard L . . . . . A36b . 163
Ménard L . . . . . A36b . 167
Meng H . . . . . C16 . 1303
Meng H . . . . . D12 . 1551
Meng H . . . . . E35 . 2749
Mengusoglu E . . . . . E25 . 2557
Mera Y . . . . . C14 . 1219
Mera Y . . . . . B35 . . 869
Mercier G . . . . . D43 . 2113
Metze F . . . . . D36 . 2043
Metze F . . . . . E15 . 2389
Meunier J . . . . . D26 . 1859
Meyer C . . . . . B14 . . 499
Meyer G . . . . . A36b . 159
Mihajlik P . . . . . C25 . 1465
Mikkola H . . . . . D35 . 1997
Mikkola H . . . . . E12 . 2303
Mikolaj W . . . . . D46 . 2275
Milone D H . . . . . B25 . . 741
Milward D . . . . . D24 . 1779
Minami Y . . . . . C16 . 1331
Minami Y . . . . . B34 . . 847
Minde T-B A31 7
Minematsu N . . . . . C16 . 1315
Minematsu N . . . . . D46 . 2255
Minematsu N . . . . . E16 . 2455
Minematsu N E36a 2811
Ming J . . . . . B46 . 1111
Ming J . . . . . B16 . . 579
Ming J . . . . . B25 . . 701
Mixdorff H . . . . . A46 . . 403
Mixdorff H . . . . . B36 . . 947
Miyajima C E36b 2837
Mizumachi M . . . . . E26 . 2607
Möbius B . . . . . C11 . 1149
Möbius B . . . . . E11 . 2285
Möbius B . . . . . B23 . . 665
Moen I . . . . . A43 . . 265
Möhler G . . . . . D45 . 2157
Möhler G . . . . . E11 . 2285
Mohri M . . . . . D14 . 1603
Molau S . . . . . E31 . 2653
Möller S . . . . . D16 . 1687
Molyneux D J . . . . . D35 . 1985
Momomura Y . . . . . E26 . 2639
Monaghan A . . . . . B15 . . 513
Montero J M . . . . . C16 . 1279
Montero J M . . . . . D45 . 2165
Montero J M . . . . . A45 . . 357
Montrésor S . . . . . E31 . 2661
Moody M . . . . . D46 . 2239
Moon S Y . . . . . E15 . 2397
Mooshammer C . . . . . C26 . 1487
Moosmüller S . . . . . A45 . . 317
Moraru D . . . . . C16 . 1291
Morel C . . . . . D36 . 2043
Moreno A . . . . . D35 . 2001
Moreno P J . . . . . D43 . 2109
Mori H . . . . . B15 . . 545
Mori K . . . . . A36a . 149
Mori K E36a 2815
Mori S . . . . . B25 . . 713
Moro Q I . . . . . D23 . 1753
Morris A C . . . . . A41 . . 225
Mou X . . . . . B12 . . 451
Mueller A F . . . . . B15 . . 549
Mukai R . . . . . E26 . 2595
Mukai R . . . . . E26 . 2599
Muller L . . . . . D25 . 1813
Muller L . . . . . D36 . 2067
Munhall K . . . . . C26 . 1511
Murahara Y . . . . . A36a . 149
Murahara Y . . . . . E26 . 2639
Murahara Y E36a 2791
Murphy P . . . . . C26 . 1495
Murray I R . . . . . E15 . 2401
Murray I R . . . . . E15 . 2405
Myrvoll T A . . . . . C15 . 1233

N

Nadeu C . . . . . C21 . 1353
Nadeu C . . . . . A41 . . 205
Nagatomo K . . . . . D44 . 2127
Naito M . . . . . B46 . 1099
Najaf-Zadeh H . . . . . D35 . 1993
Najim M . . . . . B15 . . 541
Nakadai K . . . . . C13 . 1193
Nakadai K . . . . . E26 . 2643
Nakagawa S . . . . . E25 . 2549
Nakagawa S E36a 2811
Nakagawa S E36a 2815
Nakagawa S . . . . . B34 . . 855
Nakai M . . . . . D25 . 1841
Nakajima H . . . . . A34 . . . 65
Nakamura K . . . . . E15 . 2401
Nakamura N E36a 2811
Nakamura S . . . . . B46 . 1139
Nakamura S . . . . . D15 . 1653
Nakamura S . . . . . D15 . 1661
Nakamura S . . . . . A41 . . 233
Nakamura S . . . . . E26 . 2607
Nakamura S . . . . . E26 . 2611
Nakamura S . . . . . B16 . . 571
Nakamura Y . . . . . C13 . 1197
Nakano M . . . . . C16 . 1331
Nanjo H . . . . . E24 . 2531
Narayanan S . . . . . D25 . 1845
Narayanan S . . . . . E32 . 2675
Nasr A . . . . . B25 . . 725
Natvig J E . . . . . C12 . 1163
Navratil J . . . . . C23 . 1389

Navratil J . . . . . D23 . 1757
Nefti S . . . . . D16 . 1711
Németh G . . . . . D36 . 2035
Németh G . . . . . B15 . . 517
Németh G . . . . . B15 . . 525
Neti C . . . . . B44 . 1027
Neto J . . . . . E33 . 2689
Neuvo Y . . . . . A31 . . . 13
Neuvo Y A21 7
Ney H . . . . . B46 . 1135
Ney H . . . . . C25 . 1429
Ney H . . . . . D25 . 1825
Ney H . . . . . D45 . 2205
Ney H . . . . . E31 . 2653
Nguyen N . . . . . A46 . . 407
Nguyen P . . . . . E14 . 2347
Nguyen P . . . . . B24 . . 683
Nielsen C . . . . . B42 . . 975
Niemann H . . . . . E35 . 2745
Niemann H . . . . . E35 . 2781
Nieto-Lluís V . . . . . E26 . 2615
Nieto-Lluís V . . . . . E26 . 2619
Nigra M . . . . . E13 . 2325
Niimi Y . . . . . C16 . 1283
Niimi Y . . . . . D22 . 1743
Niimi Y . . . . . D45 . 2185
Niiniluoto I . . . . . A31 . . . 11
Niklfeld G . . . . . E13 . 2341
Nils K . . . . . C22 . 1363
Niranjan M . . . . . E15 . 2377
Nishida M . . . . . C23 . 1381
Nishide R . . . . . E16 . 2455
Nishimoto T . . . . . D22 . 1743
Nishimoto T . . . . . D45 . 2185
Nishimura M . . . . . B25 . . 713
Nishiura T . . . . . E26 . 2611
Nishizaki H . . . . . E25 . 2549
Nisimura R . . . . . D44 . 2127
Nitta T . . . . . C13 . 1197
Nocera P . . . . . C15 . 1245
Noda Y . . . . . D25 . 1809
Noe B . . . . . A41 . . 201
Noé B . . . . . B11 . . 433
Nokas G . . . . . E26 . 2631
Noma K-I . . . . . D25 . 1841
Nordgård T . . . . . A35 . . 111
Nöth E . . . . . E11 . 2285
Nöth E . . . . . E35 . 2745
Nöth E . . . . . E35 . 2781
Nouza J . . . . . C16 . 1287
Nouza T . . . . . C16 . 1287
Nurminen J . . . . . D35 . 1969
Ó'Cróinín D . . . . . C21 . 1353

O

O'Shaughnessy E36b 2825
O'Shaughnessy D . . . . . D13 . 1577
Obermayer K . . . . . D26 . 1867
Och F J . . . . . D45 . 2205
Ogata J . . . . . E25 . 2577
Ogner M . . . . . C26 . 1507
Ohno S . . . . . B23 . . 661
Ohtsuka T . . . . . D46 . 2267
Okada K . . . . . E26 . 2639
Okada M . . . . . E15 . 2413
Oku T . . . . . D45 . 2185
Okuda K . . . . . D15 . 1653
Okuno H G . . . . . C13 . 1193
Okuno H G . . . . . E26 . 2643
Olaszi P . . . . . B15 . . 525
Olaszy G . . . . . B15 . . 517
Olaszy G . . . . . B15 . . 525
Olive J P . . . . . D45 . 2213
Olivier S . . . . . B16 . . 619
Olsen P . . . . . D25 . 1833
Omologo M . . . . . D13 . 1569
Omote M . . . . . E32 . 2667
Onishi S . . . . . B25 . . 693
Ono T . . . . . D22 . 1743
Onoe K . . . . . B14 . . 495
Oppermann D . . . . . D45 . 2197
Ordelman R . . . . . B45 . 1085
Orlandi M . . . . . C16 . 1327
Ortega A . . . . . D13 . 1585
Ortega A . . . . . D25 . 1845
Ortega-García J . . . . . D23 . 1753
Ostendorf M . . . . . D43 . 2117
Ostendorf M . . . . . E34 . 2715
Ostendorf M . . . . . B42 . . 987
Öster A-M E36a 2807
Otake T . . . . . A36a . 141
Otake T . . . . . B13 . . 465
Otterson S . . . . . E34 . 2715
Ou Z . . . . . B26 . . 791
Ouden, den H . . . . . A35 . . . 91
Ouellet P . . . . . D14 . 1595
Ouni K . . . . . E16 . 2471
Ouni S . . . . . A43 . . 277
Ozeki K . . . . . B45 . 1041
Ozeki K E36b 2849
Özkan M . . . . . B45 . 1053

P

Paatero T . . . . . D46 . 2271
Paavo A . . . . . A33 . . . 51
Pabon P . . . . . A33 . . . 39
Pacchiotti A . . . . . D32 . 1915
Padmanabhan M . . . . . E14 . 2359
Padmanabhan M . . . . . A44 . . 291
Paliwal K K . . . . . B46 . 1139
Paliwal K K . . . . . C15 . 1233
Paliwal K K . . . . . A41 . . 233
Paliwal K K . . . . . B16 . . 571
Paliwal K . . . . . B16 . . 591
Paliwal K . . . . . B26 . . 755
Palmer D . . . . . D43 . 2117
Palou F . . . . . A32 . . . 21
Pan S . . . . . A45 . . 333
Parandekar S . . . . . B26 . . 803
Pardo J M . . . . . C16 . 1279
Pardo J M . . . . . D45 . 2165
Pardo J M . . . . . E25 . 2553
Pardo J M . . . . . A45 . . 357
Pargellis A N . . . . . D45 . 2213
Pargellis A . . . . . B12 . . 447
Park A . . . . . D14 . 1591
Park S-W . . . . . E22 . 2491
Park Y C . . . . . E22 . 2499
Park Y-C . . . . . D35 . 2017
Park Y-C . . . . . E22 . 2491
Parthasarathy S . . . . . E15 . 2373
Pastor-i-Gadea M . . . . . E11 . 2297
Pastor-i-Gadea M . . . . . E15 . 2385
Patterson E K . . . . . C13 . 1181
Patterson E . . . . . B35 . . 873
Pearce D . . . . . A41 . . 184
Pearce D . . . . . B11 . . 425
Pearce D . . . . . B11 . . 437
Pedersen M W . . . . . C25 . 1441
Peinado A M . . . . . A41 . . 221
Peinado A M . . . . . E22 . 2503
Peinado A M . . . . . E34 . 2707
Pelachaud C . . . . . B44 . 1035
Pelaez-Moreno C . . . . . B46 . 1103
Pellegrino F . . . . . C11 . 1145
Pellegrino F . . . . . E24 . 2539
Pellom B . . . . . D36 . 2023
Pellom B . . . . . A41 . . 209
Pellom B . . . . . D45 . 2209
Pellom B . . . . . B35 . . 905
Peng H . . . . . D42 . 2087
Perez-Cordoba J L . . . . . E22 . 2503
Perez-Cordoba J L . . . . . E34 . 2707
Perrier P . . . . . C26 . 1487
Perronnin F . . . . . B24 . . 683
Petek B . . . . . C21 . 1353
Petek B . . . . . E35 . 2777
Peters S . . . . . D12 . 1559
Petersen N R . . . . . B36 . . 939
Petrinovic D . . . . . E16 . 2479
Petrovska-Delacretaz D . . . . . D23 . 1761
Petrovsky A . . . . . E26 . 2623
Pfannerer N . . . . . D36 . 2039
Pfau T . . . . . E24 . 2535
Picheny M . . . . . B14 . . 503
Pietilä S . . . . . D35 . 1965
Pitermann M . . . . . C26 . 1511
Pitz M . . . . . E31 . 2653
Platt J C . . . . . D26 . 1903
Plucienkowski J P . . . . . D13 . 1573
Plucienkowski J . . . . . D36 . 2023
Pobloth H . . . . . D35 . 1973
Podhorski A . . . . . C26 . 1519
Pokrovsky A . . . . . C16 . 1339
Polifroni J . . . . . C16 . 1331
Pollak P . . . . . D36 . 2059
Pols L C W . . . . . D36 . 2051
Pols L C W . . . . . A43 . . 269
Pols L C W A21 3
Pols L C W . . . . . B32 . . 811
Popovici C . . . . . E13 . 2325
Potamianos A . . . . . D22 . 1735
Potamianos A . . . . . D45 . 2217
Potamianos A . . . . . B12 . . 447
Potamianos G . . . . . B44 . 1027
Potamitis I . . . . . D26 . 1899
Potamitis I . . . . . E16 . 2475
Printz H . . . . . D25 . 1833
Printz H . . . . . A42 . . 251
Przybocki M . . . . . B26 . . 787
Psutka J . . . . . D25 . 1813
Psutka J V . . . . . D25 . 1813
Psutka J . . . . . D36 . 2047
Psutka J . . . . . D36 . 2067
Psutka J . . . . . D46 . 2259
Psutka J . . . . . B14 . . 487
Pucher M . . . . . E13 . 2341
Pujalte S . . . . . D35 . 2001
Purnell D W . . . . . C15 . 1249
Pusateri E . . . . . D25 . 1817

R

Radova V . . . . . D36 . 2067
Ragot S . . . . . E12 . 2319
Rahim M . . . . . C16 . 1339
Raj B . . . . . B43 . 1005
Raj B . . . . . D43 . 2109

Raj B ..... A32 ... 33
Raj B ..... B25 .. 733
Rajouani A ..... B15 .. 541
Ramabadran T ..... D26 . 1859
Ramabadran T ..... B11 .. 425
Ramaswamy G N ..... C23 . 1389
Rambow O ..... C12 . 1175
Rambow O ..... D22 . 1747
Ramos J M ..... D45 . 2165
Ramsay G ..... A33 ... 43
Randolph M ..... C25 . 1453
Rayner M ..... D24 . 1779
Rayner M ..... B25 .. 729
Reininger H ..... B26 .. 783
Reithinger N ..... D12 . 1547
Renals S ..... E15 . 2377
Renevey P ..... B46 . 1107
Renevey P ..... D26 . 1883
Renevey P ..... D26 . 1887
Reyes Gomez M J ..... A41 .. 189
Riedler J ..... D36 . 2039
Riele, te S ..... D45 . 2177
Rietveld T ..... C24 . 1399
Rietveld T ..... C24 . 1403
Rigazio L ..... E14 . 2347
Rigoll G ..... C15 . 1229
Rigoll G ..... E34 . 2711
Riis S K ..... C25 . 1441
Riley M ..... D14 . 1603
Ris C ..... E25 . 2557
Ris C ..... E26 . 2635
Riskin E ..... E34 . 2715
Ritz C H ..... D35 . 1989
Ritz C H ..... E12 . 2315
Roach M ..... E33 . 2693
Roach P E36a 2807
Rodellar-Biarge V ..... E26 . 2615
Rodellar-Biarge V ..... E26 . 2619
Rodriguez L J ..... D15 . 1665
Rodríguez García J M ..... E15 . 2393
Rodríguez-Saeta J ..... B26 .. 779
Rogati M ..... D22 . 1747
Rojc M ..... D46 . 2251
Rolland G ..... B36 .. 963
Rose G ..... D26 . 1851
Rose K ..... B16 .. 583
Rose R C ..... E14 . 2351
Rose R C ..... B11 .. 421
Rosenberg A ..... C16 . 1299
Rosenberg A ..... E15 . 2373
Rosenfeld R ..... D12 . 1563
Rothkrantz L J M ..... E16 . 2463
Rotola-Pukkila J ..... D35 . 1997
Rotola-Pukkila J ..... E12 . 2303
Rouat J ..... D33 . 1937
Rouat J E36b 2825
Roux J C ..... D36 . 2055
Rubio A J ..... D15 . 1625
Rubio A J ..... E22 . 2503
Rudnicky A ..... D12 . 1563
Rudnicky A ..... D25 . 1829
Rudnicky A ..... D43 . 2105
Rudnicky A ..... D43 . 2121
Ruoppila V T ..... D35 . 1965
Ruscitti P ..... C16 . 1339
Ruske G ..... E24 . 2535
Ruske G ..... B26 .. 751
Rusko M ..... D36 . 2059
Russell M ..... E32 . 2671
Ruwisch D ..... D26 . 1867

S

Saarinen J ..... D35 . 1969
Saarinen J ..... E35 . 2729
Sääv J ..... E35 . 2737
Sadowski J ..... D36 . 2059
Sagayama S ..... D25 . 1841
Sagerer G ..... E24 . 2527
Sagisaka Y ..... D15 . 1661
Sagisaka Y ..... A32 ... 25
Sagisaka Y ..... A34 ... 65
Sagisaka Y ..... B25 .. 693
Sagisaka Y ..... B25 .. 697
Sahakyan M ..... A45 .. 309
Saito T ..... C12 . 1171
Sakamoto M ..... C12 . 1171
Sakurada Y ..... E15 . 2409
Sakurai A ..... D46 . 2255
Salami R ..... D35 . 1997
Salami R ..... E12 . 2303
Salmela P ..... E35 . 2729
Samuelsson C ..... D44 . 2143
San-Segundo R ..... C16 . 1279
San-Segundo R ..... D45 . 2165
San-Segundo R ..... E25 . 2553
San-Segundo R ..... A45 .. 357
Sanchez V ..... E34 . 2707
Sanchez-Bote J-L ..... E26 . 2591
Sanchis A ..... E15 . 2385
Sanderson C ..... B26 .. 755
Sandri S ..... D46 . 2243
Sannier F ..... B15 .. 513
Saon G ..... B21 .. 629
Sarasola K ..... C21 . 1353
Sarikaya R ..... A41 .. 209
Sarikaya R ..... B24 .. 687
Sarikaya R ..... B35 .. 905
Saruwatari H ..... D44 . 2127
Saruwatari H ..... E26 . 2595
Saruwatari H ..... E26 . 2603
Saruwatari H ..... A45 .. 349
Saruwatari H ..... B35 .. 869
Sasajima M ..... C16 . 1311
Sasou A ..... E16 . 2443
Satheesh S ..... D35 . 2009
Sato S ..... B34 .. 859
Satoh T ..... B26 .. 759
Savariaux C ..... B36 .. 963
Savino M ..... B36 .. 943
Schaaf T ..... E25 . 2581
Schaeffler F ..... B32 .. 819
Schalk H ..... B26 .. 783
Scharenborg O ..... E15 . 2381
Scheffler K ..... A44 .. 283
Schiefer C ..... D36 . 2039
Schiel F ..... D45 . 2197
Schlosser K ..... C16 . 1339
Schlüter R ..... E31 . 2653
Schnell K ..... E16 . 2467
Schoentgen J ..... C26 . 1499
Schone P ..... C16 . 1303
Schramm H ..... C25 . 1457
Schramm H ..... B14 .. 499
Schröder M ..... B15 .. 561
Schröder M ..... A35 ... 87
Schultz T ..... E35 . 2721
Schwartz J-L ..... A36b . 163
Schwarz J ..... D45 . 2201
Schweitzer A ..... E11 . 2285
Scordilis M ..... D24 . 1783
Scott K R ..... B26 .. 747
Sedivy J ..... D25 . 1833
Segi H ..... B25 .. 709
Segura J C ..... D15 . 1625
Segura J C ..... A41 .. 221
Segura J C ..... E34 . 2707
Selouani S-A ..... D13 . 1577
Seltzer M L ..... B43 . 1005
Seneff S ..... C16 . 1331
Seneff S ..... E35 . 2761
Seneff S ..... A45 .. 353
Seneff S ..... B12 .. 451
Sepesy Maucec M ..... A42 .. 243
Seward A ..... D14 . 1607
Sfakianaki A E36a 2807
Shahshahani B ..... E13 . 2337
Sharoff S ..... D36 . 2063
Shdaifat I ..... E16 . 2431
Sheikhzadeh H ..... D26 . 1855
Shen X ..... A35 .. 123
Shen X ..... B25 .. 721
Shi Y E36a 2799
Shih C ..... B23 .. 669
Shih C ..... B36 .. 911
Shikano K ..... C14 . 1219
Shikano K ..... D15 . 1657
Shikano K ..... D16 . 1691
Shikano K ..... D44 . 2127
Shikano K ..... E26 . 2603
Shikano K ..... E26 . 2611
Shikano K ..... A45 .. 349
Shikano K ..... A45 .. 361
Shikano K ..... B35 .. 869
Shimizu A ..... B35 .. 881
Shimizu T ..... B46 . 1099
Shimodaira H ..... D25 . 1841
Shimomori T ..... C16 . 1311
Shin V ..... D33 . 1929
Shinoda K ..... D15 . 1649
Shinozaki T ..... B14 .. 491
Shirai K ..... D26 . 1875
Shirai K ..... A32 ... 25
Shirai K ..... A34 ... 65
Shire M L ..... D25 . 1797
Shriberg E ..... C22 . 1359
Shriver S ..... D12 . 1563
Sicilia-Garcia E I ..... B25 .. 701
Sienel J ..... A41 .. 201
Sienel J ..... B11 .. 433
Siivola V ..... B25 .. 737
Silva J ..... C15 . 1257
Silverman K ..... C12 . 1167
Silverman K ..... B12 .. 455
Simon-Zorita D ..... E26 . 2591
Simonsen H G ..... A43 .. 265
Simpson B D ..... B26 .. 747
Siohan O ..... E14 . 2355
Siohan O ..... B14 .. 495
Siohan O ..... B21 .. 633
Siohan O ..... B34 .. 851
Siu K-C ..... E35 . 2749
Siu M ..... B46 . 1095
Sivadas S ..... B11 .. 429
Sivakumaran P ..... B26 .. 795
Smaili K ..... A32 ... 29
Smeele P ..... D16 . 1675
Smith C L ..... B36 .. 955
Smith F J ..... B25 .. 701
Smits R ..... B13 .. 481
Sohn S M ..... E15 . 2397
Soltau H ..... E15 . 2389

Song H-E . . . . . B33 . . 841
Song Z . . . . . A34 . . . 57
Soong F . . . . . E25 . 2573
Soong F K . . . . . B14 . . 495
Soong F K . . . . . B16 . . 619
Soong F K . . . . . B21 . . 633
Soquet A . . . . . B22 . . 647
Sornlertlamvanich V . . . . . B45 . 1057
South A J . . . . . C26 . 1515
Souto N . . . . . E33 . 2689
Sovka P . . . . . D26 . 1863
Sreenivas T V . . . . . D35 . 2009
Sridharan S . . . . . C13 . 1185
Sridharan S . . . . . D46 . 2239
Srinivasamurthy N . . . . . D25 . 1845
St John Brittan P . . . . . C16 . 1295
Stadermann J . . . . . D26 . 1851
Stadermann J . . . . . E34 . 2711
Stahl V . . . . . D26 . 1851
Stapert R P . . . . . E23 . 2509
Stark L . . . . . C16 . 1299
Staroniewicz P . . . . . D36 . 2059
Stead L . . . . . C16 . 1299
Stead L . . . . . E15 . 2373
Steeneken H . . . . . D16 . 1675
Steeneken H . . . . . E22 . 2495
Steininger S . . . . . D45 . 2197
Stemmer G . . . . . E35 . 2745
Stenzel G . . . . . D45 . 2157
Stephenson T A . . . . . E35 . 2765
Stewart D . . . . . B46 . 1111
Stöber K-H . . . . . B15 . . 521
Stokes S . . . . . A46 . . 395
Stolcke A . . . . . C22 . 1359
Stolcke A . . . . . B24 . . 679
Streefkerk B M . . . . . B32 . . 811
Strik H . . . . . D21 . 1721
Strik H . . . . . D42 . 2091
Strömqvist S . . . . . A35 . . 119
Strong G . . . . . A31 . . . 15
Sturm J . . . . . E15 . 2381
Sturm J . . . . . E25 . 2585
Stuttle M . . . . . B24 . . 675
Su Y . . . . . D45 . 2153
Su Y . . . . . D45 . 2161
Sull W H . . . . . E15 . 2397
Sullivan K P H . . . . . A36a . 133
Sun X . . . . . B15 . . 537
Surendran A C . . . . . B21 . . 633
Surendran A . . . . . B26 . . 763
Suzuki N . . . . . E15 . 2413
Svaizer P . . . . . D13 . 1569
Svendsen T . . . . . C15 . 1233
Swerts M . . . . . A46 . . 383
Swerts M . . . . . A35 . . . 75
Syrdal A . . . . . C16 . 1339
Syrdal A K . . . . . B42 . . 979

T

Tabain M . . . . . B36 . . 963
Tajima K . . . . . A36a . 145
Takagi K . . . . . B45 . 1041
Takagi K E36b 2849
Takamaru K . . . . . B36 . . 935
Takeda K . . . . . B46 . 1115
Takeda K . . . . . D36 . 2027
Takeuchi Y . . . . . E15 . 2413
Tam Y-C . . . . . B16 . . 575
Tambouratzis G . . . . . D36 . 2075
Tambouratzis G . . . . . B33 . . 837
Tamura M . . . . . A45 . . 345
Tan B T . . . . . E25 . 2565
Tanaka K . . . . . C16 . 1319
Tanaka K . . . . . D25 . 1805
Tanaka K . . . . . E16 . 2443
Tanaka T E36a 2815
Tang M . . . . . A45 . . 353
Tarsaku P . . . . . B45 . 1057
Tatai P . . . . . C25 . 1465
Tatham M . . . . . D16 . 1703
Teixeira A J D S . . . . . C26 . 1483
Teixeira J P . . . . . D16 . 1707
ten Bosch L . . . . . C25 . 1421
ten Bosch L F M . . . . . B32 . . 811
Terashima R . . . . . D42 . 2095
Terken J . . . . . D45 . 2177
Terken J M B . . . . . B15 . . 529
Terken J . . . . . A35 . . . 91
Terken J M B . . . . . B36 . . 959
Tesser F . . . . . B15 . . 509
Thampanitchawong B . . . . . E35 . 2753
Thathong U . . . . . E35 . 2753
Theunissen M W . . . . . A44 . . 283
Thomas T . . . . . D23 . 1765
Thomas T . . . . . E25 . 2565
Thongprasirt R . . . . . B45 . 1057
Thunberg G C . . . . . A46 . . 399
Tihelka J . . . . . D26 . 1863
Tochinai K . . . . . A46 . . 375
Tochinai K . . . . . B36 . . 935
Toda T . . . . . A45 . . 349
Toda T . . . . . A45 . . 361
Tokuda K . . . . . D46 . 2263
Tokuda K E36b 2837
Tokuda K . . . . . A45 . . 345
Tokuda K . . . . . B26 . . 759
Tokuma S . . . . . A46 . . 391
Tokuma W . . . . . A46 . . 379
Tokuma W . . . . . A46 . . 391
Tolba H . . . . . D13 . 1577
Torre A D L . . . . . A41 . . 221
Torre Toledano D . . . . . D16 . 1695
Torres I . . . . . D15 . 1665
Torres I . . . . . D44 . 2135
Toth A . . . . . D12 . 1563
Trancoso I . . . . . D44 . 2131
Trancoso I . . . . . E33 . 2689
Traunmüller H . . . . . A46 . . 399
Trentin E . . . . . B35 . . 889
Trippel T . . . . . D36 . 2063
Tropf H S . . . . . D36 . 2059
Trouvain J . . . . . B15 . . 557
Tsai A . . . . . D45 . 2213
Tsai M-Y . . . . . C25 . 1445
Tsai W-H . . . . . C23 . 1385
Tsai W-H . . . . . B26 . . 767
Tsai W-H . . . . . B26 . . 771
Tsao Y . . . . . C15 . 1269
Tseng S-C . . . . . D45 . 2181
Tsopanoglou A . . . . . E13 . 2329
Tsui W C . . . . . D12 . 1551
Tsukada K . . . . . A45 . . 305
Tsuzaki M . . . . . D46 . 2223
Tufekci Z . . . . . C13 . 1181
Tufekci Z . . . . . B35 . . 873
Türk U . . . . . D11 . 1541
Turunen J J . . . . . E14 . 2363
Turunen M . . . . . D45 . 2189
Tychtl Z . . . . . D46 . 2259
Tzur M . . . . . E16 . 2427

U

Ueda K . . . . . D22 . 1743
Uehara T . . . . . C16 . 1311
Unoki M . . . . . E16 . 2439
Usuki N E36a 2791
Utsuro T . . . . . E25 . 2549

V

Vaich T . . . . . D25 . 1793
Vainio J . . . . . D35 . 1997
Vainio J . . . . . E12 . 2303
Vainio M . . . . . D11 . 1537
Vair C . . . . . E13 . 2325
Vallée N . . . . . A36b . 163
Vallejo J A . . . . . A45 . . 357
van Santen J . . . . . C11 . 1149
van Wijck M . . . . . A33 . . . 39
van Herwijnen O M . . . . . B15 . . 529
van Herwijnen O M . . . . . B36 . . 959
van Dinther R . . . . . C26 . 1503
van Hout R . . . . . B13 . . 481
Van Hirtum A . . . . . E16 . 2435
Van Compernolle D . . . . . D15 . 1621
Van den Heuvel H . . . . . D36 . 2051
Van den Heuvel H . . . . . D36 . 2059
Van den Dikkenberg-Pot I . . . C26 . 1471
Van Thong J-M . . . . . D25 . 1817
Van Son R J J H . . . . . D36 . 2051
Van Wijngaarden S . . . . . D16 . 1675
Van Wijngaarden S . . . . . E22 . 2495
Várkonyi-Kóczy A R . . . . . E12 . 2311
Varona A . . . . . D15 . 1665
Varona A . . . . . D44 . 2135
Vasilache M . . . . . C15 . 1265
Vasilescu I . . . . . B45 . 1065
Vaufreydaz D . . . . . C16 . 1291
Vaz F . . . . . C26 . 1483
Veldhuis R . . . . . C26 . 1503
Veldhuis R . . . . . B15 . . 521
Vermillion P . . . . . C24 . 1399
Veronik R E36a 2807
Vetter R . . . . . B46 . 1107
Vetter R . . . . . D26 . 1883
Vicsi K E36a 2807
Vidal E . . . . . E15 . 2385
Viikki O . . . . . C15 . 1265
Viikki O . . . . . E35 . 2729
Vilkman E . . . . . B36 . . 919
Villar M E36b 2845
Vinod C . . . . . C13 . 1185
Vintturi J . . . . . B36 . . 919
Visweswariah K . . . . . A42 . . 251
Vivaracho C E . . . . . D23 . 1753
Vlaj D . . . . . B46 . 1127
Vlaj D . . . . . A41 . . 193
Vlaj D . . . . . E14 . 2363
Vosnidis C . . . . . B33 . . 833

W

Wagner P . . . . . B15 . . 521
Wahlster W . . . . . D12 . 1547
Waibel A . . . . . E35 . 2721
Wakita T . . . . . D42 . 2095
Walker M . . . . . C16 . 1339
Walker M . . . . . C22 . 1371

Walker M . . . . . D22 . 1747
Wallhoff F . . . . . C15 . 1229
Walsh M . . . . . E11 . 2281
Wambacq P . . . . . D15 . 1621
Wan W . . . . . B16 . . 611
Wang C . . . . . E35 . 2761
Wang C . . . . . A45 . . 353
Wang F . . . . . D34 . 1947
Wang H-C . . . . . E31 . 2657
Wang H-C . . . . . B35 . . 877
Wang H-M . . . . . B45 . 1045
Wang H-M . . . . . B45 . 1049
Wang H-M . . . . . D34 . 1959
Wang H-M . . . . . A44 . . 299
Wang L . . . . . D12 . 1551
Wang N . . . . . C23 . 1385
Wang W-J . . . . . E35 . 2773
Wang Y-Y . . . . . D12 . 1555
Wang Z . . . . . D15 . 1629
Wang Z . . . . . B26 . . 791
Warakagoda N D . . . . . C15 . 1273
Ward T . . . . . A44 . . 287
Ward W . . . . . D24 . 1775
Ward W . . . . . D36 . 2023
Ward W . . . . . D45 . 2209
Warnke V . . . . . E35 . 2781
Weber K . . . . . B16 . . 607
Wei Y . . . . . B11 . . 425
Weinstein E . . . . . C16 . 1335
Wellekens C J . . . . . C25 . 1413
Wendt S . . . . . B16 . . 615
Wester M . . . . . D21 . 1725
Wester M . . . . . D21 . 1729
Westerdijk M . . . . . A35 . . . 87
Whittaker E . . . . . A32 . . . 33
Whittaker E . . . . . B25 . . 733
Whittaker S . . . . . C16 . 1299
Widera C . . . . . A46 . . 403
Wild T . . . . . B44 . 1023
Willett D . . . . . C15 . 1229
Willett D . . . . . B34 . . 847
Williams B . . . . . C21 . 1353
Wilson D . . . . . D43 . 2121
Wilson I . . . . . A43 . . 273
Wittenburg P . . . . . D11 . 1529
Wojdel J C . . . . . E16 . 2463
Wokurek W . . . . . A45 . . 309
Wolff M . . . . . C25 . 1433
Wong K-M . . . . . C15 . 1253
Wong Y W . . . . . D12 . 1551
Wong Y F . . . . . D12 . 1551
Wong Y W . . . . . E35 . 2741
Woodland P C . . . . . E35 . 2757
Wrede B . . . . . E24 . 2527
Wright J H . . . . . D15 . 1645
Wu C-H . . . . . D34 . 1955
Wu G . . . . . D44 . 2139
Wu J . . . . . C15 . 1261
Wu J . . . . . B25 . . 717
Wu T Y . . . . . D12 . 1551
Wu W . . . . . D25 . 1801
Wu W . . . . . D34 . 1947
Wu W . . . . . D44 . 2139
Wu W . . . . . D45 . 2153
Wu Y E36b 2841
Wutiwiwatchai C . . . . . B26 . . 775

X

Xu B . . . . . C15 . 1225
Xu B . . . . . A35 . . 123
Xu B . . . . . B25 . . 721
Xu D . . . . . B15 . . 545
Xu M . . . . . D45 . 2149
Xu Y . . . . . C16 . 1283
Xydas G . . . . . D46 . 2247

Y

Yablonsky S . . . . . D36 . 2071
Yamada H . . . . . C13 . 1197
Yamada M . . . . . C14 . 1219
Yamada M . . . . . D15 . 1657
Yamada M . . . . . B35 . . 869
Yamada S . . . . . B25 . . 697
Yamaguchi Y . . . . . A36a . 141
Yamamoto H . . . . . A32 . . . 25
Yamamoto H . . . . . B25 . . 693
Yamamoto H . . . . . B25 . . 697
Yamamoto K . . . . . B35 . . 881
Yamashita Y . . . . . B15 . . 533
Yan G-L . . . . . D34 . 1955
Yan P . . . . . D45 . 2149
Yan Y . . . . . C15 . 1237
Yang B . . . . . A36a . 137
Yano T . . . . . C16 . 1311
Yao K . . . . . B46 . 1139
Yao K . . . . . A41 . . 233
Yapanel U . . . . . D36 . 2023
Yapanel U . . . . . A41 . . 209
Yapanel U . . . . . B35 . . 905
Yasumura M . . . . . E15 . 2409
Yoma N B . . . . . C15 . 1257
Yoma N B E36b 2845
Yoo C D . . . . . D33 . 1941
Yoon S . . . . . D33 . 1941
Yoon S W . . . . . E22 . 2499
Yoon S-W . . . . . D35 . 2017
Yoshida K E36b 2849
Yoshimura T . . . . . D46 . 2263
Yoshizawa S . . . . . C14 . 1219
Yoshizawa S . . . . . D15 . 1657
Yoshizawa S . . . . . B35 . . 869
Youn D H . . . . . E22 . 2499
Youn D-H . . . . . D35 . 2017
Youn D-H . . . . . E16 . 2447
Youn D-H . . . . . E22 . 2491
Yu A-T . . . . . E31 . 2657
Yu P . . . . . D15 . 1629
Yu X . . . . . D45 . 2209
Yu X . . . . . B16 . . 611
Yuan B . . . . . C15 . 1237
Yuo K-H . . . . . B35 . . 877
Yvon F . . . . . E11 . 2293

Z

Zainkó C . . . . . D36 . 2035
Zaki A . . . . . B15 . . 541
Zamchick G . . . . . C16 . 1299
Zee E . . . . . B22 . . 643
Zgank A . . . . . E35 . 2725
Zhang B . . . . . C12 . 1159
Zhang B . . . . . D45 . 2169
Zhang G . . . . . D15 . 1617
Zhang G . . . . . D25 . 1801
Zhang J . . . . . D15 . 1617
Zhang J . . . . . D45 . 2209
Zhang J-S . . . . . D15 . 1661
Zhang L . . . . . B16 . . 595
Zhang R . . . . . D43 . 2105
Zhang R . . . . . D43 . 2121
Zhang S-W . . . . . D15 . 1661
Zhang Y . . . . . E25 . 2545
Zhao Y . . . . . C25 . 1461
Zheng F . . . . . D15 . 1617
Zheng F . . . . . D25 . 1801
Zheng F . . . . . D34 . 1947
Zheng F . . . . . D44 . 2139
Zheng F . . . . . D45 . 2149
Zheng F . . . . . D45 . 2153
Zheng F . . . . . D45 . 2161
Zheng F . . . . . A34 . . . 57
Zheng J . . . . . D16 . 1715
Zheng J . . . . . B24 . . 679
Zhou B . . . . . C14 . 1215
Zhou J . . . . . C23 . 1377
Zhou J . . . . . D34 . 1951
Zhou J E36a 2799
Zhou Q . . . . . D16 . 1715
Zhou Q . . . . . B14 . . 495
Zhu Q . . . . . A41 . . 185
Zhu W-J . . . . . A44 . . 287
Zhu X . . . . . D12 . 1563
Ziegenhain U . . . . . C25 . 1417
Zitouni I . . . . . A32 . . . 29
Zmarich C . . . . . B44 . 1035
Zolfaghari P . . . . . E16 . 2459
Zovato E . . . . . D46 . 2243
Zue V . . . . . C16 . 1331
Zue V . . . . . B12 . . 451
Zweig G . . . . . A44 . . 291

Abstracts and Technical Programme - Eurospeech 2001 - Scandinavia

Volume 1 Pages K1 - K12, 1 - 744
Volume 2 Pages 745 - 1522
Volume 3 Pages 1523 - 2278
Volume 4 Pages 2279 - 2852

Affiliations may have been abbreviated

Session notation Xn_1n_2

X = Day
- A: Monday
- B: Tuesday
- C: Wednesday
- D: Thursday
- E: Friday

n_1 = Time Slot
- 1: 09.00 - 10.50
- 2: 11.20 - 12.30
- 3: 14.00 - 15.20
- 4: 15.50 - 17.30

n_2 = Room Track
- 1: Europa Hall (ESE)
- 2: Musiksalen (Oral)
- 3: Radiosalen (Oral)
- 4: The Little Theatre (Oral)
- 5: West Hall - North (Poster)
- 6: West Hall - South (Poster)

Volume Page Session

Session A21 - Oral
Monday - 10.00 - 12.30

Keynotes

Chair: Paul Dalsgaard, CPK, Aalborg University, Denmark

Acquiring and implementing phonetic knowledge

Pols L C W
University of Amsterdam, The Netherlands

Proper early acquisition of speech and language appears to be a necessary process to reach mature speech communication. In modelling the process of natural (and pathological) speech production and speech perception, we frequently concentrate on specific aspects of phonetic knowledge. But also to improve the performance of speech technological systems, an intelligent interpretation of the abundant, but sometimes also incomplete or absent, phonetic information is highly advisable. This keynote will use the above framework to discuss the progress made in speech science and technology over the past 30 years.

Volume 1, page K-3 Session A21

Mobile Future

Neuvo Y
Nokia Mobile Phones, Finland

Mobile communications has fast become one of the key industries globally, and currently the mobile revolution extending to the Internet. Third generation cellular systems will soon be taken in use, opening up new possibilites for mobile services and multimedia. This article gives an overview of the future development of mobile communications, and highlights some of the expectations that this development places on multimedia and speech applications.

Volume 1, page K-7 Session A21

How Visual Co-Presence and Joint Attention Shape Speaking

Brennan S E
Department of Psychology, State University of New York, USA

When people in conversation refer to things, they achieve a joint focus of attention that enables them to be confident that they are both talking about the same thing. When they cannot see one another, such as over a telephone, they must use verbal means to do so. For instance, they may reuse the same expression upon repeated referring, marking the mutual belief that they have achieved a joint perspective; this process has been called entrainment. In contrast, when they can see one another, such as when they are physically co-present in the same environment, they can use visual evidence and deictic (pointing) strategies to formulate and ground references to objects. In this way, visual co-presence shapes conversation.

Volume 1, page K-11 Session A21

Session A31 - Oral
Monday - 14.00 - 15.20

ESE1 - What do Industry and Universities Expect from Each Other?: Panel discussion

Chair: Hynek Hermansky, Oregon Graduate Institute of Science and Technology, USA

Whither Speech Technology? - A Twenty-First Century Perspective

Greenberg S
International Computer Science Institute, USA

Speech-technology research lies at an historic juncture. Commercialization of the technology is likely to accelerate dramatically over the coming decade, but its scientific foundation remains uncertain. A critical shortage of qualified speech scientists and engineers looms in the absence of well-funded, challenging programs for training speech technologists and timely intervention by universities, government agencies and speech-technology companies. The speech-technology industry should collaborate closely with academic and government partners to insure an orderly expansion of academic training and research facilities required to accommodate the inevitable surge in demand for spoken-language technology. In the absence of significant academic-industry-government collaboration the pace of scientific innovation in speech research is likely to slow dramatically.

Volume 1, page 3 Session A31

3G Mobile Networks and Mobile Internet as a Promotor for New Applications - Challenges to Industry and Universities

Dobler S, Hermansson H, Minde T-B
Ericsson Research - Audio Visual Technology, Sweden

The intention of this article is to stimulate the discussion of relationships between university and industry research, especially regarding to applications in future 3rd generation mobile networks. Upcoming 3rd generation mobile networks will be briefly reviewed with respect to required speech technologies. 3G brings together high-speed radio access and IP-based services into one, powerful environment thus enabling to realize Mobile Internet. New location based services expand the capabilities of the traditional Internet, underlining that Mobile Internet is not just about making the Internet mobile. The characteristics of mobile terminals (small size, small keyboard) and the network (data rate) require e.g. automatic speech recognition with high demands on noise robustness, to create convenient user interfaces for the terminals as well as for services. On the other side, wide spread use of mobile communication opens opportunities for speech technologies for the mass market. New applications such as location-based services (e.g. finding nearest cinema and streaming of movie trailers) require to merge different technologies to catalyze these applications and to make them a success. Up to now this symbiosis is not sufficiently reflected in traditional collaboration between research at universities and in industry. An example to motivate interdisciplinary information exchange are the 'Ericsson University Days', which is a regular workshop of Ericsson research with its university partners.

Volume 1, page 7 Session A31

Universities and Industry: Marriage or Co-operation between Independent Partners?

Niiniluoto I
University of Helsinki, Finland

N/A

Volume 1, page 11 Session A31

Considerations on What Industry Expects from Universities

Neuvo Y
Nokia Mobile Phones, Finland

N/A

Volume 1, page 13 Session A31

A Perspective on Industry/University Relationships in the US

Strong G
Information Technology Research Program, NSF, USA

N/A

Volume 1, page 15 Session A31

ELRA Contribution to Bridge the Gap Between Industry and Academia

Choukri K
ELRA/ELDA, Paris, France

N/A

Volume 1, page 17 Session A31

Session A32 - Oral
Monday - 14.00 - 15.20

Linguistic Modelling: Language Model Compression

Chair: Gilles Adda, LIMSI, France

Combining word- and class-based language models: A comparative study in several languages using automatic and manual word-clustering techniques

Maltese G, Bravetti P, Crépy H, Grainger B J, Herzog M, Palou F
IBM Voice Systems, European Speech Research

This paper compares various class-based language models when used in conjunction with a word-based trigram language model by means of linear interpolation. For class-based language models where classes are automatically derived we present a comparative analysis in five languages (French, British English, German, Italian, and Spanish). With regard to classes corresponding to parts-of-speech, we present results for three languages (British English, French, and Italian). For each language, we present results for varying training corpus size and test script complexity. We achieved significant perplexity and word error rate reduction for all five languages and for several language models and recognition tasks. This work extends previous research by covering more languages and showing positive impact of these techniques with very large corpora, whereas prior work mostly focused on addressing data sparseness issues caused by small corpora.

Volume 1, page 21 Session A32

Multi-Class Composite N-gram Language Model Using Multiple Word Clusters and Word Successions

Isogai S[1], Shirai K[1], Yamamoto H[2], Sagisaka Y[2]
[1]Waseda University, Japan, [2]ATR Spoken Language Translation Research Laboratories, Japan

In this paper, a new language model, the Multi-Class Composite N-gram, is proposed to avoid a data sparseness problem in small amount of training data. The Multi-Class Composite N-gram maintains an accurate word prediction capability and reliability for sparse data with a compact model size based on multiple word clusters, so-called Multi-Classes. In the Multi-Class, the statistical connectivity at each position of the N-grams is regarded as word attributes, and one word cluster each is created to represent positional attributes. Furthermore, by introducing higher order word N-grams through the grouping of frequent word successions, Multi-Class N-grams are extended to Multi-Class Composite N-grams. In experiments, the Multi-Class Composite N-grams result in 9.5% lower perplexity and a 16% lower word error rate in speech recognition with a 40% smaller parameter size than conventional word 3-grams.

Volume 1, page 25 Session A32

Statistical Language Model Based On a Hierarchical Approach: MCnv

Zitouni I, Smaili K, Haton J-P
LORIA, France

In this paper, we propose a new language model based on dependent word sequences organized in a multi-level hierarchy. We call this model MCnv, where n is the maximum number of words in a sequence and v is the maximum number of levels. The originality of this model is its capacity to take into account dependent variable-length sequences for very large vocabularies. In order to discover the variable-length sequences and to build the hierarchy, we use a set of 233 syntactic classes extracted from the 8 French elementary grammatical classes. The MCnv model learns hierarchical word patterns and uses them to reevaluate and filter the n-best utterance hypotheses outputted by our speech recognizer MAUD. The model has been trained on a corpus of 43 million words extracted from a French newspaper and uses a vocabulary of 20000 words. Tests have been conducted on 300 sentences. Results achieved 17% decrease in perplexity compared to an interpolated class trigram model. Rescoring the original n-best hypotheses resulted in an improvement of 5% in accuracy.

Volume 1, page 29 Session A32

Quantization-based Language Model Compression

Whittaker E, Raj B
Compaq Cambridge Research Laboratory, USA

This paper describes two techniques for reducing the size of statistical back-off N-gram language models in computer memory. Language model compression is achieved through a combination of quantizing language model probabilities and back-off weights and the pruning of parameters that are determined to be unnecessary after quantization. The recognition performance of the original and compressed language models is evaluated across three different language models and two different recognition tasks. The results show that the language models can be compressed by up to 60% of their original size with no significant loss in recognition performance. Moreover, the techniques that are described provide a principled method with which to compress language models further while minimising degradation in recognition performance.

Volume 1, page 33 Session A32

Session A33 - Oral
Monday - 14.00 - 15.20

Speech Production: Voice Source

Chair: Wolfgang Hess, IKP Universität Bonn, Germany

Relations between vocal registers in voice breaks

Bloothooft G, van Wijck M, Pabon P
Utrecht University, The Netherlands

1783 modal-falsetto register breaks and 853 falsetto-modal register breaks, produced by seven untrained adult male subjects, were recorded and analyzed with respect to jumps in fundamental frequency and sound pressure level (SPL) using a computer phonetograph. SPL and relative positions of modal and falsetto registers were the most important factors underlying the results. Whereas sub- or supraglottal coupling certainly cannot explain the results, models of intrinsic non-linear behavior of the vocal folds may need to be extended for explanations of breaks outside the overlap area of the two registers.

Volume 1, page 39 Session A33

A Quasi-One-Dimensional Model of Aerodynamic and Acoustic Flow in the Time-Varying Vocal Tract: Source and Excitation Mechanisms

Ramsay G
Universite Libre de Bruxelles, Belgium

In this paper, the conservation laws of classical fluid mechanics are used to derive a quasi-one-dimensional model of fluid flow in an elastic tube of time-varying cross-sectional area, representing the human vocal tract. The global flow equations are then decomposed into aerodynamic and acoustic components, representing respiratory flow and sound propagation during speech. The nature of the coupling between the two systems is investigated, and a new interpretation of the traditional source-filter model of speech production is proposed.

Volume 1, page 43 Session A33

Spectral correlates of voice open quotient and glottal flow asymmetry : theory, limits and experimental data

Henrich N[1], d'Alessandro C[2], Doval B[2]
[1]LAM/LIMSI, France, [2]LIMSI, France

The effects of voice open quotient and glottal waveform asymmetry are studied in the spectral domain. The hypothesis that the amplitude difference of the first and second harmonics of the inverse-filtered voice signal (H1*-H2*) is a reliable spectral correlate of the open quotient is tested. Theoretical arguments and experiments are reported. In the theoretical part, analytical formulas are derived for the spectrum of several models (LF, R++,KLGLOTT88). Then it is shown that H1*-H2* is generally dependent on both open quotient and asymmetry. Domains for open quotient, asymmetry and H1*-H2* variations are given. In the experimental part, examples of voice and singing signals are analyzed. It is shown that a significant part of the spectral measurements obtained are out of the scope of the models studied.

Volume 1, page 47 Session A33

One-delayed-mass model for efficient synthesis of glottal flow

Avanzini F[1], Paavo A[2], Karjalainen M[2]
[1]Universita' di Padova, Italy, [2]Helsinki University of Technology, Finland

A lumped physical model of the glottal source is presented. Vocal folds are described as single masses but, unlike conventional one-mass models, vertical phase differences between upper and lower margins of the folds are taken into account. This is done by appropriately describing the non-linear interaction of the mechanical model with aerodynamics, resulting in a modified one-mass model, or a 'one-delayed-mass model'. Analysis on numerical simulations shows that the system behaves qualitatively as higher-dimensional ones (such as the two-mass model by Ishizaka and Flanagan); in particular, control over flow skewness is guaranteed, allowing for synthesis of realistic glottal flow waveforms. As only one degree of freedom (one mass) is needed in the model, structure and number of parameters are drastically reduced, thus making it suitable for real-time synthesis applications.

Volume 1, page 51 Session A33

Session A34 - Oral
Monday - 14.00 - 15.20

Speech Recognition and Understanding: Pronunciation and Subword Units - I

Chair: Torbjørn Svendsen, NTNU, Norway

Modeling Pronunciation Variation Using Context-Dependent Weighting and B/S Refined Acoustic Modeling

Zheng F[1], Song Z[1], Fung P[2], Byrne W[3]
[1]Tsinghua Universiy, P. R. China, [2]University of Science and Technology, Hong Kong, [3]The Johns Hopkins University, USA

The pronunciation variability is an important issue that must be faced with when developing practical automatic spontaneous speech recognition systems. By studying the initial/final (IF) characteristics of Chinese language and developing the Bayesian equation, we propose the concepts of generalized initial/final (GIF) and generalized syllable (GS), the GIF modeling method and the IF-GIF modeling method, as well as the context-dependent pronunciation weighting method. By using these approaches, the IF-GIF modeling reduces the Chinese syllable error rate (SER) by 6.3% and 4.2% compared with the GIF modeling and IF modeling respectively when the language modeling, such as syllable or word N-gram, is not used.

Volume 1, page 57 Session A34

Learning Units for Domain-Independent Out-of-Vocabulary Word Modelling

Bazzi I, Glass J
MIT Laboratory for Computer Science, USA

This paper describes our recent work on detecting and recognizing out-of-vocabulary (OOV) words for robust speech recognition and understanding. To allow for OOV recognition within a word-based recognizer, the in-vocabulary (IV) word network is augmented with an OOV model so that OOV words are considered simultaneously with IV words during recognition. We explore several configurations for the OOV model, the best of which utilizes a set of domain-independent, automatically derived, variable-length units. The units are created using an iterative bottom-up procedure where, at each iteration, the unit pairs with maximum mutual information are merged. When evaluating this method on a weather information domain, the false alarm rate of our baseline OOV model is reduced by over 60%. For example, with an OOV detection rate of 70%, the OOV false alarm rate is reduced from 8.5% to 3.2%, with only 3% relative degradation in word error rate on IV data.

Volume 1, page 61 Session A34

Pronunciation Variant Analysis using Speaking Style Parallel Corpus

Nakajima H[1], Hirano I[2], Sagisaka Y[1], Shirai K[2]
[1]ATR Spoken Language Translation Research Laboratories, Japan, [2]Waseda University, Japan

To improve the recognition accuracy for spontaneous conversational speech, we collected a corpus to study how spontaneous conversational speech differs from read style speech. The corpus consists of two parts: 1) spontaneous conversational speech and 2) read speech with the same word transcriptions as the conversational speech. In word and phone recognition experiments, it was confirmed that, for the Japanese language, the recognition of spontaneous speech is harder than that of read speech. By comparing of recognition results, we found that, both in the occurrence of errors appearing with speaking style changes, and in the types of pronunciation variants, there are differences that depend on the linguistic categories that misrecognized words belong to. We confirmed that linguistic categories also affect pronunciation variants that deteriorate the recognition accuracy.

Volume 1, page 65 Session A34

Speech recognition for huge vocabularies by using optimized sub-word units

Kneissler J, Klakow D
Philips Forschungslaboratorien, Germany

This paper describes approaches for decomposing words of huge vocabularies (up to 2 million) into smaller particles that are suitable for a recognition lexicon. Results on a Finnish dictation task and a flat list of German street names are given.

Volume 1, page 69 Session A34

Session A35 - Poster
Monday - 14.00 - 15.20

Phonetics and Phonology: Prosody and Others

Chair: Anne Bonneau, Loria, France

Factors affecting schwa-insertion in final consonant clusters in Standard Dutch

Swerts M, Kloots H, Gillis S, De Schutter G
Antwerp University, Belgium

The current paper describes a study that deals with the factors that determine the possible insertion of a schwa in final consonant clusters in Standard Dutch. Our study reveals that the absence or occurrence of such an extra vowel is dependent on regional, social and phonotactic determinants. We discuss how this finding, in combination with other results on speech variability, is important for speech technological applications, both for synthesis and recognition.

Volume 1, page 75 Session A35

Vowel Height is Intimately Associated with Stress Accent in Spontaneous American English Discourse

Hitchcock L, Greenberg S
International Computer Science Institute, USA

There is a systematic relationship between stress accent and vocalic identity in spontaneous English discourse (the Switchboard corpus of telephone dialogues). Low vowels are much more likely to be fully accented than their high vocalic counterparts. And conversely, high vowels are far more likely to lack stress accent than low or mid vocalic segments. Such patterns imply that stress accent and vowel height are bound together at some level of lexical representation. Vocalic duration appears to be the primary acoustic cue associated with stress accent, and the association between vowel height and accent level is most clearly observed in this dimension, particularly for diphthongs and the low, tense monophthongs. Together, the data suggest that vocalic duration plays an exceedingly important role in understanding spoken language.

Volume 1, page 79 Session A35

Finite State prosodic analysis of African corpus resources

Gibbon D
Universität Bielefeld, Germany

The issue of efficient language documentation, particularly with regard to minority and endangered languages, has gained in importance in recent years, as witnessed by several major funding programmes and other human language technology initiatives in the field. An application of finite state technologies to the processing of lexical tone variation in annotated corpora of African languages is described. It is shown that finite state transducers can be constructed which not only provide adequate models for contextual variation in lexical tone (including automatic downstep, downdrift, and tonal assimilations, but also that the transducers provide intuitively satisfying explications of prosodic concepts in `metrical phonology' in terms of oscillations (iterative transitions). The technique has both theoretical value in formalising typological differences in African lexical tone languages and practical value in automatically generating markup enhancements for concordance-based corpus analysis and for fundamental frequency prediction in pitch modelling.

Volume 1, page 83 Session A35

Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis

Schröder M[1], Cowie R[2], Douglas-Cowie E[2], Westerdijk M[3], Gielen S[3]
[1]DFKI, Saarbrücken, Germany, [2]Queen's University, Belfast, Northern Ireland, [3]University of Nijmegen, The Netherlands

In a database of emotional speech, dimensional descriptions of emotional states have been correlated with acoustic variables. Many stable correlations have been found. The predictions made by linear regression widely agree with the literature. The numerical form of the description and the choice of acoustic variables studied are particularly well suited for future implementation in a speech synthesis system, possibly allowing for the expression of gradual emotional states.

Volume 1, page 87 Session A35

Measuring Pitch Range

Ouden, den H, Terken J
Technische Universiteit Eindhoven, The Netherlands

The literature offers at least two methods to annotators for characterizing the pitch range of a prosodic phrase. One method is in terms of the distance between the F0 maximum of the phrase (HiF0) and the speaker's utterance-final pitch (LoF0). The other method is in terms of the distance between pitch peaks and pitch valleys in the prosodic phrase. In this paper we address two questions. The first question concerns the reliability of the different methods. Five experienced phoneticians applied both methods on a set of forty utterances taken from read aloud text. We found that reliability was higher for HiF0 than for distances between pitch peaks and valleys. The second question is whether variation that is not captured by the first approach does actually occur in pitch contours. The results suggests that the HiF0 approach captures all the variation relevant to measuring pitch range that occurs in our small corpus. We conclude that the HiF0 method is methodologically more adequate, and at the same time sufficiently powerful to represent pitch range variation adequately for read aloud text.

Volume 1, page 91 Session A35

Measuring speech rhythm

Gibbon D, Gut U
Universität Bielefeld, Germany

We address the question of rhythm variation in typologically different languages (English, said to be a stress-timed language, Ibibio, said to be a syllable-timed language) and in different varieties of the same language (British and Nigerian English). Attempts to find correlates of different rhythm types in the acoustic signal have so far not been particularly successful. We examine a number of previous studies, in search of a promising measure of rhythm, and select a recently developed measure (the Pairwise Variability Index of Low & Grabe), with minor modifications and the addition of a binary classifier for focal and nonfocal components of rhythm units. The measure and the classifier are implemented as a software tool which takes esps/waves+ label files as input, and generates statistics on durations, duration differences, the rhythm measure, and a classification of the syllables in the labeled utterance. The results show distinct differences in stress-timing and syllable-timing between Ibibio and English.

Volume 1, page 95 Session A35

Tonal alignment, scaling and slope in Italian question and statement tunes

D'Imperio M
LORIA - University of Nancy, France

Unlike in languages such as English and Standard Italian, Neapolitan Italian yes/no questions and narrow focus statements share a rising-falling (LHL) tune (D'Imperio, 2001) whose H peak "alignment" is a cue

to tune identification (D'Imperio and House, 1997). This study acoustically tested the hypothesis that all three tonal targets of the rise-fall are timed and scaled differently in questions and statements. Moreover, slope differences for both rise and fall were also tested by employing logistic regression modeling. Two Neapolitan speakers produced utterances whose target words differed in question/statement modality, syllable structure and segmental environment. The results show that all three targets within the rise-fall are timed later in questions. By contrast, no systematic slope difference was found. The exact contribution of scaling to signaling the question/statement contrast could not be determined, though. In fact, while one speaker produced higher peaks for statements, the other did not produce any scaling difference.

Volume 1, page 99 Session A35

Pragmatic temporal voice range profile as a tool in the research of speech styles

Iivonen A K
University of Helsinki, Finland

Speaker's pragmatic temporal voice range profile (TVRP) and F0 distribution histogram were used for comparing two radio speakers and two speaking styles in one speaker. It is assumed that the internal distribution of the temporal voice range, relative placement of the range on ST scale and the concentration of F0 points in certain semitone classes of the F0 histograms are connected with the speech style and the factors bedhind the style. Placing a single utterance against the background of TVRP its features can be better understood within the prosodic repertoire of the speaker.

Volume 1, page 103 Session A35

Model Based Stress Decision Method

Kim W, Kim T, Ahn S, Ko H
Korea University, Korea

This paper proposes an effective decision method focused to evaluate the ``stress position". Conventional methods usually extract the acoustic parameters and compare them to reference in absolute scale, adversely producing unstable results as testing condition changes. To cope with the environmental dependency,the proposed method is designed to be model-based that determines the stressed interval by making relative comparison over candidates. The stressed/unstressed models are then induced from normal phone models by adaptive training. The experimental results indicate that the proposed method is promising and that it is useful for automatic detection of stress positions. The results also show that generating the stressed/unstressed model by adaptive training is effective.

Volume 1, page 107 Session A35

Reduction of alternative pronunciations in the Norwegian computational lexicon NorKompLeks

Nordgård T, Foldvik A K
Norwegian University of Science and Technology, Norway

This paper describes a method for selecting one single pronunciation from a set of alternatives. Two types of reductions are described - base form reductions and inflected form reductions. The strategies are different because the inflected forms have to take care of two properties of Norwegian: The language allows for many spelling alternatives, both in base forms and inflected forms, and many morphological inflection types license pronunciation alternatives. A comprehensive pronunciation lexicon with more than 65000 base forms is used, and it is demonstrated that it is possible for users of the lexicon to derive their own stylistic preferences.

Volume 1, page 111 Session A35

The Role of Duration as a Correlate of Accent in Lekeitio Basque

Elordieta G[1], Hualde J I[2]
[1]University of the Basque Country, Spain, [2]University of Illinois, USA

Northern Bizkaian Basque shares important prosodic features with Tokyo Japanese, including the existence of a lexical distinction between accented and unaccented content words, the presence of phrase-initial rises and the consistent realization of accents as tonal falls. In this paper we investigate whether NB Basque is also like Japanese in not making use of syllable duration as a correlate of accent, as has been suggested in recent work. The analysis of experimental data from 6 native speakers of the variety spoken in the Bizkaian town of Lekeitio confirms the hypothesis that the presence of an accent on a given syllable is not manifested in an increase of its duration in this language. Other things being equal, accented and unaccented syllables do not have significantly different durations in neutral declarative sentences. More tentatively, the same results are also established for two other sentential conditions: under narrow focus and in postfocus position.

Volume 1, page 115 Session A35

Word Final Aspiration as a Phrase Boundary Cue: Data from Spontaneous Swedish Discourse

Johansson V, Horne M, Strömqvist S
Lund University, Sweden

The Swedish sound string /at/ (graphically: att) is associated with two grammatical functions: a) (part of) a subordinate conjunction and b) as an infinitive marker. Previous studies connect final lengthening and pauses with prosodic and syntactic boundaries in spoken discourse. Following these findings, this pilot study, with 5 short spontaneous discourses from 3 male speakers shows a correlation between pauses after att, and aspiration of /t/ in att. We also show a tendency for att with aspiration to be associated with the grammatical function of subordinate conjunction. Further, looking at the distribution of aspiration in the subordinate conjunction att, and in the infinitive marker att, we are able to show a tendency for the infinitive marker to be unaspirated in the normal case, while the subordinate conjunctions are characterized by final aspiration in 40 % of the cases.

Volume 1, page 119 Session A35

Study and Auto-Detection of Stress Based on Tonal Pitch Range in Mandarin

Shen X, Xu B
Institute of Automation, Chinese Academy of Sciences, P. R. China

In Mandarin, there is a special acoustic feature¡ªtonal pitch range, which is relative to stress. In this paper, we present a novel concept¡ªtonal range ratio (TRR), which is based on tonal pitch range, and make a study on the correlation between TRR and stress in Mandarin. And we developed a system to automatically detect stresses in words and sentences based on TRR in Mandarin. We obtained high success rate (92.67% in words and 82.0% in sentences). The results show that TRR has strong correlation with stress and is powerful in detecting stresses in Mandarin.

Volume 1, page 123 Session A35

Classifying emotions in speech: a comparison of methods

Amir N[1], Kerret O[2], Karlinski D[2]
[1]Holon Academic Institute of Technology, Israel, [2]Academic College of Tel Aviv Yaffo, Israel

A number of recent studies have attempted classification of emotional speech using various methods. In this paper we compare the performance of two algorithms: a classification algorithm based on euclidean distances, and a classification algorithm based on the use of neural

networks. Both perform the classification using an identical feature set, on a database of emotional speech which has been validated through subjective listening tests.

Volume 1, page 127 Session A35

Session A36a - Poster
Monday - 14.00 - 15.20

Speech Perception: First and Second Language Learning

Chair: Francisco Lacerda, Stockholm University, Sweden

Development of vowel quantity perception in late childhood

Behne D M[1], Czigler P E[2], Sullivan K P H[2]
[1]Norwegian University of Science and Technology, Norway, [2]Umeå University, Sweden

A distinction in vowel quantity is typically realized acoustically by vowel duration. Research on the perception of Swedish vowel quantity by adult native speakers supports this. It further suggests that when the duration of a vowel is relatively long (due, e.g., to inherent duration), listeners may also make use of vowel spectra to distinguish vowel quantities. The current project investigates the perceptual cues used to distinguish vowel quantities in language development by children 9 to 13 years old. Of particular interest is whether these developing listeners use spectral cues to identify the quantity of vowels which have a relatively long inherent duration. Results are compared with the findings for Swedish adults and the developmental use of vowel duration and spectra as cues for vowel quantity are described.

Volume 1, page 133 Session A36a

A study on the production-perception link of English vowels produced by native and non-native speakers

Yang B
Dongeui University, Korea

This study explored the relationship between the production of the nine English vowels and the perception of synthesized vowels by thirty-five American, Chinese, and Korean, male and female speakers. The average formant values of the ten American English speakers were employed to synthesize the nine vowels that were presented to the thirty-five speakers. The center formant values of the highest and lowest formant boundary of the same vowel quality were collected and compared to the formant values of their productions. We found that there was a strong correlation between production and perception within and across the language groups. The American, Chinese and Korean groups perceived the stimuli about the same. Individual comparison by regressional analyses of the formant frequency data of the produced vowels and the center formant values of the perceptual test led to a very remarkable r-squared value. This suggests a very lawful relationship between production and perception.

Volume 1, page 137 Session A36a

Japanese Can be Aware of Syllables and Morae: Evidence from Japanese-English Bilingual Children

Otake T, Yamaguchi Y
Dokkyo University, Japan

This study investigated the metalinguistic knowledge of the internal structure of syllables by Japanese-English bilingual children. Our recent study revealed that Japanese-English adult bilinguals could be aware of two constituents of syllable structure, syllables and morae, depending upon the nature of input materials, while monolingual speakers were aware of morae irrespective of input materials. Three experiments were conducted with 10 bilingual Japanese children, using CVCVNCV materials in Japanese, English and Spanish in order to test whether the same phenomenon could be observed by bilingual children. The subjects were asked to identify the number of chunks within the materials which were presented aurally and to stamp the number of them on a test sheet.

The results showed that they preferred moare in Japanese, but syllables in English and Spanish, suggesting that an ability to manipulate two languages freely may have a function to suppress moraic consciousness.

Volume 1, page 141 Session A36a

Neural Processes Underlying Perceptual Learning of a Difficult Second Language Phonetic Contrast

Callan D, Tajima K, Callan A, Akahane-Yamada R, Masaki S
ATR International, Japan

Neural processes underlying the perceptual learning of the English /r-l/ phonetic contrast by native Japanese speakers before and after extensive perceptual identification training using feedback was investigated using fMRI. Relative to control conditions (English /b-v/ and /b-g/ contrasts), the /r-l/ contrast showed greater brain activity as well as functional connectivity (reflecting underlying global mappings) post- relative to pre- training bilaterally in frontal and temporal brain areas involved with speech processing as well as the cerebellum and the putamen.

Volume 1, page 145 Session A36a

Human Language Identification with Reduced Segmental Information: Comparison between Monolinguals and Bilinguals

Komatsu M[1], Mori K[2], Arai T[2], Murahara Y[2]
[1]Sophia University, Japan / University of Alberta, Canada, [2]Sophia University, Japan

We conducted human language identification experiments using signals with reduced segmental information with Japanese and bilingual subjects. American English and Japanese excerpts from the OGI_TS Corpus were processed by spectral-envelope removal (SER), vowel extraction from SER (VES) and temporal-envelope modulation (TEM). With the SER signal, where the spectral-envelope is eliminated, humans could still identify the languages fairly successfully. With the VES signal, which retains only vowel sections of the SER signal, the identification score was low. With the TEM signal, composed of white-noise-driven intensity envelopes from several frequency bands, the identification score rose as the number of bands increased. Results varied depending on the stimulus language. Japanese and bilingual subjects demonstrated different scores from each other. These results indicate that humans can identify languages using a signal with drastically reduced segmental information. The results also suggest variation due to the phonetic attributes of languages and subjects fknowledge.

Volume 1, page 149 Session A36a

Session A36b - Poster
Monday - 14.00 - 15.20

Speech Perception: Miscellaneous - I

Chair: Hiroaki Kato, ATR, Japan

Coarticulatory effects in perception

Fernández S, Feijóo S
Universidad de Santiago de Compostela, Spain

The perceptual interaction between adjacent CV segments is studied in Fricative-Vowel syllables from a coarticulatory point of view. The results of a perceptual experiment with conflicting-cue stimuli are analyzed under the assumption, supported by the results, of the possible influence of F-to-V carryover coarticulation on the integration process. The DAC scale was used to estimate the degree of F-to-V carryover coarticulation in the original syllables. The magnitude of the coarticulatory effect permitted us to derive a prediction of the perceptual results based on the articulatory compatibility between the fricative and the vowel. The correlation between actual and predicted decrease in perceptual identification caused by the insertion of a conflicting transition was computed. The results show that the coarticulatory processes cannot explain the outcome of the perceptual experiment. Nevertheless, the perceptual role played by the /i/ transition can be effectively explained as a consequence of the F-to-V coarticulation.

Volume 1, page 155 Session A36b

A Case for Multi-Resolution Auditory Scene Analysis

Harding S, Meyer G
Keele University, UK

A commonly held view of auditory scene analysis is that complex auditory environments are segregated into separate perceptual streams using primitive cues that can be attended to separately. We argue that this view is inconsistent with the majority of perceptual data reported in the literature and propose an alternative model that is based on a primary, low resolution signal representation used in a passive pattern matching stage, augmented by secondary, high resolution representations that can be used in an active pattern matching stage to formulate hypotheses about the auditory scene.

Volume 1, page 159 Session A36b

Perceptual Identification and Normalization of Synthesized French Vowels from Birth to Adulthood

Ménard L[1], Schwartz J-L[1], Boë L-J[1], Kandel S[2], Vallée N[1]
[1]Université Stendhal, France, [2]Université Joseph-Fourier, France

This paper aims at exploring the invariant parameters involved in auditory normalization of French vowels. A set of 490 stimuli, including the ten French vowels produced by an articulatory model simulating seven growth stages and seven fundamental frequency values, has been submitted as a perceptual test to 43 subjects. Results confirm the important effect of the tonality distance between F1 and F0 in perceived height. Regarding place of articulation, F2-F1, and F3-F2, in Bark, appear to be good predictors of the perceived front-back dimension. Roundedness is also examined and correlated to the effective second formant, involving spectral integration of higher formants within the 3.5-Bark critical distance.

Volume 1, page 163 Session A36b

Perceptual Categorization of Maximal Vowel Spaces from Birth to Adulthood Simulated by an Articulatory Model

Ménard L, Boë L-J

Université Stendhal, France

This paper reports on an experiment aiming at determining the perceptual effects of non uniform vocal tract growth. An articulatory model was used to synthesize 342 stimuli, covering the maximal vowel space in the F1/F2 and F2/F3 dimensions, for 5 growth stages: a newborn, a 4-year-old, a 10-year-old, a 16-year-old, and a 21-year-old male speakers. Results of a categorization test of the stimuli by 40 French adult subjects reveal that for each vocal tract length, French phonological categories can be perceived. Furthermore, perceived front vowels cover a broader range in the acoustic F1/F2 space for very small vocal tracts, compared to adults. Data are interpreted in the light of the articulatory-to-acoustic mapping from a developmental point of view, and can shed light on the existence of perceptual constraints during vocal tract growth.

Volume 1, page 167 Session A36b

A study on speech over the telephone and aging

Eskenazi M, Black A
Carnegie Mellon University, USA

We describe an experiment to show how the comprehensibility of speech over the telephone is related to the age of the listener. Our intention is to show figures to prove the commonly-held belief that as we get older our hearing of information over the telephone degrades. The study was set up to determine, for all age groups from 20-29 to 80-89, whether comprehension degrades with age and with the type of speech (synthetic or natural). We gave subjects sentences containing target word pairs that they were to write down. The pairs contained more or less predictable words. Our findings, which we consider to be preliminary due to the sample size, show degradation of comprehension with age and degradation from natural speech to synthetic speech.

Volume 1, page 171 Session A36b

On the Perception of Voicing for Plosives in Noise

Chen M, Alwan A
University of California at Los Angeles, USA

Previous research has shown that the VOT and first formant transition are primary perceptual cues for the voicing distinction for syllable-initial plosives (SLP) in quiet environments. This study seeks to determine which cues are important for the perception of voicing for SLP in the presence of noise. Stimuli for the perceptual experiments consisted of naturally-spoken /CV/ syllables (six plosives in 3 vowel contexts) in varying levels of additive white Gaussian noise. In each experiment, plosives which share the same place of articulation (e.g. /p, b/) were presented to subjects in identification tasks. For each voiced/voiceless pair, a threshold SNR value was calculated. Threshold SNR values were then correlated with measurements of several acoustic parameters of the speech tokens. It was found that the VOT did not appear to influence the perception of voicing in noise as much as the first formant transition.

Volume 1, page 175 Session A36b

Predicting Visual Consonant Perception from Physical Measures

Jiang J[1], Alwan A[1], Auer E[2], Bernstein L[2]
[1]University of California at Los Angeles, USA, [2]House Ear Institute, Los Angeles, USA

The long term goal of our work is to predict visual confusion matrices from physical measurements. In this paper, four talkers were chosen to record 69 American-English Consonant-Vowel syllables with audio, video, and facial movements captured. During the recording, 20 markers were put on the face and an optical Qualisys system was used to track three-dimensional facial movements. The videotapes (with markers on the face and without sound) were presented to normal hearing viewers with average or above average lipreading ability, and visual confusion matrices were obtained. Results showed that the facial measurements were correlated with visual perception data by about 0.79 and account for about 63% of the variance.

Volume 1, page 179 Session A36b

Session A41 - Oral & Poster
Monday - 15.50 - 18.00

ESE2 - Noise Robust Recognition: Front-end and Compensation Algorithms

Chair: Børge Lindberg, CPK, Denmark

Session Introduction & Aurora Status (oral presentation, no proceedings paper)

Pearce D
Motorola Labs, UK

N/A

Volume 1, page 184 Session A41

Aurora 2 Database Training and Test Sets Description (oral presentation, no proceedings paper)

Hirsch H-G
University of Applied Sciences Niederrhein, Germany

N/A

Volume 1, page 184 Session A41

Noise Robust Feature Extraction for ASR using the Aurora 2 Database

Zhu Q, Iseli M, Cui X, Alwan A
University of California, Los Angeles, USA

Four front-end processing techniques developed for noise robust speech recognition are tested with the Aurora 2 database. These techniques include three previously published algorithms: variable frame rate analysis [Zhu and Alwan, 2000], peak isolation [Strope and Alwan, 1997], and harmonic demodulation [Zhu and Alwan, 2000], and a new technique for peak-to-valley ratio locking. Our previous work has focused on isolated digit recognition. In this paper, these algorithms are modified for recognition of connected digits. Recognition results with the Aurora 2 database show that a combination of these four techniques results in 40% error rate reduction when compared to the baseline MFCC front-end for the clean training condition, with no significant increase in computational complexity.

Volume 1, page 185 Session A41

Investigations into Tandem Acoustic Modeling for the Aurora Task

Ellis D P W, Reyes Gomez M J
Columbia University, USA

In tandem acoustic modeling, signal features are first processed by a discriminantly-trained neural network, then the outputs of this network are treated as the feature inputs to a conventional distribution-modeling Gaussian-mixture model speech recognizer. This arrangement achieves relative error rate reductions of 30% or more on the Aurora task, as well as supporting feature stream combination at the posterior level, which can eliminate more than 50% of the errors compared to the HTK baseline. In this paper, we explore a number of variations on the tandem structure: We experiment with changing the subword units used in each model (neural net and GMM), varying the data subsets used to train each model, substituting the posterior calculations in the neural net with a second GMM, and a variety of feature condition such as deltas, normalization and PCA rank reduction in the `tandem domain' i.e. between the two models.

Volume 1, page 189 Session A41

Recognition Performance of the Siemens Front-end with and without Frame Dropping on the Aurora 2 Database

Andrassy B[1], Vlaj D[2], Beaugeant C[1]
[1]Siemens, München, Germany, [2]University of Maribor, Slovenia

Following the objective of the Eurospeech special event, 'Noise Robust Recognition', the recognition results of a noise robust front-end, developed by Siemens, on the Aurora 2 database [1] are presented in this paper. The front-end was tested with and without a frame dropping algorithm. It is shown that the front-end improves the recognition results in high mismatch between training and testing by 43.90% over the reference front-end and works particularly well in conditions with high noise. Furthermore it is shown that the frame dropping mainly increases the performance of the front-end.

Volume 1, page 193 Session A41

A Multiconditional Robust Front-End Feature Extraction with a Noise Reduction Procedure Based on Improved Spectral Subtraction Algorithm

Kotnik B, Kacic Z, Horvat B
University of Maribor, Slovenia

In this paper, the procedure for feature vector extraction in multiconditionally noisy environments is presented. Proposed front-end uses time and spectral domain processing for noise reduction as well as feature extraction to create mel-cepstrum parameters and achieves a trade-off between effective noise reduction and low computational load for real-time operations. First, a novel weighting function is used to reduce the rough noise in time domain, and then a spectral subtraction method based on minimum statistics is applied to decrease the effect of additive broadband noise on speech in the spectral domain. At final stage, a feature vector, which consists of 12 mel-cepstrum parameters and the energy, is created. For evaluation of improvement of speech recognition with presented front-end, the "Aurora 2" database together with the HTK recognition toolkit have been chosen. With proposed method an average improvement in performance of 24.75% relative to the current ETSI Aurora standard was achieved.

Volume 1, page 197 Session A41

Feature Vector Selection to Improve ASR Robustness in Noisy Conditions

de Veth J[1], Mauuary L[2], Noe B[3], de Wet F[1], Sienel J[3], Boves L[1], Jouvet D[2]
[1]University of Nijmegen, The Netherlands, [2]France Télécom R&D, France, [3]Alcatel SEL, Germany

It is well known that noise reduction schemes are beneficial in ASR to reduce training-test mismatch due to noise. However, a significant mismatch may still remain after noise reduction, especially in the non-speech portions of the signals. To reduce the impact of this mismatch, two methods for discarding non-speech acoustic vectors at recognition time are investigated: variable frame rate processing and voice activity detection. Experiments are discussed for Aurora 2 and for SpeechDat Car Italian. Results show that both methods are highly effective for SpeechDat Car Italian. However, for Aurora 2, feature vector selection based on voice activity detection hardly gives a benefit, while variable frame rate processing actually lowers recognition accuracy somewhat. Several possible explanations of the different results observed for the two databases are discussed.

Volume 1, page 201 Session A41

Comparison of Spectral Derivative Parameters for Robust Speech Recognition

Macho D, Nadeu C
TALP Research Center - UPC, Spain

Recently, spectral first-derivative parameters obtained by frequency filtering (FF) have been successfully used in both clean and noisy HMM speech recognition. In this paper, two types of spectral derivative parameters, the usual FF features and the relative spectral difference (RSD) features, are compared both between them and with their second-derivative versions. Additionally, another kind of recently introduced robust speech features, the SBCOR parameters, are related theoretically with the second-derivative RSD. By experimentally comparing all those types of features in the Aurora 2.0 noisy database framework, we conclude that the first-derivative parameters are preferable to the second-derivative ones (and to the MFCC) for both clean and noisy speech recognition, and the RSD parameters show the best average performance.

Volume 1, page 205 Session A41

Robust Digit Recognition in Noise: An Evaluation Using the AURORA Corpus

Yapanel U, Hansen J H L, Sarikaya R, Pellom B
Univ. of Colorado Boulder, USA

In this paper, a variety of techniques for robust digit recognition in noise are considered using the AURORA 2.0 corpus. Current recognizers perform as well as humans in small vocabulary tasks but computer recognition performance degrades substantially when noise is introduced into the speech, while human performance is much less sensitive. To make the recognizer robust, several methodologies are employed. These include, feature processing, enhancement before recognition and model adaptation. We considered a number of processing and adaptation scenarios depending on noise type. The best performance, as expected, was obtained in matched training conditions which in general has limited applicability in real world problems. As a feature processing step, using RCCs (Root Cepstrum Coeff.) instead of MFCCs gave substantial improvement. MFCC with front-end enhancement increased performance considerably, but results were far from that obtained with matched training. When we combine the RCC with enhancement, however, we get the best results. In the next step, we employed model adaptation techniques which outperformed MFCC+enhancement and gave much closer results to the matched condition limits. However, MFCC adaptation could not outperform RCC parameterization with front-end enhancement, which we show is much more computationally efficient than model adaptation.

Volume 1, page 209 Session A41

Robust ASR Based On Clean Speech Models: An Evaluation of Missing Data Techniques For Connected Digit Recognition in Noise

Barker J P, Cooke M, Green P
Sheffield University, UK

In this study, techniques for classification with missing or unreliable data are applied to the problem of noise-robustness in Automatic Speech Recognition (ASR). The techniques described make minimal assumptions about any noise background and rely instead on what is known about clean speech. A system is evaluated using the Aurora 2 connected digit recognition task. Using models trained on clean speech we obtain a 65% relative improvement over the Aurora clean training baseline system, a performance comparable with the Aurora baseline for multicondition training.

Volume 1, page 213 Session A41

Evaluation of the SPLICE Algorithm on the Aurora2 Database

Droppo J, Deng L, Acero A
Microsoft Research, USA

This paper describes recent improvements to SPLICE, Stereo-based Piecewise Linear Compensation for Environments, which produces an estimate of cepstrum of undistorted speech given observed cepstrum of distorted speech. For distributed speech recognition applications, SPLICE can be placed at the server, thus limiting the processing that would take place at the client. We evaluated this algorithm on the Aurora2 task, which consists of digit sequences within the TIDigits database that have been digitally corrupted by passing them through a linear filter and/or by adding different types of realistic. On set A data, for which matched training data is available, we achieved a 66% decrease in word error rate over the baseline system with clean models. This preliminary result is of practical significance because in a server implementation, new noise conditions can be added as they are identified once the service is running.

Volume 1, page 217 Session A41

Model-based compensation of the additive noise for continuous speech recognition. Experiments using the AURORA II database and tasks

Segura J C, Torre A D L, Benitez M C, Peinado A M
Universidad de Granada, Spain

In this paper we apply a model-based compensation method to cancel the effect of the additive noise in Automatic Speech Recognition systems. The method is formulated in a statistical framework in order to perform the optimal compensation of the noise effect given the observed noisy speech, a model describing the statistics of the speech recorded in a clean reference environment and the estimation of the noise in the noisy recognition environment. The noise is estimated using the first frames of the sentence to be recognized and a frame-by-frame noise compensation algorithm is performed, so that the compensation procedure does not constrain real-time speech recognition systems and is compatible with emerging technologies based on distributed speech recognition. We have performed recognition experiments under noise conditions using the AURORA II database for the recognition tasks developed for thisdatabase as a standard reference. Experiments have been carried out including both, clean and multicondition training approaches. The experimental results show the improvements in the recognition performance when the proposed model-based compensation method is applied.

Volume 1, page 221 Session A41

MAP Combination of Multi-Stream HMM or HMM/ANN Experts

Morris A C, Hagen A, Bourlard H
IDIAP, Switzerland

Automatic speech recognition (ASR) performance falls dramatically with the level of mismatch between training and test data. The human ability to recognise speech when a large proportion of frequencies are dominated by noise has inspired the "missing data" and "multi-band" approaches to noise robust ASR. "Missing data" ASR identifies low SNR spectral data in each data frame and then ignores it. Multi-band ASR trains a separate model for each position of missing data, estimates a reliability weight for each model, then combines model outputs in a weighted sum. A problem with both approaches is that local data reliability estimation is inherently inaccurate and also assumes that all of the training data was clean. In this article we present a model in which adaptive multi-band expert weighting is incorporated naturally into the maximum a posteriori (MAP) decoding process.

Volume 1, page 225 Session A41

Second Order Statistics Spectrum Estimation Method for Robust Speech Recognition

Jarc B, Babic R
University of Maribor, Slovenia

A second order statistics spectrum estimation (SOSSE) method for speech enhancement is presented. DFT amplitude spectral components of noisy signal are assumed to be random values. Upon first and second order statistic values estimation of noise-only spectrum, an enhancement of noisy signal spectrum was performed. As a reference, a fast discrete cosine transform based signal subspace (FDCTSS) method was realized. The Aurora 2 database of digit sequences was used, to show methods effectiveness in improvement of speech recognition. Both methods proved well under clean training condition. The total relative improvements of 30.75% (SOSSE) and 26.31% (FDCSS) in recognition accuracy were achieved. When the multi-condition training was done the proposed SOSSE method outperformed FDCTSS method. The total relative improvements of 17.50% (SOSSE) and -4.53% (FDCTSS) were achieved.

Volume 1, page 229 Session A41

Feature Extraction and Model-Based Noise Compensation for Noisy Speech Recognition evaluated on AURORA 2 Task

Yao K, Chen J, Paliwal K K, Nakamura S
ATR Spoken Language Translation Research Labs., Japan

We have evaluated several feature-based and a model-based method for robust speech recognition in noise. The evaluation was performed on Aurora 2 task. We show that after a sub-band based spectral subtraction, features can be more robust to additive noise. We also report a robust feature set derived from differential power spectrum (DPS), which is not only robust to additive noise, but also robust to spectrum colorization due to channel effects. When the clean training set is available, we show that a model-based noise compensation method can be effective to improve system robustness to noise. Given the testing sets, as a whole, the feature-based methods can yield about 22% relative improvement in accuracy for multi-condition training task, and the model-based method can have about 63% relative performance improvement when systems were trained on clean training set.

Volume 1, page 233 Session A41

Session A42 - Oral
Monday - 15.50 - 17.30

Linguistic Modelling: Language Model Adaptation

Chair: Yoshinori Sagisaka, Waseda University GITI, Japan

Broadcast News LM Adaptation using Contemporary Texts

Federico M, Bertoldi N
ITC-Irst, Italy

This paper investigates the problem of dynamically updating the language model (LM) of a broadcast news speech recognition system, in order to cope with language and topic changes, typical of the news domain. Statistical adaptation methods are proposed that exploit written news sources which are daily available on the Internet, i.e. newswires and newspapers. Specifically, LM adaptation is performed by extending the basic lexicon, in order to minimize the out-of-vocabulary (OOV) rate, and by adapting the word probability distribution on the contemporary data. Experiments performed on 19 newscasts showed relative reductions of 58% on the OOV rate, 16% on the perplexity, and 4% on the word error rate.

Volume 1, page 239 Session A42

Topic Detection for Language Model Adaptation of Highly-Inflected Languages by Using a Fuzzy Comparison Function

Sepesy Maucec M, Kacic Z
University of Maribor, Slovenia

A new framework is proposed to construct corpus-based topic-adapted language models for large vocabulary speech recognition of highly-inflected Slovenian language. The proposed techniques can be applied to other Slavic languages, where words are formed by many different inflectional affixatation. In this article an attempt to overcome two important difficulties of highly-inflected languages (high out-of-vocabulary rate and the problem of topic detection) is described. The first problem is solved by the decomposition of words into stems and endings, and topic detection is improved by a novel approach for feature extraction based on soft comparison of words. The results of experiments on the second largest Slovenian newspaper news corpus Vecer show the decrease in perplexity by 17% in average over a general word-based model.

Volume 1, page 243 Session A42

Efficient Stochastic Finite-State Networks for Language Modelling in Spoken Dialogue Systems

Georgila K, Fakotakis N, Kokkinakis G
University of Patras, Greece

In this paper we present a novel method for creating language models for Spoken Dialogue Systems (SDS). The idea is based on combining the linguistic structure and the limited requirements for training data of grammar-based models with the robustness of stochastic models regarding spontaneous speech. Our algorithm requires a set of sentences as input, in order to train a Hidden Markov Model (HMM). Classes containing words or phrases with semantic-syntactic similarities are formed automatically and simultaneously with the construction of the HMM. The states and observations of the HMM correspond to the word/phrase classes and words/phrases respectively. The resulting HMM incorporates grammatical structure provided by large context dependencies as well as coverage of ungrammatical spontaneous sentences provided by statistical estimations. The HMM is transformed to a Stochastic Finite-State Network (SFSN), which allows for variable

history sizes with no specific upper limit. We used data from 3 different SDSs to evaluate the algorithm. The experiments carried out, resulted in precision and recall values regarding the classes formed, of 0.97 and 0.76 in average, respectively. There was also a reduction of perplexity (16.15% in average) compared to bigrams and a gain in recognition performance (keyword accuracy) of 6.2% compared to grammar-based models and 5.4% compared to bigrams.

Volume 1, page 247 Session A42

Language Models Conditioned on Dialog State

Visweswariah K, Printz H
IBM, USA

We consider various techniques for using the state of the dialog in language modeling. The language models we built were for use in an automated airline travel reservation system. The techniques that we explored include (1) linear interpolation with state specific models and (2) incorporating state information using maximum entropy techniques. We also consider using the system prompt as part of the language model history. We show that using state results in about a 20% relative gain in perplexity and about a 9% percent relative gain in word error rate over a system using a language model with no information of the state.

Volume 1, page 251 Session A42

Using Information Retrieval Methods for Language Model Adaptation

Chen L, Gauvain J-L, Lamel L, Adda G, Adda-Decker M
CNRS-LIMSI, France

In this paper we report experiments on language model adaptation using information retrieval methods, drawing upon recent developments in information extraction and topic tracking. One of the problems is extracting reliable topic information with high confidence from the audio signal in the presence of recognition errors. The work in the information retrieval domain on information extraction and topic tracking suggested a new way to solve this problem. In this work, we make use of information retrieval methods to extract topic information in the word recognizer hypotheses, which are then used to automatically select adaptation data from a very large general text corpus. Two adaptive language models, a mixture based model and a MAP based model, have been investigated using the adaptation data. Experiments carried out with the LIMSI Mandarin broadcast news transcription system gives a relative character error rate reduction of 4.3% with this adaptation method.

Volume 1, page 255 Session A42

Session A43 - Oral
Monday - 15.50 - 17.30

Speech Production: Articulation

Chair: Pascal Perrier, ICP - INPG & Université Stendhal, France

Making the Tongue Model Talk: Merging MRI & EMA Measurements

Engwall O
KTH, Sweden

Electromagnetic articulography (EMA) data collected with the Movetrack measurement system has been used to set the parameter values in a three-dimensional tongue model dynamically. The outputs from four of the receiver coils are used in the parameter control; the data from the coil on the lower incisor for the jaw height parameter and the three coils on the tongue, T1-T3, for the parameters of different parts of the tongue. The measurements of T1 control the raising and advancing of the tongue tip, those of T2 the tongue body and those of T3 the tongue dorsum movement. Rules to replicate the measured control sequences synthetically have been developed and the synthetic control sequences have been used to synthesize new fricative-vowel sequences.

Volume 1, page 261 Session A43

The Relationship between Intraoral Air Pressure and Tongue/Palate Contact during the Articulation of Norwegian /t/ and /d/

Moen I, Simonsen H G, Huseby M, Grue J
University of Oslo, Norway

Our paper addresses the question of covariation between intraoral air pressure and size of contact area between tongue and palate during the articulation of the Norwegian stop consonants /t/ and /d/. An EPG investigation of the two plosives shows a larger contact area between tongue and palate for /t/ than for /d/. An investigation of intraoral air pressure during the articulation of the two plosives shows higher air pressure for /t/ than for /d/. Presumably, the covariation between air pressure and contact area between tongue and palate may be accounted for in terms of general phonetic-physiological factors. In order to prevent air from escaping between the tongue and the palate during the closing stage of the plosive, and thus producing a fricative, a larger contact area is needed for the voiceless than for the voiced plosive since the air pressure is stronger for the voiceless than for the voiced plosive.

Volume 1, page 265 Session A43

Mechanical versus perceptual constraints as determinants of articulatory strategy

Elgendy A M, Pols L C W
University of Amsterdam, The Netherlands

This paper summarizes the results of a series of experiments conducted to investigate various aspects of normal pharyngeal articulation and the nature of pharyngeal coarticualtion. Video fiberscopic imaging, electromagnetography and acoustic analysis techniques were used to obtain empirical and quantitative data on the use of the pharynx in speech production. The overall results suggest that mechanical constraints determine to a great extent the articulatory strategy used by the speaker to achieve the perceptual/acoustic contrast essential for the process of speech encoding.

Volume 1, page 269 Session A43

Pre-Liquid Excrescent Schwa: What Happens when Vocalic Targets Conflict

Gick B, Wilson I

University of British Columbia, Canada

Sequences of high tense vowel + liquid in English often result in the percept of an intervening schwa, as in, e.g., heel, hail, hire. We argue in this paper that this apparent schwa is simply the incidental acoustic result of the tongue moving through "schwa-space" (a schwa-like position) during the transition between conflicting tongue root targets. This conflict bears on both articulatory timing relationships in syllable codas and tongue root specification for tense vowels. We present two experiments: Experiment 1 shows that excrescent schwa does not correspond with greater duration of syllable rimes; Experiment 2 shows that the tongue moves through schwa space along its trajectory in the excrescent schwa cases. Our results support a timing model whereby coda timing is determined by the relationship between syllable peak and consonant closure, but where timing is unaffected by the number of intervening vocalic events.

Volume 1, page 273 Session A43

Exploring the Null Space of the Acoustic-to-Articulatory Inversion Using a Hypercube Codebook

Ouni S, Laprie Y
LORIA, France

Our acoustic to articulatory inversion method exploits an original codebook representing the articulatory space by hypercubes. The articulatory space is decomposed into regions where the articulatory-to-acoustic mapping is linear. Each region is represented by a hypercube. The inversion procedure retrieves articulatory vectors corresponding to an acoustic entry from the hypercube codebook. The main issue is about how all the possible inverse solutions in a given hypercube could be found. As the dimension of the articulatory space is greater than the dimension of the acoustic space, the corresponding null space is sampled by linear programming to retrieve all the possible solutions. Indeed, the sampling of the null space is a crucial point because it directly controls the smoothness of articulatory trajectories recovered from the original signal. This approach permits more realistic articulatory trajectories to be obtained.

Volume 1, page 277 Session A43

Session A44 - Oral
Monday - 15.50 - 17.30

Speech Recognition and Understanding: Topic Detection and Information Retrieval

Chair: Steve Renals, Sheffield Univ, United Kingdom

Phoneme-based Topic Spotting on the Switchboard Corpus

Theunissen M W[1], Scheffler K[2], du Preez J A[1]
[1]University of Stellenbosch, South Africa, [2]University of Cambridge, United Kingdom

The field of topic spotting in conversational speech deals with the problem of identifying "interesting" conversations or speech extracts amongst large volumes of speech data. In this research, two phoneme-based topic spotting systems were evaluated on the Switchboard Corpus. Experiments [1,2] on the OGI Corpus suggested that the new Stochastic Method for the Automatic Recognition of Topics (SMART) yields a large improvement over the existing Euclidean Nearest Wrong Neighbours (ENWN) algorithm, which had outperformed competing systems in evaluations [3,4]. However, the small amount of data available for these experiments meant that more rigorous testing was required. We reimplemented the algorithm to run on the larger Switchboard Corpus, and report an improvement of SMART over ENWN characterised by a 35.8% reduction in ROC (receiver operating characteristic) error area. Statistical significance was demonstrated using a modified version of the McNemar test.

Volume 1, page 283 Session A44

Topic Styles in IR and TDT: Effect on System Behavior

Franz M, McCarley J S, Ward T, Zhu W-J
IBM T.J Watson Research Center, USA

The TREC Spoken Document Retrieval Track (SDR) and the Topic Detection and Tracking (TDT) project have annotated the same corpus with difference styles of relevance judgements, using differenct notions of topic. We compare the behavior of a topic tracking system using relevance judgements from TDT with that of the same system using relevance from the SDR in order to investigate the influence of differences document relevance judgements on the behavior of the tracking system.

Volume 1, page 287 Session A44

Extracting Caller Information from Voicemail

Zweig G, Huang J, Padmanabhan M
IBM T.J Watson Research Center, USA

In this paper we address the problem of extracting the identities and phone numbers of the callers in voicemail messages. Previous work in information extraction from speech includes spoken document retrieval and named entity detection. This task differs from the named entity task in that the information we are interested in is a subset of the named entities in the message, and consequently, the need to pick the correct subset makes the problem more difficult. Also, the caller's identity may include information that is not typically associated with a named entity. In this work, we present two information extraction methods, one based on hand-crafted rules, and one based on a maximum entropy model. We find that both systems give good performance when applied to manually-derived transcriptions, and that the maximum entropy system can reliably identify the time intervals containing phone numbers, even in the presence of significant decoding errors.

Volume 1, page 291 Session A44

A Portability Study on Natural Language Call Steering

Kuo H-K J, Lee C-H
Bell Labs, Lucent Technologies, USA

In this paper we examine the portability of the vector-based call router to a new task involving calls to the operator in the UK. One component of the router was shown to require expert knowledge and hand-tuning: the stop word list. Stop word filtering involves replacing certain words with place markers and is necessary to reduce the number of features and parameters used by the classifier. Two specific approaches that eliminates the need for stop word filtering were investigated that led to comparable classification performance: (1) using trigram, bigram, and unigram features and using SVD to reduce the number of parameters, and (2) using only unigram features and applying discriminative training to boost the performance. After discriminative training, the classification error rate was reduced by 18-30% over the baseline unigram results. Increased robustness is demonstrated by a 24-48% reduction in error rate at 20% false rejection rate.

Volume 1, page 295 Session A44

Improved Spoken Document Retrieval by Exploring Extra Acoustic and Linguistic Cues

Chen B, Wang H-M, Lee L-S
Institute of Information Science, Academia Sinica, Taiwan, ROC

In this paper, we explored the use of various extra information to improve the performance of spoken document retrieval (SDR). From the speech recognition perspective, we incorporated the acoustic stress and word confusion information into the audio indexing. From the linguistic perspective, we applied the part-of-speech information in both the audio indexing and the query representation. From the information retrieval perspective, we integrated techniques such as the query expansion by word associations and the blind relevance feedback into the retrieval process. The SDR experiments were based on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). We used the Chinese newswire text stories as query exemplars and the Mandarin Chinese audio news stories as the spoken documents. With all the above acoustic and linguistic cues applied, the average precision was improved from 0.5122 to 0.6312 for the TDT-2 collection and from 0.6216 to 0.7172 for the TDT-3 collection.

Volume 1, page 299 Session A44

Session A45 - Poster
Monday - 15.50 - 17.30

Phonetics and Phonology: Segmentals and Synthesis

Chair: Marc Swerts, Eindhoven University of Technology and Antwerp University, The Netherlands

Native vs Non-native Production of English Vowels in Spontaneous Speech: An Acoustic Phonetic Study

Tsukada K
Macquarie University, Australia

This study aims to examine acoustic characteristics of English vowels produced by 1 Australian English talker and 3 Japanese learners of English in spontaneous speech. Primary stressed vowels in multi-syllabic words were extracted from five 15-minute interview sessions. While there was a considerable overlap between different vowel categories both in native and non-native vowel spaces, centroids were more clearly separated in the former than in the latter. All three Japanese learners' vowel spaces were widely spread in the F2 direction. The Australian talker showed a moderate spectral distinction in two pairs /i - sci/ and /a - invv/. Although this appears contrary to the spectral overlap commonly reported for these pairs in Australian English, it is consistent with the notion that short vowels are more susceptible to reduction than their long counterparts which are less likely to be undershot in various consonantal contexts.

Volume 1, page 305 Session A45

Is Non-Native Pronunciation Modelling Necessary ?

Goronzy S[1], Sahakyan M[1], Wokurek W[2]
[1]Sony International (Europe), Germany, [2]IMS, University of Stuttgart, Germany

It is difficult to recognize non-native speech with speech recognition systems that are trained using native speech. While standard speaker adaptation techniques are often used in theses cases, they are not able to handle severe deviations from the expected pronunciation. Also, there has been a lot of interest in native pronunciation modelling recently. However, results often were not as good as expected. This paper investigates if a special treatment of non-native speakers is necessary. The effect of adding special pronunciation variants to the lexicon is examined. In contrast to native pronunciation modelling the results show that for the non-native case the enhanced dictionary is really necessary to obtain acceptable recognition rates. Recognition rates can be improved by up to 10% for German and even up to 28% for Italian learners of English. When combining this with MLLR adaptation, these results are further improved.

Volume 1, page 309 Session A45

Burst segmentation and evaluation of acoustic cues

Laprie Y, Bonneau A
LORIA, France

This paper investigates burst segmentation for the evaluation of acoustic cues used to identify unvoiced French stops. Unlike other works which utilize a fixed length window, our approach consists in segmenting bursts into transient and frication noise. The transient is found by minimizing the sum of spectral variances of transient and frication noise over the burst. The spectral variance criterion has the advantage of being sensitive both to energy deviations and spectral variations. Additional correction procedures augment the robustness of the segmentation against the presence of spurious noises during the closure and the determination of the voicing onset with delay. The relevance of our segmentation method has been evaluated by comparing the

characteristics of the main spectral peak in the transient segmented by our method with those of the full burst. Our experiments showed that bursts segmented by our method allow a better discrimination between the three places of articulation.

Volume 1, page 313 Session A45

The schwa in Albanian

Granser T, Moosmüller S
Institute of Acoustics, Austrian Academy of Sciences, Austria

The schwa in Albanian Introduction: In Albanian, the schwa as a phoneme is restricted to the Tosk variety (south of the Shkumbin river, generally considered as basis for the standard), whereas it is described as a back, rounded vowel in the Gheg variety (north of the Shkumbin river). Method: Recordings of spontaneous speech of 7 male speakers have been analyzed. The first two formants were calculated. In total, 570 schwa-vowels have been analyzed, 5 articulation zones have been defined. Results: With respect to the realization of the schwa in stressed position, a significant difference could be observed with regard to "within" and "outside" the borders of the Republic of Albania. Within the borders of Albania, however, no differences could be observed. Speakers from all regions display a huge amount of variability, ranging from front to back articulation. In unstressed position, a tendency towards centralization can be observed.

Volume 1, page 317 Session A45

A Testbed for Developing Multilingual Phonotactic Descriptions

Ashby S, Carson-Berndsen J, Joue G
University College Dublin, Ireland

This paper presents a testbed for developing multilingual phonotactic descriptions that employs finite state methods to represent the phonotactics of one or more languages. The motivation for this work is to make an extensive range of phonotactic descriptions of varying granularity available for speech technology applications. We discuss the design of the phonotactic testbed and how various modules may be used to generate finite state phonotactic descriptions. We provide an example multilingual application drawn from a partial sample of onset clusters spanning four language families, demonstrating how the commonalities of a broad spectrum of languages can be expressed using individual and generic phonotactic automata. We then discuss how these representations are extended via a three-tiered model to provide the basis for the feature- and event-based phonotactic automata.

Volume 1, page 321 Session A45

A Physiological Analysis of Nasals and Nasalization in Chinese

Fung W-N, Lau S-L
City University of Hong Kong, Hong Kong

This paper is a physiological investigation of the vowel nasalization in Chinese by analyzing the nasal and oral airflows for the (C):VN and (C)VN syllables in Shanghai (SH) and Hong Kong Cantonese (HKC). Results show that (i) the degree of nasalization is inversely correlated with the tongue height of the vowel followed by a syllable-final nasal in both SH and HKC; (ii) the duration of nasalization is positively correlated with the advancement of the oral closure for the syllable-final nasal in both SH and HKC; (iii) in general, 10% - 60% of the vowel is nasalized when followed by a nasal ending in SH, except for the schwa, the low-mid back vowel and the low vowel followed by a velar nasal; and (iv) in SH, the schwa, the low-mid back vowel and the low vowel followed by a velar nasal are fully nasalized.

Volume 1, page 325 Session A45

A Component by Component Listening Test Analysis of the IBM Trainable Speech Synthesis System

Donovan R E
IBM TJ Watson Research Center, USA

This paper reports on a listening test conducted to determine the impact on speech quality of each component in the IBM Trainable Speech Synthesiser. The study was originally conceived to direct future research effort to those components with the greatest potential for improvement. However, the results and conclusions regarding prosodic modification, concatenation unit length, and decision tree clustering are generally applicable and may be of wider interest.

Volume 1, page 329 Session A45

Semantic Abnormality and its Realization in Spoken Language

Pan S[1], McKeown K[2], Hirschberg J[3]
[1]IBM TJ Watson Research Center, USA, [2]Columbia University, USA, [3]AT&T Labs-Research, USA

In this paper we investigate the relationship between various lexical and prosodic features and 'semantic abnormality', the occurrence of unusual or unexpected events, in generating speech for MAGIC, which employs a Concept-to-Speech system to generate post-operative reports for patients who have undergone bypass surgery. Using the speech corpus collected for this application, we conducted empirical analysis to systematically discover significantly correlated prosodic and lexical features. The automatically learned abnormality model not only can be used in building comprehensive prosody prediction systems for Concept-to-Speech generation, but also help identify unusual information during speech analysis and understanding.

Volume 1, page 333 Session A45

TALKING FOREIGN - Concatenative Speech Synthesis and the Language Barrier

Campbell N
ATR Spoken Language Trranslation, Japan

This paper presents a solution to the problem of synthesising multi-lingual speech using waveform-concatenation speech synthesis. It presents methods for mapping the pronunciations of a target-language speaker onto the sounds available in the speech corpus of a native speaker so that the resulting synthesis produces speech which can accurately represent any foreign words encountered in a predominantly native-language text by use of multi-speaker synthesis. The methods differ depending on the language-pair and on the direction of the mapping, because in the case of one-to-many phonemic mappings, high-level features can be used, but in the many-to-one case, a physical representation of the target speech signal is required. All mappings are automatic, and the use of rule-based procedures is minimised. In this way, the methods are extensible to any language combinations. Synthesised speech samples are included with the paper so that a subjective evaluation of the results can be made.

Volume 1, page 337 Session A45

Schwa-assimilation in Danish Synthetic Speech

Jensen C
University of Copenhagen, Denmark

Assimilation of schwa into surrounding sonorant consonants is a vital feature of natural Danish speech. It varies with speaking rate and speaking style and is more likely to occur in some phonological contexts than in others. This presents some problems for the implementation of the process into a Danish text-to-speech system.

Volume 1, page 341 Session A45

Text-to-speech synthesis with arbitrary speaker's voice from average voice

Tamura M[1], Masuko T[1], Tokuda K[2], Kobayashi T[1]
[1]Tokyo Institute of Technology, Japan, [2]Nagoya Institute of Technology, Japan

This paper describes a technique for synthesizing speech with any desired voice. The technique is based on an HMM-based text-to-speech (TTS) system and MLLR adaptation algorithm. To generate speech of an arbitrarily given target speaker, speaker-independent speech units, i.e., average voice models, is adapted to the target speaker using MLLR framework. In addition to spectrum and pitch adaptation, we derive an algorithm for adaptation of state duration. We demonstrate that a few sentences uttered by a target speaker are sufficient to adapt not only voice characteristics but also prosodic features. Synthetic speech generated from adapted models using only four sentences is very close to that from speaker dependent models trained using a large amount of speech data.

Volume 1, page 345 Session A45

High Quality Voice Conversion Based on Gaussian Mixture Model with Dynamic Frequency Warping

Toda T, Saruwatari H, Shikano K
Nara Institute of Science and Technology, Japan

In the voice conversion algorithm based on the Gaussian Mixture Model (GMM), quality of the converted speech is degraded because the converted spectrum is exceedingly smoothed. In this paper, we newly propose the GMM-based algorithm with the Dynamic Frequency Warping (DFW) to avoid the over-smoothing. We also propose that the converted spectrum is calculated by mixing the GMM-based converted spectrum and the DFW-based converted spectrum, to avoid the deterioration of conversion-accuracy on speaker individuality. Results of the evaluation experiments clarify that the converted speech quality is better than that of the GMM-based algorithm, and the conversion-accuracy on speaker individuality is the same as that of the GMM-based algorithm in the proposed algorithm with the proper weight for mixing spectra.

Volume 1, page 349 Session A45

Voice Transformations: From Speech Synthesis to Mammalian Vocalizations

Tang M, Wang C, Seneff S
MIT Laboratory for Computer Science, USA

This paper describes a phase vocoder based technique for voice transformation. This method can flexibly manipulate various aspects of the input signal, e.g., pitch, duration, energy, and formant positions, without explicit F0 extraction. The modifications can be specific to any feature dimensions, and can vary over time. There are many potential applications for this technique. In concatenative speech synthesis, it can be applied to transform the voice characteristic of the speech corpus, or to smooth pitch or formant discontinuities between concatenation boundaries. The method can also be used in language learning. We can modify the prosody of the student's speech to match that from a native speaker, and use the result to guide improvements. The technique can also be used to convert other biological signals, such as killer whale vocalizations, to ones that are more appropriate for human auditory perception. Our experiments show encouraging results for all of these applications.

Volume 1, page 353 Session A45

A new Multi-Speaker Formant Synthesizer that applies Voice Conversion Techniques

Gutiérrez-Arriola J M, Montero J M, Vallejo J A, Córdoba R, San-Segundo R, Pardo J M
Universidad Politécnica de Madrid, Spain

We present a multi-speaker formant synthesizer based on parameter concatenation. The user can choose among three speakers, two males and one female. The synthesizer stores all the parameters for the basic speaker and linear transformation functions to synthesized the other two. The complete database for one speaker consists of 455 parameterized units (diphones, triphones,...) and the parameters used are pitch, formants and bandwidths and source parameters (four parameters for the LF model, and glottal noise). To get the converted speaker we store a linear transformation function for each spectral stable segment of each unit. Preliminary results show that the quality of the synthesizer is very good and that this system can help us to study and understand the speaker variability problem.

Volume 1, page 357 Session A45

Evaluation of Cross-Language Voice Conversion Based on GMM and Straight

Mashimo M[1], Toda T[1], Shikano K[1], Campbell N[2]
[1]Nara Institute of Science and Technology, Japan, [2]ATR Information Sciences Deivision, Japan

Voice conversion is a technique for producing utterances using any target speakers' voice from a single source speaker's utterance. In this paper, we apply cross-language voice conversion between Japanese and English to a system based on a Gaussian Mixture Model (GMM) method and STRAIGHT, a high quality vocoder. To investigate the effects of this conversion system across different languages, we recorded two sets of bilingual utterances and performed voice conversion experiments using a mapping function which converts parameters of acoustic features for a source speaker to those of a target speaker. The mapping functions were trained using bilingual databases of both Japanese and English speech. In an objective evaluation using Mel cepstrum distortion (Mel CD), it was confirmed that the system can perform cross-language voice conversion with the same performance as that within a single-language.

Volume 1, page 361 Session A45

Ejective Reduction in Chaha is Conditioned by More Than Prosodic Position

Coulston R
Oregon Graduate Institute / University of California San Diego, USA

This paper examines a neutralization asymmetry in Chaha ejectives, concluding that reduction is conditioned not by prosodic position alone, but also by place and manner of articulation. An acoustic examination of Chaha, a Gurage dialect of the Ethiopian Semitic language family, shows that its velar ejectives display a much stronger tendency to lose burst cues before another ejective than do alveolar ejectives in the same environment. This pattern of laryngeal neutralization provides important support for phonetically informed phonological theories. Purely prosody-based theories cannot account for this behavior but a viable alternative is found in a cue-based approach.

Volume 1, page 365 Session A45

Session A46 - Poster
Monday - 15.50 - 17.30

Speech Perception: Miscellaneous - II

Chair: William Ainsworth, Mackay Institute of Communication & Neuroscience, United Kingdom

Effects of Noise Adaptation on the Perception of Voiced Plosives in Isolated Syllables

Ainsworth W[1], Cervera T[2]
[1]Keele University, UK, [2]University of Valencia, Spain

Speech is easier to understand in continuous noise than in noise which is switched on at the beginning of the speech and off at the end. It is suggested that this is due to some adaptation process. In order to test this hypothesis a series of experiments have been performed in which the intelligibility of plosives in isolated syllables was measured as a function of the duration of the preceding noise. The spectrum of the noise was also varied. It was found that the adaptation, as measured by the mean increase in intelligibility, increased as the duration of the noise preceding the syllable was lengthened. It was also found that the adaptation varied with the centre frequency of the spectrum of the noise. The amount of adaptation was negatively correlated with the threshold of hearing.

Volume 1, page 371 Session A46

On Differential Limen of Word-based Local Speech Rate Variation in Japanese Expressed by Duration Ratio

Hiroshige M, Araki K, Tochinai K
Hokkaido University, Sapporo, Japan

Fundamental studies about differential limen (DL) for word-based speech rate variations in Japanese are described. In our previous study, the DLs are expressed by subtractive difference of mora duration. In this report, however, to fit the expression for various global speech rate, the DLs are expressed by variation ratio of mora duration. We carry out auditory tests with stimuli made by equally lengthening or shortening a duration of a word in a sentence. The subjects' focus of attention is diffused to get DLs that are used in the normal natural conversations. The obtained DLs are approximately 0.85 for acceleration and 1.18 for deceleration in variation ratio of mora duration.

Volume 1, page 375 Session A46

A Multidimensional Scaling Study of Fricatives; a Comparison of Perceptual and Physical Dimensions

Tokuma W
Seijo University, Japan

This study attempts to model the perceptual similarity data of natural English voiceless fricative syllables in terms of the auditory distance metrics using multidimensional scaling technique (MDS). First, it was proved that the perceptual configuration of nonspeech sounds is adequately modelled by Euclidean distance space. Next, the three distance metrics were analysed to show which metric is most efficient in modelling the perception of speech sounds. Finally, it was shown that the perceptual and physical configurations of five voiceless English fricatives were highly correlated. This result seems to support a model of speech perception based mainly on the general physical characteristics of speech.

Volume 1, page 379 Session A46

Reconstructing Dialogue History

Swerts M[1], Krahmer E[2]
[1]IPO, The Netherlands / CNTS, Belgium, [2]IPO, The Netherlands

This paper deals with a perceptual analysis of accent structure in Dutch to see to what extend listeners are able to reconstruct information from the previous discourse on the basis of prosodic properties of the current utterance. Using data collected in an an earlier dialogue game experiment, subjects were asked to perfom a perceptual task in which they had to reconstruct what the previous utterance was on the basis of input utterances with different accent patterns. Our results reveal that listeners are able to correctly guess the prior context for a significant number of cases, but that performance depends on the type of intonation contour of the input utterance.

Volume 1, page 383 Session A46

Timing and Interaction of Visual Cues for Prominence in Audiovisual Speech Perception

House D, Beskow J, Granström B
KTH, Sweden

The timing of both eyebrow and head movements of a talking face was varied systematically in a test sentence using an audiovisual speech synthesizer. The audio speech signal was unchanged over all sentences. 33 listeners were given the task of identifying the most prominent word in the test sentence. Results indicate that both eyebrow and head movements are powerful visual cues for prominence and that perceptual sensitivity to timing is on the order of a typical syllable duration of 100-200 ms.

Volume 1, page 387 Session A46

Modelling the Perceptual Identification of Japanese Consonants from LPC Cepstral Distances

Komatsu M[1], Tokuma S[2], Tokuma W[3], Arai T[4]
[1]Sophia University, Japan / University of Alberta, Canada, [2]Sagami Women's University, Japan, [3]Seijo University, Japan, [4]Sophia University, Japan

This study attempts to account for the perceptual phenomenon observed in Komatsu et al. [1] in terms of the spectral properties of the LPC re-synthesised stimuli. To implement this, LPC cepstral distances between re-synthesised samples and their original samples are measured. The results of the acoustic analysis and their comparison with the perceptual data indicate that there is a striking similarity in patterns between the spectral property of the Japanese consonants and their perceptual scores. This suggests that the role played by spectral information in the perception of Japanese consonants is significant across all consonant types, and also implies that even in its crudest form, it contributes significantly to their perception.

Volume 1, page 391 Session A46

Auditory-Visual Perception of Lexical Tone

Burnham D[1], Ciocca V[2], Stokes S[2]
[1]University of Western Sydney, Australia, [2]University of Hong Kong, Hong Kong

Cantonese speakers were asked to identify spoken words as one of six Cantonese words differing only in tone. Words were presented in three modes: auditory-visual (AV), auditory only (AO), and visual only (VO). Performance was equivalent in the AO and AV conditions - there was no augmentation of auditory tone perception when visual information was added. Nevertheless, performance in the VO condition was significantly above chance under certain conditions: for perceivers without phonetic training, but not those with phonetic training; for tone carried on monophthongs, but not diphthongs; for tones spoken in running speech, but not citation form; and for contour tones (involving pitch movement over time), but not level tones (involving minimal pitch movement).

Thus there is visual information for tone which is functionally relevant under certain circumstances.

Volume 1, page 395 Session A46

Syllable prominence: A matter of vocal effort, phonetic distinct-ness and top-down processing

Eriksson A, Thunberg G C, Traunmüller H
Stockholm University, Sweden

In this experiment, subjects had to rate the "prominence" of each of the syllables of 20 versions of the same utterance produced by men, women and children at various levels of vocal effort. The ratings were correlated with measurements of the SPL of the fundamental, spectral emphasis, vowel duration, F0max and F0 rise from the previous syllable. Together with ratings of the perceived vocal effort at which the utterances had been produced, these measurements were used to obtain the possible contributions of vocal effort, prosodic distinctness, and vowel duration to the perceived prominence. Together, these accounted for half of the variance. This was compared with the possible contribution of the linguistic structure of the utterance, which accounted for slightly more of the variance. The predictions of a model based on this analysis came closer to the mean than the average subject.

Volume 1, page 399 Session A46

Perceived Prominence in Terms of a Linguistically Motivated Quantitative Intonation Model

Mixdorff H[1], Widera C[2]
[1]Berlin University of Applied Sciences, Germany, [2]University of Bonn, Germany

The current study investigates the relationship between perceived syllable prominence and the F0 contour as described by a linguistically motivated model of German intonation based on the Fujisaki formula. A subcorpus of the Bonn Prosodic Database was analyzed using the F0 model, and normalized log syllable durations calculated. Analysis shows that, for accented syllables, prominences strongly correlate with the amplitude Aa of accent commands underlying the F0 movements in these syllables, whereas comparable F0 movements in unaccented syllables have little effect on prominence. The influence of Aa versus syllable duration on prominence is stronger for higher prominence levels. The fact that the F0 movement does not necessarily take place in the accented syllable proper, indicates that prominence judgment is partly guided by linguistic considerations. The results also show that F0 modeling in TTS needs to be especially accurate in accented syllables which supports the main rationale of the F0 model.

Volume 1, page 403 Session A46

Perception of Coda Voicing from Properties of the Onset and Nucleus of 'led' and 'let'

Hawkins S[1], Nguyen N[2]
[1]University of Cambridge, UK, [2]Université de Provence, France

Syllable-onset /l/ in British English is longer and often has different (usually lower) F2 frequency before a voiced coda. Five experiments explore the perceptual power of these properties and of f0. In each experiment, listeners identified as 'led' or 'let' synthetic syllables whose latter half was replaced by noise. The most reliable cue was /l/ duration; F2 frequency in the /l/ was influential mainly when the vowel quality was held constant. However, listeners learn which cues are most effective, and some choose /l/ duration rather than spectral properties relatively late in the procedure. The results support word recognition models with non-segmental lexical representation that is sensitive to systematic variation in phonetic fine detail.

Volume 1, page 407 Session A46

Auditory Filter Bank Design Using Masking Curves

Lin L L, Ambikairajah E, Holmes W H
The University of New South Wales, Australia

It is very difficult and costly to experimentally observe the motion of the basilar membrane in a fully functional cochlea with the view to obtaining amplitude response at points along the membrane. This paper presents an inexpensive method of generating psychoacoustic tuning curves from the well-known masking curves in critical band rate. We present a method for designing critical band auditory filters from the tuning curves. It is also known that the auditory filter frequency response becomes broader with increasing input signal levels and becomes narrower with decreasing signal levels. We also propose a method for designing level dependent auditory filters. The proposed filter bank is applicable to various types of signal processing required to model human auditory filtering.

Volume 1, page 411 Session A46

A New Feature Driven Cochlear Implant Speech Processing Strategy

Dashtseren E, Kitazawa S, Kitamura T
Shizuoka University, Japan

Our study focuses on the development of a new feature driven speech-processing strategy for cochlear implant system. In each cycle of stimulation of the cochlea, an electrode, corresponding to second formant frequency was chosen among 14 basilar electrodes. On the base of voiced/unvoiced decision of the speech, an electrode corresponding to first formant frequency was selected for stimulation among 6 apex electrodes. Additionally 4-6 electrodes were chosen for stimulation by maxima energy criteria. Speech intelligibility tests on multi syllable Japanese words within normal hearing listeners by acoustic simulation were provided to evaluate performance of the proposed strategy. Key words: Cochlear implant system, speech feature, acoustic simulation, speech intelligibility test.

Volume 1, page 415 Session A46

Session B11 - Oral
Tuesday - 09.00 - 10.40

ESE2 - Noise Robust Recognition: Front-end Algorithms

Chair: Hans-Guenter Hirsch, University of Applied Sciences Niederrhein

Acoustic Feature Compensation Based on Decomposition of Speech and Noise for ASR in Noisy Environments

Kim H K, Rose R C, Kang H-G
AT&T Labs-Research, USA

This paper presents a set of acoustic feature pre-processing techniques that are applied to improving automatic speech recognition (ASR) performance on the Aurora 2 noisy speech recognition task. The principal contribution of this paper is an approach for cepstrum domain feature compensation in ASR which is motivated by techniques for decomposing speech and noise that were originally developed for noisy speech enhancement. This approach is applied in combination with other feature compensation algorithms to compensating ASR features obtained from a mel-filterbank cepstrum coefficient (MFCC) front-end. Performance comparisons are made with respect to the application of the minimum mean squared error log spectral amplitude estimator (MMSE-LSA) based speech enhancement algorithm prior to feature analysis. An experimental study is presented where the feature compensation approaches described in the paper are found to reduce ASR word error rate by as much as 31% relative to uncompensated features under simulated environmental and channel mismatched conditions.

Volume 1, page 421 Session B11

A Robust Front-End Algorithm for Distributed Speech Recognition

Cheng Y M, Macho D, Wei Y, Ealey D, Kelleher H, Pearce D, Kushner W, Ramabadran T
Motorola Labs, USA

This paper presents the robust front-end algorithm that was submitted by Motorola to the ETSI STQ-Aurora DSR working group as a proposal for the Advanced DSR front-end in January 2001. The algorithm consists of a two-stage mel-warped Wiener filter, a waveform processor, a channel-normalized mel-frequency cepstral calculation and a subsystem of post-cepstral processing according to the reliability of mel-spectral components, etc. The output of this algorithm, a set of Mel-Frequency Cepstral Coefficients (MFCC), is compressed, encoded and transmitted at 4800 bps. Compared to ETSI standard MFCC front-end, the proposed algorithm delivers an improvement of 47.58% on the Aurora 2 database, which is required by this Eurospeech special event. In this paper we also give further insights about the proposal by providing performances and analyses with the Aurora SpeechDat-Car databases.

Volume 1, page 425 Session B11

Robust ASR front-end using spectral-based and discriminant features: experiments on the Aurora tasks

Benitez C[1], Burget L[2], Chen B[1], Dupont S[1], Garudadri H[3], Hermansky H[2], Jain P[2], Kajarekar S[2], Sivadas S[2]
[1]International Computer Science Institute, USA, [2]Oregon Graduate Institute Of Science and Technology, USA, [3]Qualcomm Inc., USA

This paper describes an automatic speech recognition front-end that combines low-level robust ASR feature extraction techniques, and higher-level linear and non-linear feature transformations. The low-level algorithms use data-derived filters, mean and variance normalization of the feature vectors, and dropping of noise frames. The feature vectors are then linearly transformed using Principal Components Analysis (PCA). An Artificial Neural Network (ANN) is also used to compute features that are useful for classification of speech sounds. It is trained for phoneme probability estimation on a large corpus of noisy speech. These transformations lead to two feature streams whose vectors are concatenated and then used for speech recognition. This method was tested on the set of speech corpora used for the Aurora evaluation. Using the feature stream generated without the ANN yields an overall 41% reduction of the error rate over Mel-Frequency Cepstral Coefficients (MFCC) reference features. Adding the ANN stream further reduces the error rate yielding a 46% reduction over the reference features.

Volume 1, page 429 Session B11

Noise Reduction for Noise Robust Feature Extraction for Dis-tributed Speech Recognition

Noé B[1], Sienel J[1], Jouvet D[2], Mauuary L[2], de Veth J[3], Boves L[3], de Wet F[3]
[1]Alcatel SEL, Stuttgart, Germany, [2]France Télécom R&D, DIH/IPS, France, [3]University of Nijmegen, The Netherlands

This paper describes the noise robust feature extraction meth-ods developed by France Telecom and Alcatel for the noise robust front-end standardisation of ETSI Aurora. It is shown that both noise reduction methods give a substantial improvement when compared to a standard MFCC feature extraction algorithm for speech recognition in noisy environments. In addition, blind equalisation and feature vector selection were used for further improvement of recognition performance. Results are discussed for the ETSI Aurora 2 task and the SDC-Italian task as well. It was found that the combination of noise reduction with the proposed methods is capable to achieve around 50% reduction of the error rate. In the context of the open ETSI Aurora standardisation, two propos-als were submitted based on these methods, they achieved the best results among all the proposals.

Volume 1, page 433 Session B11

Harmonic tunnelling: tracking non-stationary noises during speech

Ealey D, Kelleher H, Pearce D
Motorola Labs, UK

This paper presents a novel noise robust front-end algorithm, evaluating its performance on the Aurora 2 database. Most noise robust algorithms for speech recognition assume stationary noise, i.e. that a noise estimate taken prior to the utterance will be accurate for the duration of that utterance. However, for non-stationary noises wherein the noise spectrum can change during the utterance, there can be substantial differences between the estimated and actual noise spectra for a given frame, resulting in poor performance. The algorithm presented here provides a continuous estimate of the noise, making use of the structure of the voiced speech spectrum to sample the gaps (or “tunnels”) between the harmonic spectral peaks. Compared to the ETSI standard MFCC front-end, the proposed algorithm delivers an average improvement in performance of 43.93% on the Aurora 2 database.

Volume 1, page 437 Session B11

Session B12 - Oral
Tuesday - 09.00 - 10.40

Linguistic Modelling: Semantic Modelling

Chair: Elmar Noeth, Universitât Erlangen-Nürnberg, Germany

Resource-Limited Sentence Boundary Detection

Carter D, Gransden I
Speech Machines, UK

We examine the practical constraints imposed on the task of sentence boundary detection in speech recognizer output, by the requirements of a system that supports large-scale commercial off-line transcription of dictations. We develop and evaluate a method that observes these constraints, reformulating the best technique previously reported in order to allow the use a smoothing technique directly tailored to boundary prediction. We then show how this method can be generalized and improved upon, demonstrating significantly better performance in three different domains.

Volume 1, page 443 Session B12

Metrics for Measuring Domain Independence of Semantic Classes

Pargellis A, Fosler-Lussier E, Potamianos A, Lee C-H
Bell Labs, Lucent Technologies, USA

The design of dialogue systems for a new domain requires semantic classes (concepts) to be identified and defined. This process could be made easier by importing relevant concepts from previously studied domains to the new one. We propose two methodologies, based on comparison of semantic classes across domains, for determining which concepts are domain-independent, and which are specific to the new task. The concept-comparison technique uses a context-dependent Kullback-Leibler distance measurement to compare all pairwise combinations of semantic classes, one from each domain. The concept-projection method uses a similar metric to project a single semantic class from one domain into the lexical environment of another. Initial results show that both methods are good indicators of the degree of domain independence for a wide range of concepts, manually generated for three different tasks: Carmen (children's game), Movie (information retrieval) and Travel (flight reservations).

Volume 1, page 447 Session B12

Context-dependent Probabilistic Hierarchical Sub-lexical Modelling Using Finite State Transducers

Mou X, Seneff S, Zue V
MIT Laboratory for Computer Science, USA

This paper describes a unified architecture for integrating sub-lexical models with speech recognition, and a layered framework for context-dependent probabilistic hierarchical sub-lexical modelling using finite state transducers. Our major motivation for designing a unified architecture is to provide a framework such that probabilistic sub-lexical linguistic components can be integrated with other speech recognition components without sacrificing the flexibilities of their independent developments and configurations. We are also able to obtain a tightly coupled interface between recognizers and sub-lexical components. We present a view of using layered probabilistic models to augment context-free grammars (CFGs). It captures context-dependent probabilistic information beyond the standard CFG formalism, and provides the flexibility of developing suitable probabilistic models independently for each sub-lexical layer. Experimental results show that such an approach can achieve comparable performance to pronunciation network approaches on in-vocabulary utterances, while being able to substantially reduce errors on utterances with previously unseen words.

Volume 1, page 451 Session B12

Data-Driven Semantic Inference for Unconstrained Desktop Command and Control

Bellegarda J, Silverman K
Apple Computer, USA

At ICSLP'00, we introduced the concept of data-driven semantic inference, an approach to command and control which in principle allows for any word constructs in command/query formulation. Unconstrained word strings are mapped onto the relevant action through an automated classification relying on latent semantic analysis: as a result, it is no longer necessary for users to memorize the exact syntax of every command. The objective of this paper is to further characterize the behavior of semantic inference, using a desktop command and control task involving 113 different actions. We consider various training scenarios of increasing scope to assess the influence of coverage on performance. Under realistic usage conditions, good classification results can be obtained at a level of coverage as low as 70%.

Volume 1, page 455 Session B12

Information Extraction via Heuristics for a Movie Showtime Query System

Jansche M
The Ohio State University, USA

Semantic interpretation for limited-domain spoken dialogue systems often amounts to extracting information from utterances. For a system that provides movie showtime information, queries are classified along four dimensions: question type, and movie titles, towns and theaters that were mentioned. Simple heuristics suffice for constructing highly accurate classifiers for the latter three attributes; classifiers for the question type attribute are induced from data using features tailored to spoken language phenomena. Since separate classifiers are used for the four attributes, which are not independent, certain errors can be detected and corrected, thus increasing robustness.

Volume 1, page 459 Session B12

Session B13 - Oral
Tuesday - 09.00 - 10.40

Speech Perception: Recognition and Intelligibility

Chair: To be decided,

Recognition of (Almost) Spoken Words: Evidence from Word Play in Japanese

Otake T[1], Cutler A[2]
[1]Dokkyo University, Japan, [2]Max Planck Institute for Psycholinguistics, The Netherlands

Current models of spoken-word recognition assume automatic activation of multiple candidate words fully or partially compatible with the speech input. We propose that listeners make use of this concurrent activation in word play such as punning. Distortion in punning should ideally involve no more than a minimal contrastive deviation between two words, namely a phoneme. Moreover, we propose that this metric of similarity does not presuppose phonemic awareness on the part of the punster. We support these claims with an analysis of modern and traditional puns in Japanese (in which phonemic awareness in language users is not encouraged by alphabetic orthography). For both data sets, the results support the predictions. Punning draws on basic processes of spoken-word recognition, common across languages.

Volume 1, page 465 Session B13

Perceptual experiments on enhanced and slowed down speech sentences for second language acquisition

Colotte V, Laprie Y, Bonneau A
LORIA, France

This paper investigates the perception of speech signals that have been enhanced and slowed down selectively, with the view of improving oral comprehension for second language acquisition. Our modifications are applied on a small number of acoustic cues, i.e. bursts of unvoiced stops, unvoiced fricative noises and rapid spectral transition regions. Bursts and frication noises were amplified, and spectral transitions were amplified and slowed down. We exploit energy and spectral criteria to localize bursts and frication noises, and spectral variation function to spot rapid transitions. The perception experiment involved students who learn French as a foreign language. The subjects were asked to fill in gaps in incomplete transcriptions of 50 French sentences. The average identification rate increases from 72% up to 81% when the enhancement is applied alone, and up to 86% when the two modifications are applied simultaneously. The strengths of our approach are the robustness of acoustic cue detection and the fully automatic strategy.

Volume 1, page 469 Session B13

The Relation Between Speech Intelligibility and the Complex Modulation Spectrum

Greenberg S[1], Arai T[2]
[1]International Computer Science Institute, USA, [2]Sophia University, Japan

The amplitude and phase components of the modulation spectrum were dissociated in order to ascertain the importance of cross-spectral, envelope-modulation phase information for understanding spoken language. The dissociation was effected via local time reversals of the speech waveform (i.e., flipping the signal on its horizontal axis) at intervals ranging between 0 and 180 ms. Intelligibility declines progressively as the length of the time-reversed segment increases, down to an asymptotic trough in performance at 100 ms (4% of the words correct). Intelligibility does not correlate highly with the amplitude component of the modulation spectrum, but does coincide closely with the contour of the complex modulation spectrum, a representation that integrates the cross-spectral modulation phase and the conventional (amplitude-based) modulation spectrum into a unified representation. The results imply that intelligibility is based on both the phase and amplitude components of the modulation spectrum.

Volume 1, page 473 Session B13

Envelope Information in Speech Processing: Acoustic-Phonetic Analysis vs. Auditory Figure-Ground Segregation

Crouzet O, Ainsworth W A
Keele University, UK

Long-term envelope modulations (<100Hz) influence the identification of speech in noise. It is not clear, however, whether this influence only takes place at the level of acoustic-phonetic analysis (phonetic identification) or if envelope fluctuations may also help in auditory figure-ground segregation (e.g. separation of speech from concurrent backgrounds). An experiment is presented in which the influence of long-term envelope modulations was investigated using signals mixed with either stationary or temporally modulated noise. The better performance observed when processing speech in modulated background may be related to the listeners' ability to use envelope information in trying to follow concurrent signals independently. It is therefore predicted that, if long-term envelope modulations help to segregate speech from noisy backgrounds, this effect should be stronger when envelope information is fully available.

Volume 1, page 477 Session B13

A comparison between human vowel normalization strategies and acoustic vowel transformation techniques

Adank P, van Hout R, Smits R
University of Nijmegen, The Netherlands

Perceptual and acoustic representations of vowel data were compared directly to evaluate the perceptual relevance of several speaker normalization transformations. The acoustic representations consisted of raw F0 and formant data. The perceptual representations were obtained through an experimental procedure, with phonetically trained listeners as subjects who had to judge vowel quality on three continuous scales: vowel height, vowel advancement and vowel rounding or spreading. The raw acoustic data were transformed according to several normalization schemes. The perceptual and the acoustic representations were compared using regression techniques. A zscore-transformation of the raw data appeared to resemble the perceptual data.

Volume 1, page 481 Session B13

Session B14 - Oral
Tuesday - 09.00 - 10.40

Speech Recognition and Understanding: LVCSR - I

Chair: Xavier Aubert, Philips Research Laboratories Aachen, Germany

On Large Vocabulary Continuous Speech Recognition of Highly Inflectional Language - Czech

Ircing P[1], Krbec P[2], Hajic J[2], Psutka J[1], Khudanpur S[3], Jelinek F[3], Byrne W[3]
[1]University of West Bohemia in Pilsen, Czech Republic, [2]MFF UK, Czech Republic, [3]Johns Hopkins University, USA

A system for large vocabulary continuous speech recognition of highly inflectional language is introduced. Word-based recognition approach is compared with a morpheme-based recognition system. An experiment involving Czech N-best rescoring has been performed with encouraging results.

Volume 1, page 487 Session B14

Towards Automatic Transcription of Spontaneous Presentations

Shinozaki T, Hori C, Furui S
Tokyo Institute of Technology, Japan

This paper reports various investigations on recognizing spontaneous presentation speech in connection with the gSpontaneous Speech h national project started in 1999. Presentation speech uttered by 10 male speakers of approximately 4.5 hours duration has been recognized. Experimental results show that acoustic and language modeling based on an actual spontaneous speech corpus is far more effective than conventional modeling based on read speech. The recognition accuracy has a wide speaker-to-speaker variability according to the speaking rate, the number of fillers, the number of repairs, etc. It was confirmed that unsupervised speaker adaptation of acoustic models was effective to improve the recognition accuracy. The recognition accuracy for spontaneous speech is, however, still rather low, and there remains a large number of research issues.

Volume 1, page 491 Session B14

A Real-Time Japanese Broadcast News Closed-Captioning System

Siohan O[1], Ando A[2], Afify M[1], Jiang H[1], Lee C-H[1], Li Q[1], Liu F[1], Onoe K[2], Soong F K[1], Zhou Q[1]
[1]Bell Labs - Lucent Technologies, USA, [2]NHK - Science and Technical Research Laboratories, Japan

This paper describes a collaboration between Bell Labs and NHK (Japan Broadcasting Corp.) STRL to develop a real-time large vocabulary speech recognition system for live closed-captioning of NHK news programs. Bell Labs broadcast news recognition engine consists of a two-pass decoder using bigram language models (LM) and right biphone models during the first pass, and trigram LM with within-word triphone models in the second pass. Various pruning strategies are used to achieve real time decoding, together with a noise compensation procedure aimed at improving recognition on noisy segments of the program. The system operates in a real-time mode and delivers less than 2% of word error rate (WER) on studio news conditions and about 5% of WER on noisy news and reporter speech when evaluated on a real broadcast news program.

Volume 1, page 495 Session B14

Investigations on Conversational Speech Recognition

Beyerlein P, Aubert X, Harris M, Meyer C, Schramm H
Philips Research Laboratories Aachen, Germany

Automatic speech recognition of real-life conversational speech is a precondition for building natural human-centered man-machine interfaces. Being able to extract speech utterances from real-life broadcast news audio streams and transcribing them with an overall word accuracy of 83% we are still faced with the problem of transcribing true conversational speech in real-life (i.e. bad) background conditions. The switchboard task focusses on the latter problem. The paper summarizes a set of experimental investigations on the switchboard corpus using the Philips LVCSR system.

Volume 1, page 499 Session B14

Recent Advances in Speech Recognition System for IBM DARPA Communicator

Gao Y, Erdogan H, Li Y, Goel V, Picheny M
IBM T. J. Watson Research Center, USA

In this paper, we present methods to improve speech recognition performance of the IBM DARPA Communicator system. Our efforts for acoustic modeling include training a domain specific yet broad acoustic model, speaker clustering and speaker adaptation using feature space transforms. For language modeling, we achieved improvements by using compound words, carefully designed LM classes and adjusting the within class probabilities, using NLU state information to enhance the language model and building a language model with embedded grammar objects. Our efforts produced a relative error rate reduction of 34.6% on the test set that consists of 1173 utterances that IBM received during the NIST evaluation of the DARPA Communicator systems in June 2000. We also tested our decoding on the data from some other sites to further demonstrate the robustness of the system improvements.

Volume 1, page 503 Session B14

Session B15 - Poster
Tuesday - 09.00 - 10.40

Speech Synthesis: Systems and Prosody

Chair: Ann Syrdal, AT&T Research Labs, USA

Festival Speaks Italian!

Cosi P[1], Tesser F[2], Gretter R[2], Avesani C[1], Macon M[3]
[1]Istituto di Fonetica e Dialettologia - Consiglio Nazionale delle Ricerche, Italy, [2]Istituto Trentino di Cultura - Istituto per la Ricerca Scientifica e Tecnologica, Italy, [3]Oregon Graduate Institute for Science and Technology, USA

Finally Festival speaks Italian. In this work, the development of the first Italian version of the Festival TTS system is described. One male and one female voice for three different speech engines are considered: the Festival-specific residual LPC synthesizer, the OGI residual LPC Plug-In for Festival and the MBROLA synthesizer. The new Italian voices will be freely available for download for non-commercial purposes together with specific software modules at http://nts.csrf.pd.cnr.it/IFD/Pages/Italian-TTS.htm. This paper is devotedly dedicated to the memory of Mike Macon, whose recent passing on was really a shock to all of his friends.

Volume 1, page 509 Session B15

Multilingual TTS for Computer Telephony: The Aculab Approach

Monaghan A, Kassaei M, Luckin M, Amador-Hernandez M, Lowry D, Faulkner D, Sannier F
Aculab plc, UK

The requirements of the computer telephony (CT) industry place conflicting demands on text-to-speech (TTS) systems. Multilingual functionality and high quality output at telephone bandwidth requires detailed linguistic and acoustic analysis. At the same time, the need for robustness together with a high channel count and small memory footprint means that systems must be extremely efficient and databases must be kept small. We present a system which provides TTS for six languages, with 100 channels of highly natural output on a single DSP card.

Volume 1, page 513 Session B15

A Flexible Multilingual TTS Development and Speech Research Tool

Kiss G[1], Németh G[1], Olaszy G[2], Gordos G[1]
[1]Budapest University of Technology and Economics, Hungary, [2]Hungarian Academy of Sciences, Hungary

Diverse synthesis methods and text-to-speech (TTS) architectures are being developed and applied almost every day. This tendency raises the need for durable program systems that effectively assist research and development in this area. A flexible development system for multilingual text-to-speech and general speech research is introduced. The system was developed for use with the Multivox and Profivox concatenative speech synthesis systems, but its architecture makes it theoretically appropriate for a wide variety of purposes and different TTS systems. The system architecture and the functions of the development system are described. Keywords: TTS development system, speech research tools, system architecture, SGML derivative, object oriented design

Volume 1, page 517 Session B15

Speech Synthesis Development Made Easy: The Bonn Open Synthesis System

Klabbers E[1], Stöber K-H[2], Veldhuis R[1], Wagner P[2], Breuer S[2]
[1]IPO, Center for User-System Interaction, the Netherlands, [2]IKP, University of Bonn, Germany

This paper describes a new open source architecture for unit-selection based speech synthesis called BOSS (Bonn Open Synthesis System). It is built up modularly, with communications between modules taking place in a fixed format. This makes the addition, deletion and substitution of modules very easy. The strict separation between data and algorithms allows for the simple creation of new speech corpora for different domains and languages.

Volume 1, page 521 Session B15

Automatic Prosody Generation - a Model for Hungarian

Olaszy G[1], Németh G[2], Olaszi P[2]
[1]Hungarian Academy of Sciences, Hungary, [2]Budapest University of Technology and Economics, Hungary

In our model a complex function set is described for the three prosody components of read speech. Each of them is modelled separately by a three-step procedure. A new method, based on indirect determination of specific sound durations was developed. Final duration values are calculated from the specific durations in two further steps. F0 changes are also modelled by three levels, starting with rules on sentence level, followed by the word and syllable level, and completed by the micro intonation level. Another three level model serves the intensity structure, i.e. rules applied on sounds, on words and on the complete sentence. The three component models have influence on each other during prosody generation. Cross effects among them are also mentioned. The model can be applied in speech research and in applications (synthesis and recognition). It was tested for Hungarian. Keywords: prosody generation, three-level model, specific sound durations, word-level duration map

Volume 1, page 525 Session B15

Evaluation of PROS-3 for the assignment of prosodic structure, compared to assignment by human experts

van Herwijnen O M, Terken J M B
Technische Universiteit Eindhoven, The Netherlands

This paper describes the results of an evaluation of PROS-3, a system that assigns prosodic structure to text on the basis of the output of a syntactic parser. In order to evaluate the performance of PROS-3 as such and in combination with a revised algorithm for prosodic phrasing, we compare it to the prosodic structure as assigned by human experts. Also, the results of a perception experiment are presented, which show that listeners have the same preference of prosodic realization as we would expect on the basis of the comparison of the prosodic structures as assigned by PROS-3 and by human experts.

Volume 1, page 529 Session B15

Stochastic F0 Contour Model Based on the Clustering of F0 Shapes of a Syntactic Unit

Yamashita Y, Ishida T
Ritsumeikan University, Japan

This paper describes a stochastic modeling between an F0 contour and linguistic features of a sentence for speech synthesis. The F0 contour of a sentence is represented by concatenation of the F0 patterns of a Japanese syntactic unit, bunsetsu. A bunsetsu F0 pattern is composed of the F0 average and the F0 shape. The most probable sequence of bunsetsu F0 shapes for a sentence are found in the F0 shape database by a probabilistic measure. The probability that an F0 contour is observed for a sentence is defined by two kinds of probabilities, the F0 shape production and the F0 shape bigram. Several typical bunsetsu F0 shapes are extracted by clustering of training data and stored in the F0 shape database. The probability of the F0 shape production is computed for

each bunsetsu based on the distribution of linguistic features in the cluster.

Volume 1, page 533 Session B15

Intonational Phrase Break Prediction Using Decision Tree and N-Gram Model

Sun X[1], Applebaum T H[2]
[1]*Northwestern University, USA*, [2]*Panasonic Speech Technology Laboratory, USA*

In the current study, we propose and evaluate a new method for automatic intonational phrase break prediction based on sequences of parts-of-speech and word junctures. The proposed method uses decision trees to estimate the probability of a word juncture type (break or non-break) given a finite length window of part-of-speech values, and uses an n-gram to model the word juncture sequence. Trained on an 8,000 word database, our algorithm predicted breaks with F=77% and non-breaks with F=93%, which represents a significant improvement over the commonly used approach, which uses decision trees alone.

Volume 1, page 537 Session B15

Synthesizing Intonation of Standard Arabic Language

Zaki A[1], Rajouani A[2], Najim M[1]
[1]*Equipe Signal et Image, ENSEIRB, France*, [2]*Laboratoire d'Electronique et Etudes des Systèmes Automatiques Rabat, Morocco*

In this paper, we propose a model to generate fundamental frequency (F0) contours using neural networks. A learning procedure is proposed as an alternative to synthesis-by-rules. The generation of correct fundamental frequency contour is one of the important issues in the naturalness of automatic text-to-speech conversion systems. The proposed approach is based on a standard feed-forward multi-layer network that produces global F0 contours of sentences, directly from encoded linguistic features of standard Arabic language. Our model does not need syntactic information to produce suitable declarative intonation. TD-PSOLA synthesizer is used for validation of our results.

Volume 1, page 541 Session B15

Invariance of Relative F0 Change Field of Chinese Disyllabic Words

Xu D, Mori H, Kasuya H
Utsunomiya Univ., Japan

In automatic voice response systems where a large number of words are inserted into fixed sentences, such as in voice-guided car navigation systems, one of the most important problems is the adjustment of the fundamental frequency (F0) contour of the inserted word to suit the F0 context of the fixed sentence. The effects of intonation and tone on the F0 contours of Chinese words can be described in terms of a word-level F0 range (WF0R) and an F0 change field (F0CF). WF0R in any position of a sentence is a tone-independent general F0 range, whereas F0CF is an F0 range taking the tone combination of words into account. Relative F0CF is regulated in reference to WF0R. If WF0R is used to represent the declination of a sentence, the relative F0CF should be invariant but dependent on the tone combination of a word. This paper examines the invariance of the relative F0CF among individuals. From an analysis of four native speakers f utterances of 160 words in the initial, middle and final parts of three carrier sentences, conducted over 2 days, we show that: (1) Chinese speakers read words in the same sentence position with stable relative F0 change; (2) the relative F0CFs in the middle position of a sentence are generally the same as those in the initial position, but slightly different from those in the final position; and (3) the relative F0CFs reveal that the effects of tone on F0 contour is individual independent.

Volume 1, page 545 Session B15

Accent Label Prediction by Time Delay Neural Networks Using Gating Clusters

Mueller A F[1], Hoffmann R[2]
[1]*Siemens, Germany*, [2]*Dresden University of Technology, Germany*

In this paper a new neural network (NN) architecture for data driven prediction of accent labels---perceptual accents and pitch accents---for speech synthesis is presented. Within the proposed NN architecture, gating clusters are applied in a time delay (TD) framework. The gating clusters are used to adapt the network structure dynamically such that only available input feature vectors from the actual context window are treated. The proposed NN architecture has been successfully applied for accent label prediction on word level within our text-to-speech (TTS) system. Prediction accuracy for our German corpus was 86.1%. On an english corpus the achieved accuracy was 84.5%. This result is superior to results achieved on the same corpus with an approach based on classification and regression tree (CART) techniques[1]. The results were achieved with a simpler feature set than that used in[1]. [1] K. Ross and M. Ostendorf, "Prediction of abstract prosodic labels for speech synthesis"

Volume 1, page 549 Session B15

Transformation-Based Learning of Danish Stress Assignment

Henrichsen P J
Cophenhagen Business School, Denmark

In Danish, as in other languages, prosody assignment is fairly well described as a function of lexical and syntactic structure. So in principle, prosodic clue assignment should be open to machine learning techniques. This paper presents an experiment using transformation-based ML for unsupervised learning of Danish main stress assignment. The trained stress assigner is compared to the leading Danish text-to-speech system. In conclusion, ML for prosody assignment is advocated as an attractive alternative to naive word mapping as well as to labour-intensive grammar writing.

Volume 1, page 553 Session B15

On the Prosody of German Telephone Numbers

Baumann S, Trouvain J
University of the Saarland, Germany

Spoken telephone numbers are prosodically structured. This is reflected on various levels, such as grouping, wording and accenting. Realisation strategies employed by German speakers are used to model the prosody of telephone number production. In a listening preference test using synthetic speech two strategies used by commercial inquiry systems proved to be less acceptable than the versions based on the proposed models. These models are proposed for use in speech-synthesis-based telephone number inquiry services.

Volume 1, page 557 Session B15

Emotional Speech Synthesis: A Review

Schröder M
DFKI, Saarbrücken, Germany

Attempts to add emotion effects to synthesised speech have existed for more than a decade now. Several prototypes and fully operational systems have been built based on different synthesis techniques, and quite a number of smaller studies have been conducted. This paper aims to give an overview of what has been done in this field, pointing out the inherent properties of the various synthesis techniques used, summarising the prosody rules employed, and taking a look at the evaluation paradigms. Finally, an attempt is made to discuss interesting directions for future development.

Volume 1, page 561 Session B15

Fun or Boring? A Web-based Evaluation of Expressive Synthesis for Children

Gustafson K, House D
KTH, Sweden

Prosodic features were varied in four sentences synthesized using a developmental version of the Infovox 330 concatenated diphone Swedish male voice. The sentences were part of an interactive evaluation test carried out on a commercial website for a period of three months. 78 girls and 56 boys between the ages of 5 and 15 rated the sentences on a qualitative four-point scale. Results indicate that both girls and boys interpret large-scale F0 manipulations as representing a fun voice while longer durations are generally regarded as boring, especially by the boys. The results also confirm the feasibility of using a website for remote evaluation even with children.

Volume 1, page 565 Session B15

Session B16 - Poster
Tuesday - 09.00 - 10.40

Speech Recognition and Understanding: Articulatory and Perceptual Approaches to ASR

Chair: Nelson Morgan, ICSI, USA

Sub-Band Based Additive Noise Removal for Robust Speech Recognition

Chen J[1], Paliwal K K[2], Nakamura S[1]
[1]*ATR Spoken Language Translation Research Laboratories, Japan*,
[2]*Griffith University, Australia*

To make an automatic speech recognition system robust with respect to noise, we will probably have to solve two problems. One is the detection and identification of noise. Another is the consideration of noise effect during recognition process. In this paper, we will investigate several noise estimation approaches, such as moving average, long-term average, long-term Fourier analysis, etc. We will then introduce a sub-band based scheme to remove the noise effect from corrupted speech to make recognition system immune to additive noise. We will report on experiments on TI digits database and NOISEX database to justify the proposed approach.

Volume 1, page 571 Session B16

Development of an Asynchronous Multi-band System for Continuous Speech Recognition

Tam Y-C, Mak B
The Hong Kong University of Science and Technology, Hong Kong

Recently, multi-band automatic speech recognition (MBASR) has been proposed to combat environmental noises. We describe the two major efforts in the development of our asynchronous MBASR system for continuous speech recognition. Firstly, we successfully introduce asynchrony among sub-bands under the HMM composition framework. Secondly, the linear sub-band weightings are estimated by minimizing the string classification error among the N-best hypotheses using simulated noisy speech. When our asynchronous MBASR system is evaluated on connected TI digits with 0db additive low-pass white noise, compared with a full-band system, (1) our synchronous MBASR system reduces the absolute string error rate (SER) and word error rate (WER) by 19.8% and 14.1% respectively; (2) the introduction of asynchrony further reduces the absolute SER (WER) by 5.2%(2.5%); (3) an estimation of sub-band weightings using N-best string MCE training gives an additional reduction of absolute SER (WER) by 19.7% (5.1%).

Volume 1, page 575 Session B16

A Multi-Band Approach Based on the Probabilistic Union Model and Frequency-Filtering Features for Robust Speech Recognition

Jancovic P, Ming J
Queen's University of Belfast, UK

Multi-band approach has recently been introduced for recognition of speech corrupted by frequency-localized noise, showing higher robustness than the traditional full-band approach. However, the multi-band approach has been found to be less robust for wide-band noise than the full-band approach. In this paper, we present a multi-band recognition system based on the combination of the probabilistic union model and the frequency-filtering technique. The probabilistic union model is used to combine the features from the individual sub-bands without requiring information about the sub-band corruption. The frequency-filtering technique is used to produce the feature vector for

each sub-band, which is similar to the usual cepstral feature but does not spead the frequency-localized noise over the sub-bands. We demonstrate that this combination results in a system that is equally effective for dealing with both frequency-localized noise and wide-band noise.

Volume 1, page 579 Session B16

Split-band Perceptual Harmonic Cepstral Coefficients as Acoustic Features for Speech Recognition

Gu L, Rose K
University of California, Santa Barbara, USA

This paper presents a significant modification of our previously proposed speech recognizer's front-end based on perceptual harmonic cepstral coefficients. The spectrum is split into two frequency bands, which correspond to the harmonic and non-harmonic components. A weighting function, which depends both on the voiced/unvoiced/ transitional classification and on the prominence of harmonic structures, is applied to the harmonic band, and ensures accurate representation of the voiced and transitional speech spectral envelope. Conventional smoothed spectrum is used in the non-harmonic band. The mixed spectrum undergoes mel-scaled band-pass filtering, and the log-energy of the filters' output is discrete cosine transformed to produce cepstral coefficients. Experiments with Mandarin digit and E-set databases show significant recognition gains over plain perceptual harmonic cepstral coefficients and considerable gains over standard techniques.

Volume 1, page 583 Session B16

Error Correcting Posterior Combination for Robust Multi-Band Speech Recognition

Hagen A, Bourlard H
IDIAP, Martigny, Switzerland

In human perception, the availability of context enhances recognition and renders it more robust to noise. Even if not all phonemes in a word (or words in a sentence etc.) are correctly perceived, humans can fill in missing parts with the help of cues from the surrounding speech parts. This was proven in studies on human speech perception where recognition of words in sentences under noise was shown to outperform recognition of words in isolation or, even more drastically, of nonsense syllables under noise. A new model for quantifying the influence of contextual information on human recognition performance was recently proposed. Although the authors state that it is not a model for the recognition process itself, we will see how the ideas behind this model can be used in automatic speech recognition to extend our formerly introduced multi-band recognition systems to incorporate frequency contextual information. We will compare the new set-up to our former models such as the full combination subband approach and its approximation.

Volume 1, page 587 Session B16

Robust Parameters for Speech Recognition Based on Subband Spectral Centroid Histograms

Gajic B[1], Paliwal K[2]
[1]Norwegian University of Science and Technology, Norway, [2]Griffith University, Australia

In this paper we propose a new speech parameterization framework that efficiently combines frequency and magnitude information from the short-term power spectrum of speech. This is achieved through computation of subband spectral centroid histograms (SSCH). Relationship between the proposed method and auditory based speech parameterization methods is discussed. An experimental study on an automatic speech recognition task has shown that the proposed method outperforms the conventional speech front-ends in presence of different types of additive noise, while it performs comparably in the noise-free conditions. In the case of car noise, our method also outperforms the computationally expensive auditory based methods, while having simplicity and low computational cost similar to the conventional front-ends.

Volume 1, page 591 Session B16

Pseudo-Articulatory Representations and the Recognition of Syllable Patterns in Speech

Edmondson W, Zhang L
University of Birmingham, UK

This paper presents an account of syllable structure as the basis for organizing articulatory activity. This contrasts with the serial organization of more conventional phonetic segments. It is demonstrated that working with syllables in this way can provide the basis for linguistically motivated speech recognition using the previously reported notion of the Pseudo-Articulatory Representation (PAR).

Volume 1, page 595 Session B16

ASR - Articulatory Speech Recognition

Frankel J, King S
University of Edinburgh, UK

We propose that using a continuous trajectory model to describe an articulatory-based feature set will address some of the shortcomings inherent in the hidden Markov model (HMM) as a model for speech recognition. The articulatory parameters allow us to explicitly model effects such as co-articulation and assimilation. A linear dynamic model (LDM) is used to capture the characteristics of each segment type. These models are well suited to describing smoothly varying, continuous, yet noisy trajectories, such as we find present in speech data. Experimentation has been based on data for a single speaker from the MOCHA corpus. This consists of parallel acoustic and recorded articulatory parameters for 460 TIMIT sentences. We report the results of classification and recognition tasks using both real and recovered articulatory parameters, on their own and in conjunction with acoustic features.

Volume 1, page 599 Session B16

Efficient Decoding Strategy for Conversational Speech Recognition Using State-Space Models for Vocal-Tract-Resonance Dynamics

Ma J[1], Deng L[2]
[1]BBN Technologies, USA, [2]Microsoft Company, USA

In this paper, we present an efficient strategy for likelihood computation and decoding in a continuous speech recognizer using underlying state-space dynamic models for the hidden speech dynamics. The state-space models have been constructed in a special way so as to be suitable for the conversational or casual style of speech where phonetic reduction abounds. The interacting multiple model (IMM) state estimation algorithm for switching state-space models is first introduced, which uses a merging strategy derived from Bayes's rule to meet the challenge of exponential growth in the switching combination. Then one specific dynamic-programming based decoding algorithm, incorporating the merging strategy, are derived. It successfully overcomes the exponential growth in the original search paths by using the path-merging strategy. Evaluation experiments on conversational speech using the Switchboard corpus demonstrate that the use of the new decoding strategy is capable of reducing the recognizer's word error rate compared with the baseline recognizers, including the HMM system and the state-space dynamic model using the HMM-produced phonetic boundaries.

Volume 1, page 603 Session B16

HMM2- Extraction of Formant Structures and their Use for Robust ASR

Weber K[1], Bengio S[2], Bourlard H[1]
[1]IDIAP, Martigny / EPFL, Lausanne, Switzerland, [2]IDIAP, Martigny, Switzerland

As recently introduced [ftp://ftp.idiap.ch/pub/reports/2000/rr00-30.ps.gz], an HMM2 can be considered as a particular case of an HMM mixture in which the HMM emission probabilities (usually estimated through Gaussian mixtures or an artificial neural network) are modeled by state-dependent, feature-based HMM (referred to as frequency HMM). A general EM training algorithm for such a structure has already been developed [ftp://ftp.idiap.ch/pub/reports/2000/rr00-11.ps.gz]. Although there are numerous motivations for using such a structure, and many possible ways to exploit it, this paper will mainly focus on one particular instantiation of HMM2 in which the frequency HMM will be used to extract formant structure information, which will then be used as additional acoustic features in a standard Automatic Speech Recognition (ASR) system. While the fact that this architecture is able to automatically extract meaningful formant information is interesting by itself, empirical results will also show the robustness of these features to noise, and their potential to enhance state-of-the-art, noise-robust HMM-based ASR.

Volume 1, page 607 Session B16

Auditory model based speech recognition in noisy environment

Yu X, Wan W, Lun D
The Hong Kong Polytechnic University, Hong Kong

This paper presents a new speech feature, the ASBF speech feature based on the mathematical model of inner ear of human auditory system. This new speech feature is extracted using both mathematical model of inner ear and primary auditory nerve processing model of human auditory system, and it can track the speech formants effectively. In the experiment, the performance of MFCC and the ASBF are compared in both clean and noisy environments using left-to-right CDHMM with 6 states and 5 Gaussian mixtures. The experimental result shows that the ASBF is much more robust to noise than MFCC. When only 5 dimension is used in ASBF vector, the recognition rate is approximately 38.6% higher than the traditional MFCC with 39 dimension in the condition of S/N=10dB with white noise.

Volume 1, page 611 Session B16

Forward Masking for Increased Robustness in Automatic Speech Recognition

Wendt S, Fink G A, Kummert F
Universitaet Bielefeld, Germany

In automatic speech recognition MFCC or LPCC are features commonly used today. However, their calculation considers only a few features of the auditory system. On the assumption that the human representation of speech is an optimal representation, considering more features of the auditory system might lead to a better performance of automatic speech recognition systems. In this paper a model proposed by Strope and Alwan (see references), which relies on the human acoustic perception and allows to consider the effect of forward masking, is incorporated after some modifications into an automatic speech recognition system with a MFCC-based front-end. The extended system is evaluated on recognition tasks, that are closer to real recognition than (connected) digit recognition commonly used in the literature. The evaluations show an increased robustness of the speech recognition system with forward masking on all recognition tasks, but especially on data recorded in noisy environments.

Volume 1, page 615 Session B16

An Auditory System-Based Feature for Robust Speech Recognition

Li Q, Soong F K, Olivier S
Bell Labs, Lucent Technologies, USA

An auditory feature extraction algorithm for robust speech recognition in adverse acoustic environments is presented. The feature computation is comprised of an outer-middle-ear transfer function, FFT, frequency conversion from linear to the Bark scale, auditory filtering, nonlinearity, and discrete cosine transform. The feature is evaluated in two tasks: connected-digit recognition and large vocabulary continuous speech recognition. The tested data were under various noise conditions, including handset and hands-free speech data in landline and wireless communications with additive car and babble noise. Compared with the LPCC, MFCC, MEL-LPCC, and PLP features, the proposed feature has an average 20% to 30% string error rate reduction on the connected-digit task, and 8% to 14% word error rate reduction on the Wall Street Journal task in various additive noise conditions.

Volume 1, page 619 Session B16

Session B21 - Oral
Tuesday - 11.10 - 12.50

ESE2 - Noise Robust Recognition: Robust systems - What helps ?

Chair: David Pearce, Motorola, USA

Experiments with the Philips continuous ASR system on the AURORA noisy digits database

Lieb M, Fischer A
Philips Research Labs Aachen, Germany

With this work we evaluate the Philips continuous speech recognition system on the standardized AURORA noisy digit string recognition task. A variety of noise robust algorithms, ranging from spectral subtraction during the feature extraction stage, to adaptation techniques in the HMM-decoding stage, are applied and their effects are presented. Detailed experimental results show the contribution of the single approaches to the overall system performance. By thoroughly combining the best performing of the standard algorithms, we achieve significant improvements for the matched training as well as for the non-matched condition scenarios.

Volume 1, page 625 Session B21

Robust Digit Recognition in Noisy Environments: the IBM Aurora 2 System

Saon G, Huerta J, Jan E-E
IBM T.J. Watson Research Center, USA

In this paper we describe some experiments on the Aurora 2 noisy digits database. The algorithms that we used can be broadly classified into noise robustness techniques based on a linear-channel model of the acoustic environment such as CDCN and its novel variant termed Alignment-based CDCN (ACDCN, proposed here), and techniques which do not assume any particular knowledge about the structure of the environment or noise conditions affecting the speech signal such as discriminant feature space transformations and speaker/channel adaptation. We present recognition experiments for both the clean training data and the multi-condition training data scenarios.

Volume 1, page 629 Session B21

Evaluating the Aurora Connected Digit Recognition Task -- A Bell Labs Approach

Afify M, Jiang H, Korkmazskiy F, Lee C-H, Li Q, Siohan O, Soong F K, Surendran A C
Bell Labs, Lucent Technologies, USA

Connected digit recognition has always been an ideal task for fundamental research in speech recognition due to its low complexity and potential applicaitons. In Bell Labs we have developed a number of techniques targeting directly or indirectly at connected digit recognition. For the Aurora task, we study a few such algorithms for the entire spectrum of the issues, including feature extraction, context-dependent digit modeling, minimum classification error acoustic modeling, unsupervised noise compensation, and utterance verification. We show how each component contributes to the reduction of digit recognition and verification errors. Average over all three test sets we obtained 84.6% and 91.3% digit accuracies for clean- and multi-condition training, respectively. This represents an average of 48.6% error rate reduction when compared to the official Aurora baseline results.

Volume 1, page 633 Session B21

Session B22 - Oral
Tuesday - 11.10 - 12.30

Phonetics and Phonology: Segmentals

Chair: Ian Maddieson, University of California, Berkeley

Liaison and schwa deletion in French: an effect of lexical frequency and competition?

Fougeron C, Goldman J-P, Frauenfelder U H
Univ. of Geneva, Switzerland

This study aims to determine whether the production of the lexical variants created by the phonological processes of liaison and schwa deletion in French are conditioned by factors linked to lexical recognition. We hypothesise that the realisation of these variants would be favoured for words which are lexically "salient" in term of frequency and in their lexical neighbourhoods. This claim was tested by examining a speech corpus for the effects of lexical frequency, neighbourhood density and neighbourhood frequency on the production of liaison (both in linking and linked words and their co-occurrence) and elision. Overall the results do not support our hypothesis: lexical frequency and competition do not appear to influence strongly whether liaison and elision are realised or not.

Volume 1, page 639 Session B22

An Acoustical Analysis of the Vowels in Beijing Mandarin

Zee E[1], Lee W-S[2]
[1]City University of Hong Kong, Hong Kong, [2]University of Hong Kong, Hong Kong

The study is a spectral analysis of the vowels and syllabic approximants in Beijing Mandarin. It presents: (i) the average F1, F2, F3 values for the resonant sounds, (ii) the vowel ellipses for the resonant sounds, showing their relative positions in the F1/F2 plane, (iii) the vowel diagrams, showing the F-patterns of the first three formant frequencies of the resonant sounds, (iv) the formant trajectories for the rhotic schwa, showing that the vowel in the V syllables is actually a sequence of a plain schwa and a rhotic schwa, and (v) the diagrams of the average vowel positions for the vowels followed by a nasal ending, showing that the effect of the nasal ending on the F1 and F2 of the vowel sounds varies accoring to vowel type and nasal type.

Volume 1, page 643 Session B22

Discriminant analysis of nasal vs. oral vowels in French: comparison between different parametric representations

Delvaux V, Soquet A
Université Libre de Bruxelles, Belgium

The purpose of this paper is to investigate the realization of the [nasal] contrast in French by performing an acoustic analysis of naturally spoken nasal and oral vowels, and by carrying out discriminant analysis on these data. Results consistently show that generic parametric representations allow to reliably discriminate between nasals and orals. A specific issue addressed in this paper is the relationship between phonetic and phonological nasalization in French.

Volume 1, page 647 Session B22

Whispery voiced nasal stops in Rwanda

Demolin D, Delvaux V
Université Libre de Bruxelles, Belgium

The paper describes the main phonetic characteristics (acoustic, aerodynamic and articulatory) of Kinyarwanda prenasalized stops, focussing on voiceless consonants. We conclude from instrumental observations that the phonetic description of these sounds should be redefined. Consonants previously described as voiceless prenasalized stops in Rwanda are in fact whispery voiced nasal stops. Finally, the paper shows that the description of these sounds raises several important questions about nasal venting and the control of the velum closure.

Volume 1, page 651 Session B22

Session B23 - Oral
Tuesday - 11.10 - 12.30

Speech Production: Prosody - I

Chair: Louis C.W. Pols, Univ. of Amsterdam, The Netherlands

Prominence correlates. A study of Swedish

Fant G[1], Kruckenberg A[1], Liljencrants J[1], Botinis A[2]
[1]*KTH, Sweden,* [2]*University of Skövde, Sweden*

This is a summary of studies of word and syllable prominence in Swedish performed during several years. A unique feature is the correlation of observed acoustic data with a continuously scaled parameter of perceived prominence. Besides the established parameters of duration, F0, intensity, and spectral tilt we have also data on true subglottal pressure. Studies of co-variation within the set of acoustic parameters reveal some interesting relations, some of which can be related to the production mechanism. The major part of the material derives from prose reading, but we have also data from contrasting "lab type" sentences. Some systematic differences appear. Our findings have applications in the development of text-to-speech rules

Volume 1, page 657 Session B23

Quantitative Analysis of the Effects of Emphasis Upon Prosodic Features of Speech

Ohno S[1], Fujisaki H[2]
[1]*Tokyo University of Technology, Japan,* [2]*Science University of Tokyo, Japan*

While it is known that emphasis is represented mainly by fundamental frequency, speech rate, and source intensity, few studies have been published on the relative roles of these variables in expressing the degree of emphasis. The present paper introduces the relative speech rate and the relative source intensity of a target utterance against a reference utterance, and formulates the processes of their generation by quantitative models that are in line with the model that has been established for the fundamental frequency contour. This makes it possible to compare the effects of emphasis on the three variables in quantitative terms, as well as to compare the effects of various degrees of emphasis. Analysis of English utterances by a native speaker and a non-native speaker indicated the influence of emphasis on the three variables in quantitative terms, and also clarified the difference between native and non-native speakers.

Volume 1, page 661 Session B23

Towards a Model of Target Oriented Production of Prosody

Dogil G, Möbius B
University of Stuttgart, Germany

A new paradigm for prosody research is presented, inspired by the speech production model recently proposed by Guenther, Perkell, and colleagues. This research paradigm aims at generalizing the production model by extending it from a predominantly segmental perspective to a new theory of the production of prosody. Speech movements in the prosodic domain are interpreted as intonational gestures that are planned to reach and traverse perceptual target regions. Evidence from F0 alignment studies suggests that the perceptual targets can be approximately represented by regions in a multidimensional acoustic-temporal space. These studies also indicate that segmental, spectral, temporal, and prosodic structure are co-produced in such a way as to mutually support and enhance, and not impair, the perceptual targets. Furthermore, examples of multi-level mappings between invariant and variable targets in the domain of prosody are provided, and a dichotomy of phonemic and postural prosodic settings is discussed.

Volume 1, page 665 Session B23

Prosody Control for Speaking and Singing Styles

Shih C, Kochanski G P
Bell Laboratories, Lucent Technologies, USA

By proper control of prosody, text-to-speech systems already have the capability to imitate distinctive speaking styles. We show two examples where we can capture the critical features: the singing style of Dinah Shore and the speaking style of Martin Luther King Jr. The styles are described by Stem-ML tags (soft template mark-up language), which offers the flexibility needed to control accent shapes, phrasal pitch contours, and amplitude profiles, for speech as well as for singing.

Volume 1, page 669 Session B23

Session B24 - Oral
Tuesday - 11.10 - 12.30

Speech Recognition and Understanding: Acoustic Modelling - I

Chair: Andrej Ljolje, AT&T Labs, USA

A Mixture of Gaussians Front End for Speech Recognition

Stuttle M, Gales M J F
Cambridge University, UK

This paper describes a feature extraction technique based on fitting a Gaussian mixture model (GMM) to the speech spectral envelope. The features obtained (the component means, variances and priors) represent both the the general shape of the spectrum and provide information on the position of the spectral peaks. As the features select peaks in the spectrum they are related to the formant amplitudes, locations and bandwidths. Results using the Resource Management corpus, a medium vocabulary task are presented. Although by themselves the GMM features do not outperform MFCC features, systems combining the GMM systems with a standard frontend are shown to give a reduction in word error rate.

Volume 1, page 675 Session B24

Improved Maximum Mutual Information Estimation Training of Continuous Density HMMs

Zheng J, Butzberger J, Franco H, Stolcke A
SRI International, USA

In maximum mutual information estimation (MMIE) training, the currently widely used update equations derive from the Extended Baum-Welch (EBW) algorithm, which was originally designed for the discrete hidden Markov model (HMM) and was extended to continuous Gaussian density HMMs through approximations. We derive a new set of equations for MMIE based on a quasi-Newton algorithm, without relying on EBW. We find that by adopting a generalized form of the MMIE criterion, the H-criterion, convergence speed and recognition performance can be improved. The proposed approach has been applied to a spelled-word recognition task leading to a 21.6% relative letter error rate reduction with respect to the standard Maximum Likelihood Estimation (MLE) training method, and showing advantages over the conventional MMIE approach in terms of both training speed and recognition accuracy.

Volume 1, page 679 Session B24

Maximum-Likelihood Training of a Bipartite Acoustic Model for Speech Recognition

Perronnin F, Kuhn R, Nguyen P, Junqua J-C
Panasonic Speech Technology Laboratory, USA

This paper describes a context-dependent model that supports extremely rapid speaker adaptation. The model, called "Eigencentroid plus Delta Trees" (EDT), incorporates prior knowledge about speaker space and has modest memory requirements. The paper gives the formulae for training EDT models and performs a detailed entropy analysis to show how EDT and speaker-independent models trained on experimental data differ from each other. Phoneme recognition results on the TIMIT database are also given. EDT yields 12.1% relative error rate reduction (ERR) for supervised adaptation on three sentences, 11.2% ERR for unsupervised adaptation on three sentences, and 10.4% ERR for self-adaptation on a single sentence.

Volume 1, page 683 Session B24

Analysis of the Root-Cepstrum for Acoustic Modeling and Fast Decoding in Speech Recognition

Sarikaya R, Hansen J
RSPL-CSLR, University of Colorado-Boulder, USA

Root-cepstral analysis has been proposed previously for speech recognition in car environments. In this paper, we focus on an alternative aspect of Root-cepstrum as it applies to discriminative acoustic modeling and fast speech recognizer decoding. We compare Root-cepstrum to Mel-Frequency cepstrum Coefficients (MFCC) in terms of their noise immunity during modeling and decoding speed. Our experiments use the SPINE~cite{HAN00} corpus which is composed of clean and noisy data with a 5K vocabulary size. Experiments were performed that allow pair-wise comparisons of acoustic models across different feature sets and acoustic units. We observed that for 84% of the phonemes, the average distance to all other acoustic units is increased in the Root-cepstrum domain compared to MFCC resulting in a sharp acoustic model set. Therefore, the ambiguity in the Root-cepstrum space is reduced. Large vocabulary noisy speech recognition experiments showed a 27.5% reduction in real--time processing factor (RTF) compared to MFCC features while improving overall recognition accuracy.

Volume 1, page 687 Session B24

Session B25 - Poster
Tuesday - 11.10 - 12.30

Linguistic Modelling: Language Models - I

Chair: Kazuhiko Ozeki, Univ. of Electro-Communications, Tokyo, Japan

Structured language model for class identification of out-of-vocabulary words arising from multiple word-classes

Onishi S, Yamamoto H, Sagisaka Y
ATR Spoken Language Translation Research Laboratories, Japan

A structured language model (STLM) is proposed to cope with out-of-vocabulary (OOV) words coming from multiple word-classes. The STLM aims at independently modeling the classes without interference and identifying the class of words arising from multiple word-classes. The STLM consists of the conventional word-class N-gram and the sets of the independent-trained class-specific sub-word N-grams. We made an experimental language model by using STLM for the two similar proper-noun classes and performed the speech recognition experiments. The results show that any OOV word of the one class is never misrecognized as that of the other class. The results show that the STLM could integrate the multiple different statistical language models with no interference.

Volume 1, page 693 Session B25

New Language Models Using Phrase Structures Extracted From Parse Trees

Jitsuhiro T[1], Yamamoto H[1], Yamada S[2], Sagisaka Y[1]
[1]ATR Spoken Language Translation Research Laboratories, Japan,
[2]NTT Communication Science Laboratories, Japan

This paper proposes a new speech recognition scheme using three linguistic constraints. Multi-class composite bigram models are used in the first and second passes to reflect word-neighboring characteristics as an extension of conventional word n-gram models. Trigram models with constituent boundary markers and word pattern models are both used in the third pass to utilize phrasal constraints and headword co-occurrences, respectively. These two models are made using a training text corpus with phrase structures given by an example-based Transfer-Driven Machine Translation (TDMT) parser. Speech recognition experiments show that the new recognition scheme reduces word errors 9.50% from the conventional scheme by using word-neighboring characteristics, that is only the multi-class composite bigram models.

Volume 1, page 697 Session B25

Triggering Individual Word Domains in N-Gram Language Models

Sicilia-Garcia E I, Ming J, Smith F J
Queen's University of Belfast, UK

We present a new method of introducing domain knowledge into an n-gram language model. It is based on a combination of language models for individual word domains. Each word model is built from an individual corpus which is formed by extracting those subsets of the entire training corpus which contain that significant word. When testing, significant words are extracted from a cache and their models are combined with a global language model. Different methods of combining the models are described; one simple method based on combining frequencies rather than probabilities gives promising results and provides a relatively simple method of introducing domain information into an n-gram language model. A 20% reduction in language model perplexity over the standard 3-gram approach is obtained which is similar to results

obtained with other more complex domain models. The model also requires a small cache compared with other models requiring a cache.

Volume 1, page 701 Session B25

A Structured Statistical Language Model conditioned by Arbitrarily Abstracted Grammatical Categories based on GLR Parsing

Akiba T, Itou K
National Institute of Advanced Industrial Science and Technology, Japan

This paper presents a new statistical language model for speech recognition, based on Generalized LR parsing. The proposed model, the Abstracted Probabilistic GLR (APGLR) model, is an extension of the existing structured language model known as the Probabilistic GLR (PGLR) model. It can predict next words from arbitrarily abstracted categories. The APGLR model is also a generalization of the original PGLR model, because PGLR can be considered to be a special case of APGLRs that predict the next words from the least abstracted grammatical categories, namely the terminal symbols. The selection of the abstraction level is arbitrary; we show several strategies to define the level. The experimental results show that the proposed model performs better than the original PGLR model for speech recognition.

Volume 1, page 705 Session B25

Speech Recognition of Broadcast Sports News

Matsui A, Segi H, Kobayashi A, Imai T, Ando A
NHK(Japan Broadcasting Corp.), Japan

This paper shows that a domain-dependent language model and state-skipped HMMs can achieve improvements in word recognition accuracy on a broadcast sports news transcription task. Although a domain-dependent language model is much better than a general model in terms of word error rate, the smaller training corpus for a special topic relative to the general news corpus leads to problems especially in higher-order n-gram probability estimation. In this paper, we tried a linear interpolation technique to smooth out unreliable higher-order n-gram probabilities using more reliable lower-order n-gram probabilities. We also applied a language model adaptation technique by using news manuscripts on sports topics. For acoustic modeling, we added two state-skipping paths to three-state HMMs to deal with phonemes of duration less than three frames. Overall, we reduced the word error rate from 15.1% to 5.8%, and achieved sufficient performance to realize real-time subtitling services.

Volume 1, page 709 Session B25

Improvement of a Structured Language Model: Arbori-context Tree

Mori S, Nishimura M, Itoh N
IBM Research, Tokyo Research Laboratory, IBM Japan, Japan

In this paper we present an extention of a context tree for a structured language model (SLM), which we call an arbori-context tree. The state-of-the-art SLM predicts the next word from a fixed partial tree of the history tree, such as two exposed heads, etc. An arbori-context tree allows us to select an optimum partial tree of a history tree for the next word prediction depending on the effectiveness in the similar way that a context tree selects the length of the history (n of n-gram). The experiment we conducted showed that the test set perplexity of the SLM based on an arbori-context tree (79.98) was lower than that of the SLM with a fixed history (101.56).

Volume 1, page 713 Session B25

Smoothing Issues in the Structured Language Model

Kim W, Khudanpur S, Wu J
The Johns Hopkins University, USA

The Structured Language Model (SLM) recently introduced by Chelba and Jelinek is a powerful general formalism for exploiting syntactic dependencies in a left-to-right language model for applications such as speech and handwriting recognition, spelling correction, machine translation, etc. Unlike traditional N-gram models, optimal smoothing techniques -- discounting methods and hierarchical structures for back-off -- are still being developed for the SLM. In the SLM, the statistical dependencies of a word on immediately preceding words, preceding syntactic heads, non-terminal labels, etc., are parameterized as overlapping N-gram dependencies. Statistical dependencies in the parser and tagger used by the SLM also have N-gram like structure. Deleted interpolation has been used to combine these N-gram like models. We demonstrate on two different corpora -- WSJ and Switchboard -- that more recent modified back-off strategies and nonlinear interpolation methods considerably lower the perplexity of the SLM. Improvement in word error rate is also demonstrated on the Switchboard corpus.

Volume 1, page 717 Session B25

The Study Of The Effect Of Training Set On Statistical Language Modeling

Shen X, Xu B
Institute of Automation, Chinese Academy of Sciences, P.R. China

In this work, we make a study on the effect of training set on statistical language modeling (SLM). A corpus selection system based on perplexity is presented. It is tested in two experiments: one is to select optimal training corpus for generating a domain-specific SLM; the other one is for generating an optimal SLM for a LVCSR system. The results show that the training corpus is important for the capability of SLM and our corpus selection system is powerful for optimal corpus selection. With the help of this system, we generated a SLM for a LVCSR system, which contributed 14.5%--17.7% relative character error reduction.

Volume 1, page 721 Session B25

Stochastic Finite State Automata Language Model triggered by Dialogue States

Esteve Y[1], Bechet F[1], Nasr A[2], De Mori R[1]
[1]LIA - University of Avignon, France, [2]LIM - University of Aix-Marseille II, France

Within the framework of Natural Spoken Dialogue systems, this paper describes a method for dynamically adapting a Language Model (LM) to the dialogue states detected. This LM combines a standard n-gram model with Stochastic Finite State Automata (SFSAs). During the training process, the sentence corpus used to train the LM is split into several hierarchical clusters in a 2-step process which involves both explicit knowledge and statistical criteria. From the same sentence corpus, SFSAs are extracted in order to model longer contexts than the ones used in the standard n-gram model. A first decoding process calculates a word-graph as well as a first sentence hypothesis. This first hypothesis will be used to find the optimal sub-LM. Then, a rescoring process of the word graph using this LM is performed. By adapting the LM to the dialogue state detected, we show a statistically significant gain in WER on a dialogue corpus collected by France Telecom R&D.

Volume 1, page 725 Session B25

A Baseline Method for Compiling Typed Unification Grammars into Context Free Language Models

Rayner M[1], Dowding J[2], Hockey B A[2]
[1]Netdecisions, UK, [2]RIACS, USA

This paper presents a minimal enumerative approach to the problem of compiling typed unification grammars into CFG language models, a prototype implementation and results of experiments in which it was

used to compile some non-trivial unification grammars. We argue that enumerative methods are considerably more useful than has been previously believed. Also, the simplicity of enumerative methods makes them a natural baseline against which to compare alternative approaches.

Volume 1, page 729 Session B25

Comparison of Width-wise and Length-wise Language Model Compression

Whittaker E, Raj B
Compaq Cambridge Research Laboratory, USA

In this paper we investigate the extent to which Katz back-off language models can be compressed through a combination of parameter quantization (width-wise compression) and parameter pruning (length-wise compression) methods while preserving performance. We compare the compression and performance that is achieved using entropy-based pruning against that achieved using only parameter quantization. We then compare combinations of both methods. It is shown that a broadcast news language model can be compressed by up to 83% to only 12.6Mb with no loss in performance on a broadcast news task. Compressing the language model further by quantization to 10.3Mb resulted in only a 0.4% degradation in word error rate which is better than can be achieved through entropy-based pruning alone.

Volume 1, page 733 Session B25

Large Vocabulary Statistical Language Modeling for Continuous Speech Recognition in Finnish

Siivola V, Kurimo M, Lagus K
Helsinki University of Technology, Finland

Statistical language modeling (SLM) is an essential part in any large-vocabulary continuous speech recognition (LVCSR) system. The development of the standard SLM methods has been strongly affected by the goals of LVCSR in English. The structure of Finnish is substantially different from English, so if the standard SLMs are directly applied, the success is by no means granted. In this paper we describe our first attempts of building a LVCSR for Finnish and the new SLMs that we have tried. One of our objective has been the indexing and recognition of broadcast news, so special issues of our interest are topic detection, word stemming and modeling words that are poorly covered in the training data. Our new methods are based on neural computing using the self-organizing map (SOM) which has recently been shown to successfully extract and approximate latent semantic structures from massive text collections.

Volume 1, page 737 Session B25

A New Technique Based on Augmented Language Models to Improve the Performance of Spoken Dialogue Systems

López-Cózar R[1], Milone D H[2]
[1]Granada University, Spain, [2]Universidad Nacional de Entre Ríos, Argentine

This paper presents a new technique that aims to improve the performance of spoken dialogue systems by using the so-called augmented language models. We define an augmented language model as a compound of a language model and a set of values concerning parameters that can influence the speech recognition when the language model is used. The diverse language models used by a dialogue system can be very different, in terms of perplexity for example. Then, the aim of the technique is to find and use the combination of values concerning the different parameters that leads to the best recognition results when the different language models are used by a dialogue system. The technique has been applied to a dialogue system for the fast food domain. The results show that when the augmented language models are used the system's performance is enhanced. In the experiments we have achieved a reduction of 9,33% in the word error rate and an increment of 11,26% in the sentence understanding.

Volume 1, page 741 Session B25

Session B26 - Poster
Tuesday - 11.10 - 12.30

Speaker Recognition: Identification, Verification and Tracking. Speech Recognition and Understanding: Language Identification

Chair: Jean-Francois Bonastre, LIA-Universite d'Avignon, France

The Influence of Vocal Effort on Human Speaker Identification

Brungart D S[1], Scott K R[1], Simpson B D[2]
[1]Air Force Research Laboratory, USA, [2]Veridian, USA

Although many of the acoustic cues used for speaker identification change systematically with the voice level of the talker, little is known about the influence vocal effort has on the identification of individual talkers by human listeners. In this experiment, listeners were trained to identify four different same-sex talkers speaking at one of three different levels of vocal effort (whispered, conversational, or shouted). They were then tested on their ability to identify the same four talkers speaking at the other two levels of vocal effort. The results show that the whispering talkers were harder to identify than the conversational talkers, and that the conversational talkers were harder to identify than the shouting talkers. The results also show that listeners who were trained to identify individual talkers speaking at one level of vocal effort had difficulty identifying the same talkers when they were speaking at a different level of vocal effort.

Volume 2, page 747 Session B26

Improving Speaker Recognition Using Phonetically Structured Gaussian Mixture Models

Faltlhauser R, Ruske G
Technische Universitaet Muenchen, Germany

Gaussian Mixture Models (GMM) are highly suitable for speaker identification and verification. Nevertheless these models try to represent primarily the distribution of the available training data - neglecting any possible phonetic information which might be of worth. In our paper we present a recognition system using multiple speaker GMMs based on phonetic classes. By introducing 'phonetic' mixture coefficients a weighting of phoneme classes with respect to speaker recognizability can be achieved. The implicit integration in the probability computation avoids the need for a phonetic labeling during recognition. The mixture weights can be learned in a training phase. Model training was examined applying MAP enrolment and the recently reported Eigenvoice approach. Especially for the latter a phonetic separation is advantageous. Recognition error reductions up to 15% relatively were achieved. Furthermore, the multiple GMM approach is particularly effective for speaker enrolment with sparse training data.

Volume 2, page 751 Session B26

Information Fusion for Robust Speaker Verification

Sanderson C, Paliwal K
Griffith University, Australia

In this paper we have studied two information fusion approaches, namely feature vector concatenation and decision fusion, for the task of reducing error rates in a speaker verification system used in mismatched conditions. Three types of features are fused: Mel Frequency Cepstral Coefficients (MFCC), MFCC with Cepstral Mean Subtraction (CMS) and Maximum Auto-Correlation Values (MACV). We have used the mismatch sensitivity of Linear Prediction Cepstral Coefficients (LPCC) as a speech quality measure for selecting the weight of the contribution of the MFCC modality in the adaptive decision fusion approach. We show that in most cases concatenation fusion is superior to decision fusion. The results lead us to propose a hybrid fusion approach in which two combinations of concatenation fusion are further fused using adaptive decision fusion. The hybrid system is shown to have the lowest error rates on both clean and noisy speech.

Volume 2, page 755 Session B26

A Robust Speaker Verification System against Imposture Using an HMM-based Speech Synthesis System

Satoh T[1], Masuko T[1], Kobayashi T[1], Tokuda K[2]
[1]Tokyo Institute of Technology, Japan, [2]Nagoya Institute of Technology, Japan

This paper describes a text-prompted speaker verification system which is robust to imposture using synthetic speech generated by an HMM-based speech synthesis system. In the verification system, text and speaker are verified separately. Text verification is based on phoneme recognition using HMM, and speaker verification is based on GMM. To discriminate synthetic speech from natural speech, an average of inter-frame difference of the log likelihood is calculated, and input speech is judged to be synthetic when this value is smaller than a decision threshold. Experimental results show that the false acceptance rate for synthetic speech was reduced drastically without significant increase of the false acceptance and rejection rates for natural speech.

Volume 2, page 759 Session B26

Sequential Decisions for Faster and More Flexible Verification

Surendran A
Bell Labs, Lucent Technologies, USA

Most speaker verification systems wait to collect a complete utterance from a speaker before making a decision. Faster verification can be achieved if decisions are made sequentially on smaller ``chunks" of data. In this paper we present a sequential decision making algorithm in a connected digit application and discuss its properties. We show that sequential decisions, apart from requiring shorter utterances and fewer computations on the average, add another dimension of flexibility over current systems: within some limitations, they provide the ability to systematically tradeoff between the performance of the system and the amount of data needed to make a decision. Thus they make a speaker verification system work faster and be more flexible in real applications.

Volume 2, page 763 Session B26

Background Learning of Speaker Voices for Text-independent Speaker Identification

Tsai W-H[1], Chu Y C[2], Huang C-S[2], Chang W-W[3]
[1]Philips Research East Asia-Taipei / National Chiao Tung University, Taiwan, ROC, [2]Philips Research East Asia-Taipei, Taiwan, ROC, [3]National Chiao Tung University, Taiwan, ROC

This study provides a novel learning mechanism, the so-called background learning, to the problem of text-independent speaker identification (speaker ID). Unlike the conventional speaker ID, the proposed system does not rely on enrollment data of clients in construction of speaker-specific models, but instead attempts to learn speakers' voices via clustering and parametric modeling of off-line collected data with no label of speaker identity. This eliminates the necessity of enrolling a large amount of speech data from clients. To permit such unsupervised learning, an efficient algorithm for blind clustering of speech utterances based on speaker characteristics is developed. Experimental results demonstrated that when very limited enrollment data is available, the speaker-ID performance achieved with

the background learning could emulate that of using abundant enrollment data.

Volume 2, page 767 Session B26

Explicit Exploitation of Stochastic Characteristics of Test Utterance for Text-independent Speaker Identification

Tsai W-H[1], Chang W-W[2], Huang C-S[3]
[1]Philips Research East Asia-Taipei / National Chiao Tung University, Taiwan, ROC, [2]National Chiao Tung University, Taiwan, ROC, [3]Philips Research East Asia-Taipei, Taiwan, ROC

In this paper, a novel speaker-identification (speaker-ID) technique based on explicit exploitation of stochastic characteristics of test utterance is proposed. Unlike the conventional approach which hypothesizes the identity of a test speaker by determining which client's model maximizes the likelihood for the test utterance, it is aimed to bilaterally compare test speaker's voices with client speakers' voices instead of simply taking the unilateral likelihoods into account. We study two approaches respectively based on cross likelihood ratio and Bayesian information criterion to accomplish this aim. Performance of the proposed approaches was evaluated by close-set text-independent speaker ID experiments and was shown to be superior to that of the conventional approach based on maximum likelihood decision rule.

Volume 2, page 771 Session B26

Improvement of Speaker Verification for Thai Language

Wutiwiwatchai C, Achariyakulporn V, Kasuriya S
National Electronics and Computer Technology Center, Thailand

There are many strategies proposed for speaker verification (SV) system, both in text-dependent (fixed-text) and text-independent (free-text) domains. To convey an appropriate algorithm for Thai speech, several consecutively improvement methods are compared in this paper including the dynamic time warping (DTW) matching and Gaussian mixture model (GMM) based systems. We firstly developed a system based on the conventional scoring algorithm. This system is improved by the incorporation of many scoring algorithms such as the cohort normalization, the global speaker model (GSM), and a new approach, namely, global anti-speaker model (GASM). Experiments are set up for Thai numeral speech and the results show an improving tendency of each algorithm.

Volume 2, page 775 Session B26

Speaker Identification for Car Infotainment Applications

Rodríguez-Saeta J[1], Koechling C[2], Hernando J[1]
[1]UPC, Spain, [2]Robert Bosch, Germany

Car applications demand more and more the use of speech technologies. Drivers must concentrate on controlling the car and the non-use of hands makes the voice a valuable tool. Here we analyze the possibility of identifying the user of a car through her/his voice in order to develop some useful applications, and establish preferences, some of them related to music. The identification will be done in parallel to speech commands which will be given to devices in the car in the future. Once the user is identified, the system loads a personal profile. It includes music preferences which can be downloaded from the Internet databases using e.g. MPEG-7 .

Volume 2, page 779 Session B26

A System for Text Dependent Speaker Verification - Field Trial Evaluation and Simulation Results

Schalk H[1], Reininger H[2], Euler S[3]
[1]Universitaet Frankfurt, Germany, [2]ATIP, Germany, [3]Robert Bosch, Germany

In a speaker verification system, an identity claim is made by an unknown speaker. An utterance of this speaker and a model of the speaker whose identity is claimed is compared. If the model and the utterance match well the claim is accepted otherwise it is rejected. Thus, two classes of errors can occur in a speaker verification system: false acceptances and false rejections. The Equal Error Rate (EER) is often used as a performance measure of a verification system and results if the system parameters are adjusted in such a way that the two kinds of errors are equal. To model a speaker, reference utterances of this speaker are recorded in an enrollment session. Modelling itself can be done either in the signal domain using Dynamic Time Warping (DTW) or with a stochastic model of the speaker using Hidden Markov Models (HMM).

Volume 2, page 783 Session B26

Speaker Recognition in a Multi-Speaker Environment

Martin A, Przybocki M
National Institute of Standards and Technology, USA

We discuss the multi-speaker tasks of detection, tracking, and segmentation of speakers as included in recent NIST Speaker Recognition Evaluations. We consider how performance for the two-speaker detection task is related to that for the corresponding one-speaker task. We examine the effects of target speaker speech duration and the gender mix within test segments on results for these tasks. We also relate performance results for the tracking and segmentation tasks, and look at factors affecting segmentation performance.

Volume 2, page 787 Session B26

A New DP-Like Speaker Clustering Algorithm

Ou Z, Wang Z
Tsinghua Univ., P. R. China

In this paper we propose a new segment-synchronous speaker clustering algorithm based on the Bayesian Information Criterion (BIC), which is motivated by the Dynamic Programming (DP) idea. Compared with the commonly used agglomerative speaker clustering methods, the proposed algorithm is faster for lack of distance-matrix building and more reasonable as it avoids in some degree the simple irrevocable merging fashion. Moreover it facilitates online speaker clustering, which is important for real-time transcription applications (e.g., broadcast news, teleconferences etc.). In our experiments on 1997 Hub4 Mandarin broadcast news development data, unsupervised speaker adaptation with this DP-like clustering achieved 17.66% relative reduction in Character Error Rate (CER) from the baseline, as much as with the clustering by the true speaker identities.

Volume 2, page 791 Session B26

On the Use of the Bayesian Information Criterion in Multiple Speaker Detection

Sivakumaran P, Fortuna J, Ariyaeeinia A M
University of Hertfordshire, UK

An efficient scheme, based on the Bayesian information criterion (BIC), for the detection of speaker changes in an audio stream is introduced and investigated. BIC has been the subject of considerable attention in recent years due to its effectiveness for speaker change detection (SCD) as well as the detection of other forms of acoustic changes. A main difficulty in BIC-based SCD has been reported to be that of the computational complexity. The scheme proposed here tackles this problem by reducing the computational load in the previously proposed algorithms significantly, without compromising their effectiveness. The paper describes the new scheme thoroughly and analyses its performance. Experiments are based on 3 hours of broadcast news with 416 speaker changes. With this data, the proposed scheme has been found to be

capable of running in about 0.06 times real-time whilst keeping the rate of each of misdetection and false alarm close to 9%.

Volume 2, page 795 Session B26

Preliminary experiments on language identification using broadcast news recordings

Benarousse L, Geoffrois E
DGA/CTA/GIP, France

This article presents experiments on language identification using Broadcast News recordings, for which large amounts of data are available. The system uses a Broadcast News partitioner developed by LIMSI to extract the speech segments from raw signals. These segments are then transcribed using a language-independent HMM acoustic model. Phonotactic models are trained for each language, and used to score the transcription of the test signals. Training was conducted on recordings from three monolingual radios (about 17h of signal per language) and tests were made on signals from other radios. We also investigated a rejection strategy to improve the identification results. Without any rejection, the error rates range from 13.8% (5s segments) to 4.3% (45 s segments). Rejecting 1/3 of the data improves these rates by 78% for 10s segments.

Volume 2, page 799 Session B26

Multi-Stream Statistical N-Gram Modeling With Application To Automatic Language Identification

Kirchhoff K, Parandekar S
University of Washington, USA

Most state-of-the art automatic language identification systems are based on phonotactic information, i.e. languages are identified on the basis of probabilities of phone sequences extracted from the acoustic signal. This approach ignores the potential advantages to be gained from a richer representation of the acoustic signal in terms of parallel streams of subphonemic events. In this paper we develop an alternative approach to language identification which is based on parallel streams of phonetic features and sparse modeling of statistical dependencies between these streams. We present results on the OGI-TS database and show that the feature-based system outperforms a comparable phone-based system significantly while using fewer parameters. Moreover, the feature-based system exhibits a markedly better performance on very short test signals (< 3 seconds).

Volume 2, page 803 Session B26

Session B31 - Demonstrations
Tuesday - 14.00 - 15.20

ESE3 - Imagination 2001

Chair: Gerrit Bloothooft, University of Utrecht, The Netherlands

Human language technologies are coming of age. ELSNET challenged the young generation to demonstrate or simulate at Eurospeech 2001 - Scandinavia new, imaginative, creative or even crazy ideas for innovative applications using speech and language technology. The winner of the contest will win 5000 Euro contributed by Hewlett-Packard European Research Labs. Proposals in every area of communication were welcomed, all modalities included, as long as speech and language play a significant role. These could involve imaginative applications in the fields of aids for the handicapped, games, education, wireless communication systems, performing arts, internet, agents and avatars, ambient intelligence, etc. All registrants to Eurospeech 2001 - Scandinavia, less than 35 years of age at the time of the conference, could participate (also as group effort). At Eurospeech 2001 - Scandinavia a jury will judge the ideas of feasibility, creativeness and originality. The jury consists of Sadaoki Furui, Julia Hirschberg, Joseph Mariani, Hans Kamperman, and Roger Tucker. The winner will be announced during the closing ceremony of the conference.

Volume 2, page 808 Session B31

Session B32 - Oral
Tuesday - 14.00 - 15.20

Phonetics and Phonology: Prominence and Timing

Chair: Gøsta Bruce, Dept. of Linguistics and Phonetics, Sweden

Up to what level can acoustical and textual features predict prominence

Streefkerk B M[1], Pols L C W[1], ten Bosch L F M[2]
[1]Univ. of Amsterdam, the Netherlands, [2]L & H, Belgium

In this paper both acoustical as well as textual correlates of prominence are discussed. Prominence, as we use it, is defined at the word level and is based on listener judgments. A selection of useful acoustic input features is tested for classification of prominent words, with the help of Feed Forward Nets. We use spoken sentences from many different speakers, taken from the Dutch Polyphone corpus of telephone speech. For an independent test set of 1,000 sentences about 72% of the words are correctly classified as prominent or not. At the text input level we also developed an algorithm, using linguistic/syntactical features derived from text only, to predict prominence. The prediction agrees with the perceived prominence in 82.6% of the cases.

Volume 2, page 811 Session B32

Linguistic Factors Affecting Timing in Korean with Application to Speech Synthesis

Chung H, Huckvale M
University College London, UK

This paper describes the results of a study of the phonetic and phonological factors affecting the rhythm and timing of spoken Korean. Stepwise construction of a CART model was used to uncover the contribution and relative importance of phrasal, syllabic, and segmental contexts. The model was trained from a corpus of 671 read sentences, yielding 42,000 segments each annotated with 69 linguistic features. On reserved test data, the best model showed a correlation coefficient of 0.73 with a RMS prediction error of 26 ms. Analysis of the classification tree during and after construction shows that phrasal structure had the

greatest influence on segmental duration. Strong lengthening effects were shown for the first and last syllable in the accentual phrase. Syllable structure and the manner features of surrounding segments had smaller effects on segmental duration. The model has application within Korean speech synthesis.

Volume 2, page 815 Session B32

Measuring Rhythmic Deviation in Second Language Speech

Schaeffler F
University of Munich, Germany

This study deals with the question of whether recently provided methods to determine the rhythm class of languages can be transferred to foreign-accented speech. Therefore read German speech of Venezuelan Spanish native speakers was compared with read speech of a native German control group by means of four different measurements. Three of the four applied measurements showed significant differences between the two groups, with one of the differences contradicting earlier expectations. The study has shown that the measurements can be successfully transferred to foreign-accented speech, but slightly modified measurements are suggested.

Volume 2, page 819 Session B32

Good timing: Place-dependent voice onset time in ejective stops

Maddieson I
University of California, Berkeley, USA

Voice onset time after voiceless unaspirated stops demonstrates a dependence on place of articulation, most reliably being shorter for labial and coronal than for velar stops. Some of the proposed explanations for this pattern suggest that a parallel dependence is not be expected for aspirated or ejective stops. However, similar patterns do occur with both aspirated and unaspirated stops. Cho and Ladefoged (1999) have suggested that ejectives do not follow the same trend, but they had little data on bilabial ejectives to compare with more plentiful data on velars. This paper contributes more material to this debate with expanded data on Yapese and the first published material on ejective VOT in Nez Perce. The results suggest that ejectives have a similar pattern to plosives and that therefore a unified explanation for all three types of stops should be sought.

Volume 2, page 823 Session B32

Session B33 - Oral
Tuesday - 14.00 - 15.20

Speech Synthesis: Concatenation - I

Chair: Thierry Dutoit, Faculte Polytechnique de Mons, Belgium

Design of an optimal continuous speech database for text-to-speech synthesis considered as a Set Covering Problem

Francois H, Boeffard O
IRISA, Université Rennes 1, ENSSAT, France

Text-to-speech synthesis can be carried out by concatenation of acoustic units obtained from a continuous speech database. This paper presents the optimization of such as database according to phonetic criteria. A large corpus of texts is assembled (311 572 sentences), phonetized automatically and condensed (12 217 sentences) to retain only 10 tokens of the most frequent triphonemes. This is a NP-hard problem of set covering. It has been solved in an approximate way using a greedy algorithm. The condensed database covers 25% of the initial distinct triphonemes, each being represented by 10 tokens at least, which allows 95% of the triphoneme tokens of the initial corpus to be covered. The distribution of the triphonemes remains proportional to their initial statistical appearance.

Volume 2, page 829 Session B33

Use of Clustering Information for Coarticulation Compensation in Speech Synthesis by Word Concatenation

Vosnidis C, Digalakis V
Technical University of Crete, Greece

The Weather Report Synthesizer is a speech synthesis system for weather forecasts in Greek. Instead of trying to improve the synthesis quality of PSOLA based diphone concatenation speech synthesizers, we have chosen to use words as the synthesis units. This approach has the advantage of low complexity and quick implementation, and achieves better speech quality due to the fact that the synthesis units inherently possess the necessary prosodic feature diversity. The selection of the optimal sequence of words that form the synthesized speech, however, presents the greatest challenge in the synthesis process. Several features are taken into consideration during the selection, but we have identified Coarticulation at the edges of consecutive words to have the greatest effect on the quality of the synthesized utterance. We present a novel method for evaluating a measure on coarticulation effects among pairs of words, based on feature clustering information obtained from a current SR system.

Volume 2, page 833 Session B33

Reducing spectral mismatches in concatenative speech synthesis via systematic database enrichment.

Founda M, Tambouratzis G, Chalamandaris A, Carayannis G
Institute for Language and Speech Processing, Greece

This paper presents work performed for the Time-Domain TTS system, which is being developed at the ILSP for the Greek language. It focuses on the enhancement of the synthetic speech quality, by reducing the spectral mismatches between concatenated segments. To that end, a study has been performed to determine the distance that can best predict when a spectral mismatch is audible. Experimentation with different spectral distances has taken place and the distance with the best performance has been used in order to systematically enrich the segment database, which initially contained only one instance per segment. Results of this procedure indicate a substantial improvement in the synthetic speech quality.

Volume 2, page 837 Session B33

Hansori 2001 - Corpus-based Implementation of the Korean Hansori Text-to-Speech Synthesizer

Ferencz A, Choi S-W, Song H-E, Koo M-W
Korea Telecom R & D Group, Korea

The improvement of Text-to-Speech (TTS) synthesizers' speech quality and naturalness is a continuous concern of researchers worldwide. The present paper gives a brief introduction of several previous Hansori TTS systems and is introducing our approach on experimenting, adopting and implementing corpus-based techniques for the system. We are focusing on corpus selection, on the optimal unit search criteria, on the missing unit (triphone unit) replacement handling, and we present some useful development tool options included in the trial version of the system.

Volume 2, page 841 Session B33

Session B34 - Oral
Tuesday - 14.00 - 15.20

Speech Recognition and Understanding: LVCSR - II

Chair: Sadaoki Furui, Tokyo Inst. of Technology, Japan

Time and Memory Efficient Viterbi Decoding for LVCSR using a Precompiled Search Network

Willett D, McDermott E, Minami Y, Katagiri S
NTT Communication Science Laboratories, Japan

In this paper, we present our recently developed time-synchronous speech recognition decoder, which adopts the idea of representing the search space of Large Vocabulary Continuous Speech Recognition (LVCSR) in a single precompiled network. In particular, we outline our approaches for time and memory efficient Viterbi decoding in this scenario. This includes reducing the fast memory needs by keeping the search network on disk and only loading the required parts on demand. Evaluations are carried out on a difficult Japanese LVCSR task which involves a back-off trigram language model and full cross-word dependent triphone acoustic models. Time and memory efficiency enables the real-time Viterbi decoding of entire lecture speeches in a single time-synchronous pass with a search error of less than 1%.

Volume 2, page 847 Session B34

A New Verification-Based Fast Match Approach to Large Vocabulary Speech Recognition

Liu F, Afify M, Jiang H, Siohan O
Bell Laboratories, Lucent Technologies, USA

This paper proposes a new fast match algorithm for large vocabulary continuous speech recogniton. By viewing the fast match as a verification problem we develop a likelihood ratio score to be used instead of conventional likelihood based fast matches. We also improve the computation through incremental calculation and the use of SSE for Intel instruction set. When used in a 20K Japanese broadcast news task the proposed fast match leads to about 30-40% improvement in speed with almost no degradation in the recognition accuracy.

Volume 2, page 851 Session B34

A Fast Calculation Method in LVCSRS by Time-Skipping and Clustering of Probability Density Distributions

Nakagawa S, Horibe Y
Toyohashi University of Technology, Japan

In this paper, we propose a rapid output probability calculation method in HMM based large vocabulary continuous speech recognition systems (LVCSRS). This method is based on time-skipping of calculation, clustering of probability density distributions, and pruning of calculation. Only distributions covering input feature vectors with high probabilities are used to calculate output probabilities strictly, and representative distributions for other distributions are used to calculate them approximately. Here a skipping method for likelihood calculation is adopted in the time domain. Using the rapid calculation method by clustering of probability density distributions, the recognition time in a LVCSRS system was reduced by about 40%. Using a pruning method of likelihood calculations on the way, it was further reduced by 25%. Finally, using time-skipping, the calculation time, furthermore, was reduced by 15% without compromising recognition accuracy.

Volume 2, page 855 Session B34

Speech Recognition of Japanese News Commentary

Homma S, Kobayashi A, Sato S, Imai T, Ando A
NHK (Japan Broadcasting Corp.), Japan

This paper describes some improvements in speech recognition of broadcast news commentary in Japanese. Since news commentary speech has different linguistic and acoustic features from read speech, it gives lower word recognition accuracy. In this paper we apply to news manuscripts some rules which represent the linguistic features of news commentaries, and generate word sequences for language model adaptation. We also use a large volume of transcriptions of news programs as training texts. Acoustic models are speaker-adapted and their structures are changed so as to recognize relatively short phonemes, because we found the speech rate of news commentary is sometimes much faster than that of read speech. Furthermore, by using a decoder that can handle cross-word triphone models, we reduced the word error rate by 32%.

Volume 2, page 859 Session B34

Session B35 - Poster
Tuesday - 14.00 - 15.20

Speech Recognition and Understanding: Noise Robustness - I

Chair: Alex Acero, Microsoft Research, USA

A comparison of LPC and FFT-based acoustic features for noise robust ASR

de Wet F, Cranen B, de Veth J, Boves L
University of Nijmegen, The Netherlands

Within the context of robust acoustic features for automatic speech recognition (ASR), we evaluated mel-frequency cepstral coefficients (MFCCs) derived from two spectral representation techniques, i.e. the fast Fourier transform (FFT) and linear predictive coding (LPC). ASR systems based on the two feature types were tested on a digit recognition task using continuous density hidden Markov phone models. System performance was determined in clean acoustic conditions as well as in different simulations of adverse acoustic conditions. The LPC-based MFCCs outperformed their FFT counterparts in most of the adverse acoustic conditions that were investigated in this study. A tentative explanation for this difference in recognition performance is given.

Volume 2, page 865 Session B35

Unsupervised Noisy Environment Adaptation Algorithm Using MLLR and Speaker Selection

Yamada M[1], Baba A[2], Yoshizawa S[2], Mera Y[1], Lee A[1], Saruwatari H[1], Shikano K[1]
[1]Nara Institute of Science and Technology, Japan, [2]Laboratories of Image Information Science and Technology, Japan

An unsupervised acoustic model adaptation algorithm using MLLR and speaker selection for noisy environments is proposed. The proposed algorithm requires only one arbitrary utterance and environmental noise data. The adaptation procedure is composed of the following four steps. (1) Speaker selection from a large number of database speakers is carried out using GMM speaker models based on one arbitrary utterance. (2) Initial speaker adapted HMM acoustic models are calculated from the HMM sufficient statistics of the selected speakers, where the sufficient HMM statistics are pre-calculated and stored. (3) A small subset of the clean speech database from the selected speakers and the environment noise data are superimposed. (4) MLLR adaptation is carried out using the noise-superimposed speech database from the selected speakers. The proposed algorithm is evaluated in a 20k vocabulary dictation task for newspaper in noisy environments. We attain 85.7% word correct rate in 25dB SNR, which is slightly better than the matched model by the E-M training using noise superimposed whole speech database. The proposed algorithm is also 7% better than the HMM composition algorithm.

Volume 2, page 869 Session B35

Applying Parallel Model Compensation with Mel-Frequency Discrete Wavelet Coefficients for Noise-Robust Speech Recognition

Tufekci Z, Gowdy J, Gurbuz S, Patterson E
Clemson University, USA

Interfering noise severely degrades the performance of a speech recognition system. The Parallel Model Combination (PMC) technique is one of the most efficient techniques for dealing with such noise. Another method is to use features local in the frequency domain. Recently, we proposed Mel-Frequency Discrete Wavelet Coefficients (MFDWCs) as speech features local in frequency domain. In this paper, we discuss using PMC along with MFDWC features to take advantage of both noise compensation and local features (MFDWCs) to decrease

the effect of noise on recognition performance. In addition we discuss the effect of increasing the number of the noise model mixture component on the performance of the Mel-Frequency Cepstral Coefficients (MFCCs) and MFDWCs. We evaluate the performance of MFDWCs using the NOISEX-92 database for various noise types and noise levels. We also compare the performance of these versus MFCCs and both using PMC for dealing with additive noise.

Volume 2, page 873 Session B35

Linear Interpolation of Cepstral Variance for Noisy Speech Recognition

Hwang T-H[1], Yuo K-H[2], Wang H-C[2]
[1]Industrial Technology Research Institute, Taiwan, ROC, [2]National Tsing Hua University, Taiwan, ROC

Speech model combination with the background noise has been shown effective to improve the pattern classification rate of noisy speech. The model combination can be performed by the addition of the spectral statistics such as the means and the variances. Since the speech feature for pattern classification has to be expressed in the cepstral domain, the combined spectral statistics have to be transferred into the cepstral domain for speech recognition. In our previous study, we have proposed a direct adaptation scheme of the cepstral variance that is without the mapping from the spectral domain to the cepstral domain. In this paper, an improved version to perform the adaptation is proposed. From the study, it is observed that the adapted variance can be expressed as a linear interpolation of the speech and the noise variances to obtain a comparable recognition rate that is obtained with the mapping process. Due to the direct adaptation of the variances, a lot of computation can be reduced to perform the environmental adaptation.

Volume 2, page 877 Session B35

Evaluation of a Generalized Dynamic Cepstrum in Distant Speech Recognition

Matsumoto H, Shimizu A, Yamamoto K
Shinshu University, Japan

This paper examines the effectiveness of a generalized dynamic cepstrum in distant speech recognition. The generalized dynamic cepstrum (DyMFGC) is based upon the forward masking on the generalized logarithmic spectrum instead of the log-spectrum, which intends to make it robust to additive noise as well as convolutional noise. Digit recognition tests were carried out in a relatively quiet and small sized office environment. Under white noise environments, the DyMFGC outperforms the dynamic cepstrum on the logarithmic spectrum and the MFCC with cepstral mean normalization. It also maintains the word accuracy of 90% to 95% within a 1m distance from a source. In speech babble noise environments, the performance of the DyMFGC is approximately the same as that of the dynamic cepstrum on the logarithmic amplitude scale.

Volume 2, page 881 Session B35

Robust Speech/Non-Speech Detection using LDA applied to MFCC for continuous Speech Recognition

Martin A, Damnati G, Mauuary L
France Telecom R&D, France

Continuous speech recognition applications need precise detection because the number of words to recognize is unknown and vocabulary words can be short. The speech/non-speech detection must be robust to the boundary precision. In this work, a new approach to evaluate detection algorithm for continuous speech recognition is presented. The speech/non-speech detection using energy parameter combined with a Linear Discriminant Analysis (LDA) applied to Mel Frequency Cepstrum Coefficients (MFCC) is compared to the algorithm based on signal to noise ratio (SNR). The LDA applied to MFCC for speech/non-speech detection improves recognition performance in noisy environment and for continuous speech recognition applications.

Volume 2, page 885 Session B35

Toward Noise-tolerant Acoustic Models

Trentin E, Gori M
Universita' di Siena, Italy

Acoustic models relying on hidden Markov models (HMMs) are heavily noise-sensitive: recognition performance drops whenever a significant difference in acoustic conditions holds between training and test environments. The relevance of developing acoustic models that are intrinsically robust has to be stressed. Robustness to noise is related to the generalization capabilities of the model. Artificial neural networks (ANNs) appear to be a promising alternative, but they historically failed as a general paradigm for speech recognition. This paper faces the problem by (i) investigating the recognition performance of the ANN/HMM hybrid proposed by the authors over tasks with noisy signals, and (ii) proposing an explicit ``soft'' weight grouping technique, capable to improve its robustness. Experiments over noisy speaker-independent connected-digits strings are presented. In particular, results on the VODIS II/SpeechDatCar database, collected in a real car environment, show the dramatic gain over the standard HMM, as well as over Bourlard and Morgan's hybrid.

Volume 2, page 889 Session B35

Noise Estimation Without Explicit Speech, Non-speech Detection: a Comparison of Mean, Modal and Median Based Approaches

Evans N W D, Mason J S
University of Wales, UK

Automatic speech recognition performance tends to be degraded in noisy conditions. Spectral subtraction is a simple, popular approach of noise compensation. In conventional spectral subtraction noise statistics are updated during speech gaps and subtracted from a corrupt signal during speech intervals. Some means of explicit speech, non-speech detection is therefore essential. Recent proposals have avoided the problem of speech, non-speech detection by continually updating noise estimates whether speech is present or not. In this paper, we evaluate two such approaches of noise estimation and compare their performance with standard noise estimation in hand-labelled speech gaps. Experimental results are reported with the conventional spectral subtraction framework on a 1500 speaker database. Results confirm that such approaches of noise estimation which do not rely on explicit speech, non-speech detection compare favourably with conventional noise estimation approaches.

Volume 2, page 893 Session B35

Evaluation of Front-End Features and Noise Compensation Methods for Robust Mandarin Speech Recognition

Chengalvarayan R
Lucent Technologies, USA

This paper describes speaker-independent speech recognition experiments concerning acoustic front-end processing on a telephone database that was recorded in various dialect regions in China. In this paper, three different features based on human voice production, perception and auditory systems have been evaluated for Mandarin speech recognition. Experimental comparisons showed that auditory-filtered cepstral coefficients outperforms the other type of features. When speech recognizers are deployed in telephone services, they often encounter variable acoustic mismatches which significantly deteriorate their performance. Three different channel equalization techniques have been explored in this study to decrease this mismatch, hence improving

the recognition accuracy. We present results with various noise compensation methods based on hierarchical cepstral mean subtaction and signal bias removal.

Volume 2, page 897 Session B35

ALGONQUIN: Iterating Laplace's Method to Remove Multiple Types of Acoustic Distortion for Robust Speech Recognition

Frey B J[1], Deng L[2], Acero A[2], Kristjansson T[3]
[1]University of Toronto, Canada, [2]Microsoft Research, USA, [3]University of Waterloo, Canada

We show how an iterative form of Laplace's method can be used to estimate the log-spectrum of clean speech from the log-spectrum of noisy, distorted speech, using a time-varying mixture model of the log-spectra of the clean speech, noise, channel distortion and noisy speech. We use this method, called ALGONQUIN, to denoise speech features and then feed these features into a large vocabulary speech recognizer whose WER on the clean WSJ data is 4.9%. When 10dB of time-varying airplane engine noise is added to the data, the recognizer obtains a WER of 28.8%. ALGONQUIN reduces the WER to 12.6%, well below the WER of 25.0% obtained by spectral subtraction, and close to the WER of 9.7% obtained by retraining the recognizer on training data corrupted by the exact same noise. If ALGONQUIN is used to denoise the noisy training data before the recognizer is retrained, the WER drops to 8.5%. For 10dB of white noise, spectral subtraction reduces the WER from 55.1% to 33.8%. ALGONQUIN reduces the WER to 14.2%. The recognizer trained on noisy data obtains a WER of 14.0%, whereas the recognizer trained on noisy data denoised by ALGONQUIN obtains a WER of 9.9%.

Volume 2, page 901 Session B35

Robust Speech Recognition in Noise: An Evaluation using the SPINE Corpus

Hansen J H L, Sarikaya R, Yapanel U, Pellom B
Univ. of Colorado Boulder, USA

In this paper, methodologies for effective speech recognition are considered along with evaluations of an NRL speech in noise corpus entitled SPINE. When speech is produced in adverse conditions that include high levels of noise, workload task stress, and Lombard effect, new challenges arise concerning how to best improve recognition performance. Here, we consider tradeoffs in (i) robust features, (ii) front-end noise suppression, (iii) model adaptation, and (iv) training and testing in the same conditions. The type of noise and recording conditions can significantly impact the type of signal processing and speech modeling methods that would be most effective in achieving robust speech recognition. We considered alternative frequency scales (M-MFCC, ExpoLog), feature processing (CMN, VCMN, LP-vs-FFT MFCCs), model adaptation (PMC), and combinations of gender dependent with gender independent models. For the purposes of achieving effective speech recognition performance, computational speed and availability of adaptation data greatly impacts final recognition performance. In particular, while reliable algorithm formulations for addressing specific types of distortion can improve recognition rates, these algorithms cannot reach their full potential without proper front-end algorithm data processing to direct compensation. While parallel banks of speech recognizers can improve recognition performance, their significant computational requirements can render the recognizer useless in actual speech applications.

Volume 2, page 905 Session B35

Session B36 - Poster
Tuesday - 14.00 - 15.20

Speech Production: Prosody - II

Chair: Sarah Hawkins, University of Cambridge, United Kingdom

Automated modeling of Chinese intonation in continuous speech

Kochanski G, Shih C
Bell Laboratories, Lucent Technologies, USA

We built and trained a model of intonation in continuous Mandarin speech based on the Stem-ML model of interacting accents. With this model, we found that we can accurately reproduce the intonation of the speaker using only one accent template for each lexical tone category. The resulting parameters are interpretable, and we find that the fitted model is consistent with linguistic expectations. Stem-ML is a phenomenological model of the muscle dynamics and planning process that controls the tension of the vocal folds. It describes the interactions between nearby tones or accents.

Volume 2, page 911 Session B36

Prediction of intonation patterns of accented words in a corpus of read Swedish news through pitch contour stylization

Frid J
Lund University, Sweden

This paper describes an initial attempt at the construction of a data-driven model of Swedish intonation. The study is mainly concerned with model-building and prediction of the intonation patterns of accented words in a corpus of read news in Swedish. Extraction of pitch information is achieved by performing a stylization of the pitch contours. The information is used to build a model for the prediction of pitch patterns using linguistic features such as accent type and position of stress. The model is tested against unseen data from the same corpus. The evaluation is done by numerical comparisons. The RMSE between predicted and original contours for the different categories ranges between 3.7 and 31.4 Hz. The results are quite promising for future studies.

Volume 2, page 915 Session B36

The use of fundamental frequency raising as a strategy for increasing vocal intensity in soft, normal, and loud phonation

Alku P[1], Vintturi J[2], Vilkman E[3]
[1]Helsinki University of Technology, Finland, [2]Helsinki University Central Hospital, Finland, [3]Oulu University, Finland

A method is presented to estimate the effect of intentional raising of fundamental frequency (F0) on vocal intensity. The method, Energy of the Synthesised Period (ESP), is based on computation of the energy of a hypothetical speech sound synthesised using a single period of the glottal volume velocity waveform and a digital filter that models the vocal tract. Both the glottal flow and the vocal tract filter are estimated by inverse filtering. The results show that, in producing loud voice, speakers use F0 to increase the number of glottal closures per time unit, which increases rapid fluctuations in the speech pressure waveform, which, in turn, raises vocal intensity. The average increase of sound pressure level due to this active use of F0 was approximately 4 dB in loud speech.

Volume 2, page 919 Session B36

Prosodic Interactions on Segmental Durations In Greek

Botinis A[1], Fourakis M[2], Bannert R[3]
[1]University of Skövde, Sweden, [2]The Ohio State University, USA, [3]Umeå University, Sweden

The present study is an experimental investigation of the effects of prosodic variables on segmental durations in Greek. Nonsense disyllabic CVCV words were produced in a carrier sentence under different conditions of stress, focus and tempo. The results indicate: (1) the intrinsic durations of vowels are rather canonical in the order /iu<eo<a/; (2) the adjacent consonant /s/ shows complementary duration tendencies; (3) stress has bigger effect on the vowel than the consonant; (4) focus has no major effects; (5) tempo has also bigger effect on the vowel than the consonant. In summary, stress has a bigger effect on both consonant and vowel durations than tempo whereas the effects of focus are in question.

Volume 2, page 923 Session B36

Study on Factors Influencing Durations of Syllables in Mandarin

Chu M[1], Feng Y[2]
[1]Microsoft Research China, P. R. China, [2]Institute of Acoustics, Chinese Academy of Sciences, P. R. China

This paper studies factors that have influences on durations of syllables in Mandarn. Six factors are investigated by comparing means of duration in different categories in the preparatory study. The vector space is reduced from 299250 to 1000 by removing factors that do not cause significant changing in means and merging levels in each factor that have similar means. In order to remove influences from various initials and finals, durations of syllables are divided by the mean duration of the corresponding base syllables. The sum of squared residuals analysis is performed on the normalized duration. The results reveal that boundary index has the largest influence on syllable durations. Tone identity comes next. Though, position in word is not a very important single factor, it influences duration together with boundary index.

Volume 2, page 927 Session B36

A Comparative Study of Pauses in Dialogues and Read Speech

Gustafson-Capkova S[1], Megyesi B[2]
[1]Stockholm University, Sweden, [2]KTH, Sweden

This study aims to investigate the length, frequency and position of various types of pauses in three different speech styles: elicited spontaneous dialogues, professional reading and non-professional reading.

Volume 2, page 931 Session B36

Detecting Japanese Local Speech Rate Deceleration in Spontaneous Conversational Speech Using a Variable Threshold

Takamaru K, Hiroshige M, Araki K, Tochinai K
Hokkaido University, Japan

The variable threshold(VT), which detects the speech rate deceleration, is proposed. The VT varies dynamically depending upon the duration of previous mora in the utterance. The VT should not change rapidly because listener cannot perceive small variations of mora duration. Thus, a set of functions with time constants which decide response speed of the VT is introduced. We apply the VT to six sentences of spontaneous conversational speech. The auditory test of detecting local speech rate deceleration is carried out for the evaluation. The possibility of detecting the local speech rate deceleration by the VT is indicated.

Volume 2, page 935 Session B36

Modelling Fundamental Frequency in First Post-tonic Syllables in Danish Sentences

Petersen N R
University of Copenhagen, Denmark

The work reported in the paper continues previous research on the description of Danish intonation in mathematical terms using a linear regression model. The present paper focuses on the first post-tonic syllable which constitutes the fundamental frequency peak in the stress group in Standard Danish. The results indicate that--in contradistinction to the stressed syllables--the F0 variation over the sentence in the first post-tonics is explained mainly by their position in the sentence, and to a much lesser degree by their position in the prosodic phrase.

Volume 2, page 939 Session B36

Non-Finality and Pre-Finality in Bari Italian Intonation: a Preliminary Account

Savino M
Politecnico di Bari, Italy

In this paper, a preliminary account of intonational strategies used by Bari Italian speakers in signal non-finality and pre-finality in discourse organisation is presented and discussed. Results obtained from auditory and instrumental analysis of speech material elicited with different methods (Map Task dialogues and monologues, lists readings) show that a rich inventory of intonational choices is available to Bari Italian speakers for conveying subordination relationships within a sequence of information (in a route describing task, this is normally a sequence of instructions and/or explanations), but also for signalling in advance the end of the sequence. Moreover, these results represent a further contribution to the development of an autosegmental-metrical account of the Bari Italian intonation system.

Volume 2, page 943 Session B36

Building An Integrated Prosodic Model of German

Mixdorff H[1], Jokisch O[2]
[1]Berlin University of Applied Sciences, Germany, [2]Dresden University of Technology, Germany

The intellegibility and naturalness of synthetic speech strongly depends on its prosodic quality. Departing from works by Mixdorff on a linguistically motivated model of German intonation based on the Fujisaki model, the current paper presents statistical results concerning the relationship between linguistic and phonetic information underlying an utterance and its prosodic features. Statistical analysis yields, inter alia, the following pairs of strongest single factor - prosodic feature: boundary depths (right) - syllable duration; boundary depths (left) - phrase command magnitude Ap; accent type (intoneme) - accent command amplitude Aa. These results were employed for training an FFNN-based integrated prosodic model predicting syllable durations along with syllable-aligned Fujisaki control parameters. Correlations between trained and predicted parameters suggest synergy effects, as they are mostly higher than correlations yielded when predicting parameters individually from the same set of input features using a regression model. Informal listening tests with resynthesis examples showed encouraging results.

Volume 2, page 947 Session B36

A Model of F0 Contour for Arabic Affirmative and Interrogative Sentences

Ibrahim O A G[1], El-Ramly S H[1], Abdel-Kader N S[2]
[1]Ain Shams University, Egypt, [2]Cairo University, Egypt

This Paper presents the results of analyzing the global contour of the fundamental frequency (F0) for Arabic sentences and developing a model that represents it. The work concentrated on analyzing only affirmative and interrogative isolated 'read-loud' sentences. The work is divided into two parts: 1) Extracting the common characteristics and differences between the F0 plots of affirmative and interrogative Arabic sentences, and 2) Analyzing the effect of change in sentences length on the characteristics and differences extracted in the first part of the study. The model obtained can be easily implemented in speech synthesizers to improve its intonation.

Volume 2, page 951 Session B36

Variation in Final Lengthening as a Function of Topic Structure

Smith C L, Hogan L A
University of New Mexico, USA

This experiment shows that for an English speaker reading aloud, the topic structure of the text affects the amount of lengthening in sentence-final words. The speaker lengthened words less at the end of sentences that were followed by another sentence elaborating on the topic of the first, than at the end of sentences where the subsequent sentence added new information or switched topics. These results show that speech durations are affected by larger-scale linguistic organization, in addition to the well-known effects of local and phrasal structure. Modeling variation at the text or discourse level could improve the comprehensibility of longer passages of synthesized text.

Volume 2, page 955 Session B36

Do speakers realize the prosodic structure they say they do?

van Herwijnen O M, Terken J M B
Technische Universiteit Eindhoven, The Netherlands

In this paper we describe a study in which a comparison was made between prosodic structures as realized in a spoken version of a text and as assigned by annotators of this text on paper. The prosodic structures were assigned by experts. This study puts to test the strategy of annotating text on paper to obtain a HUMAN reference of the prosodic structure that would be assigned when reading text aloud. This strategy is less time consuming than the often used analysis of spoken versions to obtain the assigned prosodic structure. The results of the comparison described here show that speakers are fairly capable of predicting what prosodic structure they would assign when reading text aloud.

Volume 2, page 959 Session B36

Coarticulatory Effects at Prosodic Boundaries: Some Acoustic Results

Tabain M, Rolland G, Savariaux C
Institut de la Communcation Parlée, France

Acoustic data are presented from a prosodic database containing data from 3 French speakers. The prosodic boundaries examined are the Utterance, the Intonational Phrase, the Accentual Phrase, and the Word. The aim is to study the interaction of coarticulatory effects with prosodic effects. The vowel /a/ before the prosodic boundary and the consonants /b d g f s S/ after the prosodic boundary are examined. It is found that the vowel duration is greatly affected by the strength of the prosodic boundary, but consonant duration less so. The duration of the fricative consonants is more stable than the stop consonants. Formant values suggest that /a/ is lower and more back the stronger the prosodic boundary, and that the vowel is more likely to reach its low target following a bilabial consonant /b f/. Based on an examination of formant values, the velar stop /g/ appears to have much variability in the front-back dimension. Finally, there is a strong negative correlation between duration and mean velocity of the formant transition, and this effect is related to the prosodic boundary.

Volume 2, page 963 Session B36

Generating Duration from a Cognitively Plausible Model of Rhythm Production

Barbosa P
IEL/UNICAMP, Brazil

A dynamical model of rhythm production is presented. The model is meant to generate segmental duration from the interplay between a dynamical rhythmic system and a gestural score representation. The rhythmic level is being implemented by a coupled-oscillator system (composed by a syllabic and a phrase stress oscillator) that delivers V-to-V-size beats to the gestural score. The model is able to automatically generate segment and pause acoustic duration according to speech rate input. The coupling of both oscillators as well as the interaction between the rhythmic system and a linguistic description of sentences is achieved by a recurrent neural network. The network delivers syllable-size normalized durations, which are then statistically distributed among the segments. The model exhibits cognitively plausible language universal and language-specific phonetic properties that are in complete disagreement with output-oriented techniques of speech generation which do not take into account the underlying speech production mechanism.

Volume 2, page 967 Session B36

Session B41 - Demonstrations
Tuesday - 15.50 - 17.30

ESE3 - Imagination 2001 - Continued

Chair: Gerrit Bloothooft, University of Utrecht, The Netherlands

Human language technologies are coming of age. ELSNET challenged the young generation to demonstrate or simulate at Eurospeech 2001 - Scandinavia new, imaginative, creative or even crazy ideas for innovative applications using speech and language technology. The winner of the contest will win 5000 Euro contributed by Hewlett-Packard European Research Labs. Proposals in every area of communication were welcomed, all modalities included, as long as speech and language play a significant role. These could involve imaginative applications in the fields of aids for the handicapped, games, education, wireless communication systems, performing arts, internet, agents and avatars, ambient intelligence, etc. All registrants to Eurospeech 2001 - Scandinavia, less than 35 years of age at the time of the conference, could participate (also as group effort). At Eurospeech 2001 - Scandinavia a jury will judge the ideas of feasibility, creativeness and originality. The jury consists of Sadaoki Furui, Julia Hirschberg, Joseph Mariani, Hans Kamperman, and Roger Tucker. The winner will be announced during the closing ceremony of the conference.

Volume 2, page 972 Session B41

Session B42 - Oral
Tuesday - 15.50 - 17.30

Speech Synthesis: Concatenation - II

Chair: Julie Vonwiller, Appen Pty Limited, Australia

Must Diphone Synthesis be so Unnatural?

Barry W[1], Nielsen C[2], Andersen O[2]
[1]University of the Saarland, Germany, [2]Aalborg University, Denmark

An English utterance was synthesized in four versions using sets of diphones produced under four different prosodic and contextual conditions. The synthesis used either accented di-phones only or appropriately located accented and unaccented diphones, with each of these conditions being repeated using neutral-context and differentiated-context diphones. They were presented to two listener groups, a native English and a non-native group for paired comparison acceptability judge-ments. The results show a massive preference for the stress- and context-differentiated condition. Both stress and context had a significant effect on acceptability judgements, but con-text-differentiation raised acceptability more strongly than stress-differentiation. Both the native and the main sub-group of non-native listeners judged the stimuli in essentially the same way.

Volume 2, page 975 Session B42

Phonetic Effects on Listener Detection of Vowel Concatenation

Syrdal A K
AT&T Labs - Research, USA

Concatenative speech synthesis quality depends in part on the minimization of audible discontinuities between two successive concatenated units. This study focuses on human detection of concatenation discontinuities in synthetic speech. A phonetic analysis compared the perceptual results from two voices -- one female and one male. Neither a comprehensive phonetic analysis nor a comparison of discontinuity detection between voices has been reported previously. Although discontinuities were generally more detectable for the female than the male, there were many similarities between results obtained from the two speakers. A reliably higher detection rate was observed for diphthongs than for monophthong vowels. Post-vocalic consonants influenced concatenation discontinuities significantly more than pre-vocalic consonants, and post-vocalic sonorants were associated with higher detection rates than post-vocalic non-sonorants. The differences in discontinuity detection among vowels and consonantal contexts for both voices consistently suggest that highly audible discontinuity is related to concatenation in regions of spectral change.

Volume 2, page 979 Session B42

Variable-length acoustic units inference for text-to-speech synthesis

Boeffard O
IRISA, Université Rennes 1, ENSSAT, France

The best voices in text-to-speech synthesis are currently obtained via acoustic units concatenation-based systems. In such systems, the choice of units whose concatenations will produce an acoustic message is a crucial stage. Moreover, it can be observed that current TTS systems use acoustic units which most often correspond to variable-length phonetic descriptions. In this article, an original framework is proposed which allows the automatic determination of an optimum set of variable-length acoustic units.

Volume 2, page 983 Session B42

Unit Selection for Speech Synthesis Using Splicing Costs with Weighted Finite State Transducers

Bulyko I, Ostendorf M
University of Washington, USA

In this paper we describe how unit selection for concatenative speech synthesis can be implemented efficiently for sub-phonetic units using weighted finite state transducers (WFST). We also introduce splicing costs as a measure to indicate which unit boundaries are particularly good or poor joint points. Splicing costs extend the flexibility offered by the unit selection paradigm. Through a perceptual experiment we demonstrate an improvement in speech quality achieved by using splicing costs during unit selection.

Volume 2, page 987 Session B42

Cantonese Text-To-Speech Synthesis Using Sub-Syllable Units

Law K M, Lee T, Lau W
The Chinese University of Hong Kong, Hong Kong

This paper describes our recent investigation on the use of both intra-syllable and cross-syllable acoustic units for Cantonese text-to-speech synthesis. In our previous work, isolated monosyllable units were used for concatenative speech synthesis of Cantonese. The synthetic speech was considered to be unnatural in such a way that there was an obvious lack of perceptual continuity. The proposed system adopts an acoustic inventory that covers all legitimate intra-syllable and cross-syllable acoustic units. Synthetic speech produced via concatenation of such sub-syllable units better captures the pertinent transitory effects that are crucial to perceived naturalness. Different strategies are used to concatenate speech segments with different acoustic-phonetic properties. Subjective listening test shows a noticeable performance improvement that is accounted for mainly by smoother transition between sonorant segments.

Volume 2, page 991 Session B42

Session B43 - Oral
Tuesday - 15.50 - 17.30

Signal Analysis: Microphone Arrays & Source Localisation

Chair: Maurizio Omologo, ITC-IRST, Italy

Blind Speech Separation of Moving Speakers Using Hybrid Neural Networks

Koutras A, Dermatas E, Kokkinakis G
University of Patras, Greece

In this paper we present a novel method for Blind Speech Separation of convolutive speech signals of moving speakers in highly reverberant rooms. The separation network used is a hybrid neural network, which performs separation of convolutive speech mixtures in the time domain, without any prior knowledge of the propagation media, based on the Maximum Likelihood Estimation (MLE) principle. The proposed method improves significantly (more than 13% in all adverse mixing situations) the performance of a phoneme-based continuous speech recognition system and therefore can be used as a front-end to separate simultaneous speech of speakers who are moving in reverberant rooms.

Volume 2, page 997 Session B43

Computationally Efficient Frequency-Domain Combination of Acoustic Echo Cancellation and Robust Adaptive Beamforming

Herbordt W, Buchner H, Kellermann W
University Erlangen-Nuremberg, Germany

For hands-free acoustical human/machine interfaces, e. g. for automatic speech recognition or teleconferencing systems, microphone arrays using robust Generalized Sidelobe Cancellers (GSCs) in conjunction with acoustic echo cancellation (AEC) can be efficiently applied for optimum communication. This contribution devises a new structure for combining AEC and GSC. It reduces the computational complexity by more than a factor of ten relative to a time-domain arrangement, increases convergence speed, and preserves positive synergies.

Volume 2, page 1001 Session B43

Calibration of Microphone Arrays for Improved Speech Recognition

Seltzer M L[1], Raj B[2]
[1]Carnegie Mellon University, USA, [2]Compaq Computer Corporation, USA

We present a new microphone array calibration algorithm specifically designed for speech recognition. Currently, microphone-array-based speech recognition is performed in two independent stages: array processing, and then recognition. Array processing algorithms designed for speech enhancement process the waveforms before recognition. These systems make the assumption that the best array processing methods will result in the best recognition performance. However, recognition systems interpret a set of features extracted from the speech waveform, not the waveform itself. In our calibration method, the filter parameters of a filter-and-sum array processing scheme are optimized to maximize the likelihood of the recognition features extracted from the resulting output signal. By incorporating the speech recognition system into the design of the array processing algorithm, we are able to achieve improvements in word error rate of up to 37% over conventional array processing methods on both simulated and actual microphone array data.

Volume 2, page 1005 Session B43

Improving Simultaneous Speech Recognition in Real Room Environments Using Overdetermined Blind Source Separation

Koutras A, Dermatas E, Kokkinakis G
University of Patras, Greece

In this paper we present a novel solution to the Overdetermined Blind Speech Separation (OBSS) problem for improving speech recognition accuracy of N simultaneous speakers in real room environments using M (M>N) microphones. The proposed OBSS system uses basic NxN Blind Speech Separation networks that process in parallel all different combinations of the available mixture signals in the frequency domain, resulting to lower computational complexity and faster convergence. Extensive experiments using an array of two to ten microphones and two simultaneous speakers in a simulated real room, showed that when the number of the microphones increases beyond two, the separation performance is improved and the phoneme recognition accuracy of an HMM based decoder increases drastically (more than 6%). Therefore, the introduction of more microphones than speakers is justified in order to improve speech recognition accuracy in multi simultaneous speaker environments.

Volume 2, page 1009 Session B43

Real-time Sound Source Localization and Separation System and Its Application to Automatic Speech Recognition

Asano F, Goto M, Itou K, Asoh H
AIST, Japan

A real-time sound localization/separation system for near-field sound sources was constructed and evaluated in a real office environment. As for the sound localization, the experimental results showed that the direction of the two sources was estimated with high accuracy while the range of the sources was estimated with moderate accuracy. As for the sound separation, a recognition rate of 70% for an on-line recognizer on a network and of 90% for an off-line recognizer were achieved, respectively.

Volume 2, page 1013 Session B43

Session B44 - Oral
Tuesday - 15.50 - 17.30

Speech Recognition and Understanding: Audio-Visual Processing

Chair: Hervé Bourlard, IDIAP/Swiss Federal Institute of Technology, Switzerland

An Efficient Lipreading Method Using the Symmetry of Lip

Lee J[1], Kim J[2]
[1]Waseda Univ., Japan, [2]Chonnam National University, Korea

An efficient method to reduce the amount of feature data for real-time automatic image-transform-based lipreading is proposed. Image-transform-based approach obtaining a compressed representation of image pixel values of speaker's mouth is reported to show superior lipreading performance. However, since this approach produces many feature vectors relevant to lip information, it requires much computation time for lipreading even when principal component analysis (PCA) is applied. To reduce the computational load efficiently, we propose an algorithm that utilizes the symmetry of the lip. The proposed method reduces the amount of required feature vectors up to 51% compared to the original one. Also, it improves the recognition rates by compensating the variation of illumination. With our database (22 words, 70 talkers) recorded in a natural environment, our method achieved an accuracy of 53.5% for visual-only speaker independent word recognition task. The extracted features are modeled by hidden Markov models with Gaussian mixture distributions.

Volume 2, page 1019 Session B44

Comparing Audio- and A-Posteriori-Probability-Based Stream Confidence Measures for Audio-Visual Speech Recognition

Heckmann M[1], Wild T[1], Berthommier F[2], Kroschel K[1]
[1]Universität Karlsruhe, Germany, [2]Institut de la Communication Parlée, France

During the fusion of audio and video information for speech recognition, the estimation of the reliability of the noise affected audio channel is crucial to get meaningful recognition results. In this paper we compare two types of reliability measures. One is the use of the statistics of the phoneme a-posteriori probabilities and the other is the analysis of the audio signal itself. We implemented the entropy and the dispersion of the probabilities and, from the audio-based criteria, the so called Voicing Index. To test the criteria a hybrid ANN/HMM audio-visual recognition system was used and 5 different types of noise at 12 SNR levels each were added to the audio signal. The best sigmoidal fit for each criterion between the fusion parameter and the value of the criterion over all noise types and SNR values was performed. The resulting individual errors and the corresponding averaged relative errors are given.

Volume 2, page 1023 Session B44

Large-vocabulary audio-visual speech recognition by machines and humans

Potamianos G, Neti C, Iyengar G, Helmuth E
IBM T.J. Watson Research Center, USA

We compare automatic recognition with human perception of audio-visual speech, in the large-vocabulary, continuous speech recognition (LVCSR) domain. Specifically, we study the benefit of the visual modality for both machines and humans, when combined with audio degraded by speech-babble noise at various signal-to-noise ratios (SNRs). We first consider an automatic speechreading system with a pixel based visual front end that uses feature fusion for bimodal

integration, and we compare its performance with an audio-only LVCSR system. We then describe results of human speech perception experiments, where subjects are asked to transcribe audio-only and audio-visual utterances at various SNRs. For both machines and humans, we observe approximately a 6 dB effective SNR gain compared to the audio-only performance at 10 dB, however such gains significantly diverge at other SNRs. Furthermore, automatic audio-visual recognition outperforms human audio-only speech perception at low SNRs.

Volume 2, page 1027 Session B44

Evaluation of an Automatically Obtained Shape and Appearance Model For Automatic Audio Visual Speech Recognition

Daubias P, Deleglise P
LIUM, France

In this paper, we first present a shape and appearance model for Audio-Visual Automatic Speech Recognition. The shape model is a template (mean shape) and a set of deformation vectors to transform it into any possible shape. The global appearance model is a neural network trained to classify 5*5 colour image blocks as from skin, lips or inside of mouth. Both parts of this model were built automatically (without hand-labelling). Appearance model was built using speech bimodality (acoustic information). We then propose several measures for quality evaluation of lip location. Finally, we show the classification results obtained using a hand-labelled and two automatically built appearance models of the lips.

Volume 2, page 1031 Session B44

An approach to an Italian Talking Head

Pelachaud C[1], Magno-Caldognetto E[2], Zmarich C[2], Cosi P[2]
[1]Universita di "La Sapienza", Italy, [2]C.N.R. of Padova , Italy

Our goal is to create a natural talking face with, in particular, lip-readable movements. Based on real data extracted from an Italian speaker with the ELITE system, we have approximated the data using radial basis functions. In this paper we present our 3D facial model based on MPEG-4 standard and our computational model of lip movements for Italian. Our experiment is based on some phonetic-phonological considerations on the parameters defining labial orifice, and on identification tests of visual articulatory movements.

Volume 2, page 1035 Session B44

Session B45 - Poster
Tuesday - 15.50 - 17.30

Linguistic Modelling: Language Models - II

Chair: To be decided,

Pause Information for Dependency Analysis of Read Japanese Sentences

Takagi K, Ozeki K
The University of Electro-Communications, Japan

The work presented in this paper is devoted to the modeling of distribution of post-phrase pause duration for dependency analysis of read Japanese sentences. The pause information is incorporated into a dependency structure parser, which handles numerical information as part of linguistic knowledge. A series of our previous works has shown that the duration of pauses is constantly effective to improve the parsing accuracy among other prosodic features. This paper aims to improve the parsing accuracy by reforming pause distribution functions to better fit the actual distribution of pause duration. First, the pause distribution is divided into two parts, each of which is represented by a separate model: one by a discrete probability model, and the other by a bimodal p.d.f. Secondly, pause duration is calculated in log scale. All the experiments show that these modifications improve parsing accuracy.

Volume 2, page 1041 Session B45

An HMM/N-gram-based Linguistic Processing Approach for Mandarin Spoken Document Retrieval

Chen B, Wang H-M, Lee L-S
Institute of Information Science, Academia Sinica, Taiwan, ROC

In this paper an HMM/N-gram-based linguistic processing approach for Mandarin spoken document retrieval is presented. The underlying characteristics and different structures of this approach were extensively investigated. The retrieval capabilities were verified by tests with indexing features of word- and syllable(subword)-levels and comparison with the conventional vector space model approach. To further improve the discrimination capabilities of the HMMs, both the expectation-maximization (EM) and minimum classification error (MCE) training algorithms were introduced in training. The information fusion of indexing features of word- and syllable-levels was also investigated. The spoken document retrieval experiments were performed on the Topic Detection and Tracking Corpora (TDT-2 and TDT-3). Very encouraging retrieval performance was obtained.

Volume 2, page 1045 Session B45

Probabilistic Concept Verification for Language Understanding in Spoken Dialogue Systems

Lin Y-C, Wang H-M
Industrial Technology Research Institute, Taiwan

In the past researches, several kinds of information have been explored to assess the confidence measure or to select the confidence tag for a word/phrase. However, the contextual confidence information is little touched. In this paper, we propose a concept-based probabilistic verification model to integrate the contextual confidence information. In this model, a concept is verified not only according to its acoustic confidence measure but also according to neighboring concepts and their confidence levels. Experimental results show that the proposed model significantly outperforms the model using only confidence measures. The error rate of confidence tag is reduced from 17.7% to 15.12%, which corresponds to an error reduction rate of 14.5%.

Volume 2, page 1049 Session B45

Turkish Word Segmentation Using Morphological Analyzer

Külekcý M O[1], Özkan M[2]
[1]*National Research Institute of Electronics & Cryptology, Turkey*,
[2]*Bogazici Univ., Turkey*

This paper describes an algorithm to segment an input Turkish string without any spaces, which may be an output of a speech-to-text application, into words by using morphological analyser. It is quite possible to use the algorithm on other languages, which has a morphological analysis component, as well. Turkish morphological analyser is designed and implemented as the linguistic engine of the algorithm. The construction of the analyser proposes a technique that attempts to achieve group vise morpheme recognition instead of searching suffixes one by one in a word.

Volume 2, page 1053 Session B45

Thai Grapheme-To-Phoneme Using Probabilistic GLR Parser

Tarsaku P, Sornlertlamvanich V, Thongprasirt R
NECTEC, Thailand

Many difficulties in the Thai language such as the absence of boundary word, linking syllables in pronunciation, and homographs are challenging us in developing a Thai Grapheme-to-Phoneme (G2P) converter. Presently there are a couple Thai G2P systems which are proposed in ruled-based and decision-tree approach. The rule-based approach has a drawback in the limitation of employing the context. The decision-tree approach is somehow able to capture the local context for making the decision. On the contrary, the Probabilistic Generalized LR (PGLR) approach is reported that both the global and local context are efficiently captured in the probabilistic model. In this paper, we implement a Thai G2P system based on the PGLR approach. The result of experiment shows 90.44% of word accuracy in case of ignoring vowels length and 72.87% of word accuracy in case of exact match evaluation. These results are superior to those of rule-based and decision-tree approaches.

Volume 2, page 1057 Session B45

Aligning Prosody and Syntax in Property Grammars

Blache P, Hirst D
LPL-CNRS, France

We propose in this paper a new approach for representing the prosody/syntax interface. We use for this a particular formalism, called {em Property Grammars}, in which all information is represented by means of constraints. We show how alignment constraints can implement such an interface. One of the interests of these constraints, in comparison with other approaches such as optimality theory, is the possibility of representing different information at the same level (allowing then a parallel treatment of prosody and syntax). This discussion is illustrated with the example of dislocated constructions.

Volume 2, page 1061 Session B45

From Perceptual Designs to Linguistic Typology and Automatic Language Identification : Overview and Perspectives

Barkat M, Vasilescu I
University of Lyon 2, France

In the last years, researches in human identification of languages benefit from a special attention as an alternative to improve the robustness of automatic systems. In this scope, two fundamental goals are pursued: first, to highlight the perceptual strategies used by subjects during an experimental language and/or dialectal identification task and secondly, to identify a set of discriminative cues corresponding to different linguistic levels. Methodologically speaking, two different approaches emerge: one can find experimental designs using natural speech aiming at determining global discriminative cues and/or designs based on modified-speech aiming at isolating specific linguistic levels. These experiments lack of methodology as for the choice of studied languages as well as for subjects' mother tongue. This aspect being all the more worsened by a hazy spotting and a limited exploitation of discriminative criteria. We suggest an original approach based on a two-step methodology integrating to perception "genetic" considerations and resulting into the modeling of perceptually identified discriminative cues.

Volume 2, page 1065 Session B45

Morphological Approaches for an English Pronunciation Lexicon

Fitt S
Univ. of Edinburgh, UK

Most pronunciation lexica for speech synthesis in English take no account of morphology. Here we demonstrate the benefits of including a morphological breakdown in the transcription. These include maintaining consistency, developing the symbol set and providing the environmental description for allophones and phonetic variables. Our approach does not use a full morphological generator, but includes morphological boundaries in the lexicon.

Volume 2, page 1069 Session B45

An Embodiment Paradigm for Speech Recognition Systems

Joue G, Carson-Berndsen J
University College Dublin, Ireland

The problems of conventional speech recognition approaches include incomplete linguistic knowledge and inability to deal with underspecification. These issues can be addressed by understanding the constraints of speech to predict speech tendencies. Understanding what constraints exist requires an embodied view of speech and that the traditional disembodied view of speech is the fundamental limitation on the robustness of many speech systems. Viewing speech as a form of embodied cognition, or within context of its production and use, provides important insights for speech recognition. In making this claim, this paper briefly outlines a strongly embodied account of cognition and develops from that an embodiment paradigm for speech recognition. The embodiment paradigm proposed leads to both an explanatory and descriptive account of linguistic structure. It simplifies the view of speech structure for automatic speech recognisers, by considering only the most directly relevant motivations or constraints influencing communication and thus speech.

Volume 2, page 1073 Session B45

Multi-Parser Architecture for Query Processing

Kui X[1], Fuliang W[1], Helen M[2], Luk P C[2]
[1]*Intel China Research Center, P. R. China*, [2]*The Chinese University of Hong Kong, P. R. China*

Natural language queries provide a natural means for common people to interact with computers and access to on-line information. Due to the complexity of natural language, the traditional way of using a single grammar for a single language parser leads to an inefficient, fragile, and often very big language processing system. Multi-Parser Architecture (MPA) intends to alleviate these problems, and the modularized MPA also has the advantage of easier portability to new domains and distributed computing. In this paper, we investigate the effect of using different types of parsers on different types of query data in MPA. Three data sets and two types of sub-parsers, particularly a predictive cascading composition for pre-compiled Earley parsers , have been examined. Results show that partitioning grammars leads to superior speed

performance for the Earley-style parser across the three data sets. GLR parser is faster than Earley parser in the partitioned case, but it can lead to an excessive memory usage for the un-partitioned case.

Volume 2, page 1077 Session B45

Two-Stage Probabilistic Approach to Text Segmentation

Chen Y-C, Lin Y-C
Industrial Technology Research Institute, Taiwan

For telephone-based spoken dialogue systems, the responses to users should be specific and short. Therefore, it is highly demanded to segment a topical text into specific event segments which can be use to answer users¡| queries. However, the lexical cohesion approach, which has been widely used to segment text into topics, is not suitable for segmenting text into smaller units, like events. In this paper, we present a two-stage approach to partition text into event segments. In the first stage, a trigram chunk tagger is used to label the segmentation tags. In the second stage, the unreliable segmentation tags are detected and then verified by a probabilistic verification model. Compared with the chunk tagger, the verification model can explore more contextual information and is less sensitive to the sparseness of training data. Experimental results show that the proposed two-stage approach significantly outperforms the chunk tagger approach. The improvements on precision and recall rates are 27% to 83% in different testing tasks.

Volume 2, page 1081 Session B45

Lexicon Optimization for Dutch Speech Recognition in Spoken Document Retrieval

Ordelman R, Hessen, van A, Jong, de F
University of Twente, The Netherlands

In this paper, ongoing work concerning the language modelling and lexicon optimization of a Dutch speech recognition system for Spoken Document Retrieval is described: the collection and normalization of a training data set and the optimization of our recognition lexicon. Effects on lexical coverage of the amount of training data, of decompounding compound words and of different selection methods for proper names and acronyms are discussed.

Volume 2, page 1085 Session B45

Evaluation of Recent Speech Grammar Standardization Efforts

Brøndsted T
Aalborg University, Denmark

The "Voice Browser " activity within the W3C consortium addresses the need for standards for speech grammars, dialogue descriptions etc. in distributed systems. This paper discusses the consortium's recent speech grammar working draft specification. The W3C specification is based on the Javatm Speech Grammar Format (JSGF) defined by Sun Microsystems and is - with all good and bad qualities - characterized by traditions of formal language theory. In a constructive spirit, we suggest some possible improvements based on natural language theory. The suggestions concern compound feature-based semantic presentations, lexicon structure, ambiguity, and exhaustive parsing,

Volume 2, page 1089 Session B45

Session B46 - Poster
Tuesday - 15.50 - 17.30

Speech Recognition and Understanding: Noise Robustness - II

Chair: Chin Hui Lee, Bell Labs, Lucent Tech., USA

Robust Speech Recognition against Packet Loss

Siu M, Chan Y C
Hong Kong University of Science and Technology, Hong Kong

Recognizing speech transmitted over mobile or computer networks poses new challenges such as packet loss in transmission. Viterbi algorithm, the most common speech recognition approach, seaches for the most likely state sequence that explains all observation. However, because it implicitly sums the log observation probabilities, the resulting solution is sensitive to outlier frames. In this paper, we propose a robust approach that searches the state sequence that best explains x percent of the observation and is insensitive to the corruption of a limited number of frames. We evaluated the proposed algorithm on the TI-digits task. With 10% of the data loss, the proposed algorithm achieves improvement of 71.6% for isolated digit recognition and 32.2% for connected digit recognition.

Volume 2, page 1095 Session B46

Rapid CODEC Adaptation for Cellular Phone Speech Recognition

Naito M, Kuroiwa S, Kato T, Shimizu T, Higuchi N
KDDI R&D Laboratories, Japan

Along with the popularization of cellular phone, it becomes important issue to improve recognition accuracy for cellular phone speech input. However, the distortion caused by current low-bit rate speech corder is nonlinear. Therefore, it is difficult to compensate these distortion by only applying conventional CMN which assuming distortion as stationary linear transfer on spectrum domain. In this paper, to improve the accuracy of speech recognition over cellular-phone network, we investigate the use of CODEC-dependent acoustic model and rapid CODEC-adaptation using model selection based on maximum likelihood criterion. These method reduce degradation of recognition performance due to difference in CODEC by 33%.

Volume 2, page 1099 Session B46

A robust front-end for ASR over IP and GSM networks: an integrated scenario

Gallardo-Antolin A, Pelaez-Moreno C, Diaz-de-Maria F
University Carlos III, Madrid, Spain

Both for the transmission over GSM and IP networks, voice must be encoded at the originating end and subsequently decoded at the receiving end. This lossy coding produces a quality deterioration, that though acceptable for a human being, seriously affects the performance of Automatic Speech Recognizers (ASR) when they are not specifically designed for operating under those conditions. The authors have already introduced and tested a new robust front-end which improves ASR performances by simulating both networks environments. Here we complete some previous results including realistic GSM models, compare both scenarios and put forward and integrated scenario where mobile GSM devices require the services of an ASR facility situated into an IP network.

Volume 2, page 1103 Session B46

Robust Speech Recognition using Missing Feature Theory and Vector Quantization

Renevey P, Vetter R, Krauss J
CSEM, Switzerland

This paper addresses the problem of speech recognition in noisy conditions when low complexity is required like in embedded systems. In such systems, vector quantization is generally used to reduce the complexity of the recognition systems (e.g. HMMs). A novel approach for vector quantization based on the missing data theory is proposed. This approach allows to increase the robustness of the system against the noise perturbations with only a small increase of the computational requirements. The proposed algorithm is composed of two parts. The first part consists in dividing the spectral temporal features of the noisy signal into two subspaces: the unreliable (or missing) features and the reliable (or present) features. The second part of the proposed approach consists in defining a robust distance measure for vector quantization that compensates for the unreliable features. The proposed approach obtains similar results in noisy conditions than a more classical approach that consists in adapting the codebook of the vector quantization to the noisy conditions using model compensation. However the computation requirements are lower in the proposed approach and it is more suitable for a low complexity speech recognition system.

Volume 2, page 1107 Session B46

Modeling the Mixtures of Known Noise and Unknown Unexpected Noise for Robust Speech Recognition

Ming J, Jancovic P, Hanna P, Stewart D
Queens University of Belfast, UK

Real-world noise may be a mixture of known or trainable noise and unknown unexpected noise. This paper investigates the combination of the conventional noise-reduction techniques with the probabilistic union model to deal with this type of mixed noise for robust speech recognition. In particular, we have developed a multi-environment system to remove the known or trainable acoustic mismatch across different environments. The novelty of this system, in contrast to other multi-environment models, is that the acoustic model for each environment is built upon the probabilistic union model, so that this system is also capable of accommodating further unknown unexpected noise within a specific environment. We have tested the new system for connected digit recognition in different environments, each involving an environment-specific noise and some unknown untrained noise. The results indicate that the new system offers significantly improved performance for the environments involving unknown additional noise, in comparison to a baseline multi-environment system.

Volume 2, page 1111 Session B46

Robust Speech Recognition based on Selective Use of Missing Frequency Band HMMs

Kawamura T, Takeda K, Itakura F
Nagoya University, Japan

In this paper, we propose a multi-stream approach that selectively uses Missing Frequency Band HMMs (MFB-HMM) that is trained on the band-eliminated speech. This makes the model insensitive to the noise in the missing frequency band. With multiple MFB-HMMs of different missing frequency bands, the proposed recognition system is robust in various types of noise conditions. Recognition experiments show that the selective use of the MFB-HMMs is very effective in narrow band noise condition even if the noise is unstationary, however, the improvements of the performance to general noisy conditions, e.g. in-car noise and music sound, are not as high as in the narrow band noise case. The results of the experiments also show that the optimal selection of the MFB-HMM significantly improves the performance regardless of the type of the noise; therefore, the model selection measure is the key issue in this method.

Volume 2, page 1115 Session B46

A New Method for Speech Recognition in the Presence of Non-stationary, Unpredictable and High-level Noise

Masuda-Katsuse I
Kyushu Institute of Design / Institute of Systems & Information Technologies, Japan

We propose a new method for speech recognition in the presence of non-stationary, unpredictable and high-level noise by extending PreFEst (Predominant-F0 Estimation Method) which was developed to estimate melody and bass lines from music signals. The proposed method does not need to know about noise characteristics in advance and does not even estimate them in its process. A small set of evaluations demonstrates the feasibility of the method by showing a good performance even when the background noise is real and non-stationary noise and the SNR is less than 10 dB.

Volume 2, page 1119 Session B46

A Computational Efficient Real Time Noise Robust Speech Recognition Based on Improved Spectral Subtraction Method

Kotnik B, Kacic Z, Horvat B
University of Maribor, Slovenia

In this paper, a speech enhancement method is presented, which uses spectral and time domain processing and achieves a trade-off between effective noise reduction and low computational load for real-time operations. First, a spectral subtraction method is used to reduce the effect of additive broadband noise on speech. Then, a novel weighting function is used to reduce a residual "musical noise" in time domain. This weighting function is a compound of a short-time zero crossing value and a short-time energy of speech signal. For evaluation of improvement of speech recognition the Slovenian SpeechDat FDB, the German SpeechDat FDB and SpeechDat-Car, as well as the Spanish SpeechDat FDB databases together with the HTK recognition toolkit were used. Word recognition accuracy in connected digits recognition task was improved by 8.7% for Slovenian FDB, by 5.1% for Spanish FDB, by 3.2% for German SpeechDat-Car, and by 2% for German SpeechDat FDB database.

Volume 2, page 1123 Session B46

The Use of Noisy Frame Elimination and Frequency Spectrum Magnitude Reduction in Noise Robust Speech Recognition

Vlaj D, Kacic Z, Horvat B
University of Maribor, Slovenia

In this paper the procedure for feature vector extraction and the problems, which must be solved, by defining the feature vectors, which contain only the information about the speech signal are described. A new procedure of feature extraction which is based on the frame elimination and frequency spectrum reduction for the noisy part of the speech signal is proposed. For all tests the Slovenian telephone speech database SpeechDat II was used. The connected digits were used for both, training and testing. There were 800 speakers used for training and 200 for testing. The word recognition accuracy was increased for 3.1 percentage points with the new procedure, and this was achieved, when the number of Gaussian mixtures was four times smaller than with the ordinary method. The results obtained are especially encouraging for the systems where the size of the available memory and processing power are limited (for example, mobile phones).

Volume 2, page 1127 Session B46

Combined Linear Regression Adaptation and Bayesian Predictive Classification for Robust Speech Recognition

Chien J-T
National Cheng Kung University, Taiwan, ROC

The uncertainty in parameter estimation due to the adverse environments deteriorates the speech recognition performance. It becomes crucial to incorporate the parameter uncertainty into decision so that the classification robustness can be assured. In this paper, we propose a linear regression based Bayesian predictive classification (LRBPC) for robust speech recognition. This framework is constructed under the paradigm of linear regression adaptation of HMM's. Because the regression mapping between HMM's and adaptation data is ill posed, we properly characterize the uncertainty of regression parameters using a joint Gaussian distribution. A predictive distribution is derived to set up the LRBPC decision. Such decision is robust compared to the plug-in maximum a posteriori decision adopted in the maximum likelihood linear regression (MLLR). Since the specified distribution belongs to the conjugate prior family, the evolutionary hyperparameter is established. With the hyperparameter, the LRBPC achieves significantly better performance than MLLR adaptation in car speech recognition.

Volume 2, page 1131 Session B46

Quantile Based Histogram Equalization for Noise Robust Speech Recognition

Hilger F, Ney H
Lehrstuhl fuer Informatik VI, RWTH Aachen - University of Technology, Germany

This paper describes an approach to increase the noise robustness of automatic speech recognition systems by, transforming the signal after Mel scaled filtering, to make the cumulative density functions of the signal's values in recognition match the ones that where estimated on the training data. The cumulative density functions are approximated using a small number of quantiles. Recognition tests on several databases showed significant reductions of the word error rates. On a real life database recorded in driving cars with a large mismatch between the training and testing conditions the relative reductions of the word error rates where over 60%.

Volume 2, page 1135 Session B46

Sequential Noise Compensation by A Sequential Kullback Proximal Algorithm

Yao K[1], Paliwal K K[2], Nakamura S[1]
[1]ATR Spoken Language Translation Research Labs., Japan, [2]Griffith University, Australia

We present a sequential noise compensation method based on the sequential Kullback proximal algorithm, which uses the Kullback-Leibler divergence as a regularization function for the maximum likelihood estimation. The method is implemented as filters. In contrast to sequential noise compensation method based on the sequential EM algorithm, the convergence rate of the method and estimation error after convergence can be adjusted by a relaxation factor, where the sequential EM algorithm corresponds to the particular case of the presented algorithm. Through experiments on parameter estimation and speech recognition in noise, we verified the efficacy of the algorithm.

Volume 2, page 1139 Session B46

Session C11 - Oral
Wednesday - 09.00 - 10.40

ESE4 - SIGshow

Chair: Gerrit Bloothooft, University of Utrecht, The Netherlands

Education on the Web: Launch of Three New Websites (no proceedings paper)

Eriksson A[1], Bloothooft G[2]
[1]Stockholm University, Sweden, [2]Utrecht University, The Netherlands

The ISCA Special Interest Group on Education will launch three websites during the presentation: (1) JEWELS (Joint European Website for Education in Language and Speech, sponsored by ELSNET and supported by ISCA, EACL and Socrates projects) which provides information on contents of studies, educational materials and tools, links to sites and courses, and information on European Educational support programmes, (2) a new professional website of the European Masters in Language and Speech, and (3) the website of EduSIG itself.

Volume 2, page 1144 Session C11

SPeaker and Language characterization (SpLC): A Special Interest Group (SIG) of ISCA

Bonastre J-F[1], Magrin-Chagnolleau I[1], Euler S[2], Pellegrino F[3], André-Obrecht R[4], Mason J[5], Bimbot F[6]
[1]LIA University of Avignon, France, [2]Robert-C Bosch, Germany, [3]DDL, University of Lyon II, France, [4]IRIT, University Paul Sabatier, France, [5]University of Swansea, UK, [6]IRISA/INRIA, France

Last year, SPLC - an ISCA Special Interest Group centered around Speaker and Language Characterization born. The aims of this paper are to present the SPLC SIG, its the objectives and the work done during the first year.

Volume 2, page 1145 Session C11

SProSIG: A Special Interest Group on Speech Prosody (no proceedings paper)

Hirst D[1], Bel B[1], Campbell N[2]
[1]CNRS, Aix en Provence, France, [2]ATR, Kyoto, Japan

This presentation will present the activities of SProSIG since its creation in January 2000 including the setting up of an email list, dedicated web pages, and the planning of an International Conference on Speech Prosody (Speech Prosody 2002) to be held in Aix en Provence in April 2002. A number of ideas for future activities will also be presented for discussion.

Volume 2, page 1144 Session C11

The ISCA Special Interest Group on Speech Synthesis

Campbell N[1], Hess W[2], Möbius B[3], van Santen J[4]
[1]ATR Spoken Language Translation Research Laboratories, Japan, [2]University of Bonn, Germany, [3]University of Stuttgart, Germany, [4]Oregon Graduate Institute, USA

This paper describes the constitution and activities of the ISCA Speech Synthesis Special Interest Group, SynSIG. It summarises past achievements and suggests ways in which future development could be maintained. The aims of the Special Interest Group on Speech Synthesis are to promote the study and diffusion of knowledge about speech synthesis in general, in a number of ways including: dedicated web pages, a mailing list, a bibliographic database, organisation of workshops on specific themes, exchange of students, and helping to co-ordinate sessions on speech synthesis in international conferences and workshops. The international and multi-disciplinary nature of the SIG also provides

a means for diffusing information both to and from the different research communities involved in the synthesis of various languages.

Volume 2, page 1149 Session C11

Auditory Visual Speech Processing

Massaro D W
University of California, Santa Cruz, USA

This paper provides an overview of the developments in Auditory Visual Speech Processing, a Special Interest Group within Eurospeech. I hope that this discussion will be informative and useful to readers in a variety of fields, including psychology, speech science, animation, psycholinguistics, human-machine interaction, hearing-impaired communication, and numerous other fields which also share in this fruitful intersection.

Volume 2, page 1153 Session C11

Session C12 - Oral
Wednesday - 09.00 - 10.40

Speech Synthesis: Prosody

Chair: Nick Campbell, ATR, Japan

Training Prosodic Phrasing Rules for Chinese TTS Systems

Chen W, Lin F, Li J, Zhang B
Tsinghua University, P. R. China

This paper describes several experiments designed to train prosodic phrasing models for Chinese TTS systems and to investigate the underlying rules that control Chinese prosody. First, we collected 559 sentences from news programs and built a large corpus for modeling Chinese prosody. Second, we selected 20 features and used classification and regression trees (CART) and transformational rule-based learning (TRBL) techniques to generate phrasing rules automatically. Lastly, we propose a computer aided error-driven method of designing rule templates, and integrate it into the TRBL algorithm. The experimental results show that we achieve a high success rate of 94.5%, and we also get a set of well comprehensible rule templates which may give us insights into the relationship between Chinese syntax and prosody.

Volume 2, page 1159 Session C12

Intonation Modelling with a Lexicon of Natural F0 Contours

Heggtveit P O, Natvig J E
Telenor R&D, Norway

We describe a new approach for generating Norwegian intonation in text to speech synthesis. The method is based on a phonological representation of utterances. The overall f0 contour of an utterance is synthesised by concatenation of stored f0 contours corresponding to accent units. Candidate accent units are found by searching a lexicon derived from natural speech and selecting the unit that is the best match with respect to the properties of the target accent units of the utterance to be synthesised. A formal subjective test confirms that the new approach leads to more natural speech than a former rule based method, but the quality is still inferior to intonation copied from natural speech.

Volume 2, page 1163 Session C12

Smooth Contour Estimation in Data-Driven Pitch Modelling

Silverman K[1], Bellegarda J[1], Lenzo K[2]
[1]Apple Computer, USA, [2]Carnegie-Mellon University, USA

Apple's next-generation text-to-speech system in MacOS X uses a superpositional pitch model, comprising a relatively smooth underlying F0 contour and a separate contribution from the influence of the phonetic segments. This paper focuses on the data-driven modelling of the underlying contour, based on electroglottographic signals obtained from a corpus of reiterant speech. F0 extraction from such signals leads to more accurate characteristic shapes, as objectively illustrated by a typically low mean absolute frequency deviation (between 2 and 3 Hz) between original and synthetic F0 contours. This in turn supports a better (both more complete and more realistic) model of F0 behavior. Experimental results illustrate the improved prosodic representation resulting from this F0 model.

Volume 2, page 1167 Session C12

Generating F0 Contours by Statistical Manipulation of Natural F0 Shapes

Saito T, Sakamoto M

IBM Research, Tokyo Research Laboratory, Japan

This paper proposes a method of generating F0 contours from natural F0 segmental shapes for speech synthesis. The extracted shapes of F0 units are basically kept unchanged, by eliminating any averaging operation in the analysis phase and minimizing modification operations in the synthesis phase. The use of gkept-unchanged h F0 shapes has a great potential to incorporate a wide variety of speaking styles in the same framework, including not only read-out speech, but also dialogue and emotive speech. A linear-regression statistical model is proposed here to gmanipulate h the stored raw F0 shapes for building them up to a sentential F0 contour. Through experimental evaluations, the proposed model turns out to provide a robust F0 contour prediction. By using the model, linguistically derived information of a sentence can be directly mapped, in a purely data-driven manner, to acoustic F0 values of the sentential intonation contour for a trained speaker.

Volume 2, page 1171 Session C12

Learning Prosodic Features using a Tree Representation

Hirschberg J, Rambow O
AT&T Labs Research, USA

We describe experiments designed to learn associations between two types of intonational features, pitch accent and phrasing, from a tree-based corpus annotated with various intonational and syntactic features, for a concept-to-speech system. We show that using novel tree-based features improves the quality of boundary prediction over using only the linear order-based features normally used in text-to-speech.

Volume 2, page 1175 Session C12

Session C13 - Oral
Wednesday - 09.00 - 10.40

Applications: Multimodal Applications

Chair: Nikos Fakotakis, Univ. of Patras, Greece

Lip-Reading from Parametric Lip Contours for Audio-Visual Speech Recognition

Gurbuz S, Patterson E K, Tufekci Z, Gowdy J N
Clemson University, USA

This paper describes the incorporation of a visual lip tracking and lip-reading algorithm that utilizes the affine-invariant Fourier descriptors from parametric lip contours to improve the audio-visual speech recognition systems. The audio-visual speech recognition system presented here uses parallel hidden Markov models (HMMs), where a joint decision, using an optimal decision rule, is made after processing. This work describes the extraction of affine-invariant Fourier descriptors (AI-FDs) from parametric lip contour data. Finally, this work validates the use of optimal weight selection, which is based on the noise type and signal-to-noise ratio (SNR) for joint audio-visual automatic speech recognition (JAV-ASR).

Volume 2, page 1181 Session C13

An Investigation of HMM Classifier Combination Strategies for Improved Audio-Visual Speech Recognition

Lucey S, Sridharan S, Vinod C
Queensland University of Technology, Australia

The combining of independent audio and visual HMM classifiers (late integration) has been shown to out perform the combination of audio and visual features in a single HMM classifier (early integration) when either or both modalities are presented with distortion for the task of speech recognition. Theoretical foundations for the optimal combination of these audio and video classifiers are still unclear. In this paper a number of strategies for combining these classifiers are investigated. An argument for using a hybrid of the sum and product rules is made based on empirical, theoretical and heuristic evidence.

Volume 2, page 1185 Session C13

Combining Multi-Party Speech and Text Exchanges over the Internet

Bernsen N O, Dybkjær L
University of Southern Denmark, Denmark

Bilateral or group text chatting over the Internet has become a favoured pastime for many people across the world. Yet it would seem that, in general, text chat is a severely impoverished mode of on-line communication compared to, e.g., fully situated human-human spoken conversation, video conferencing, or even speaking over the telephone. This paper explores what happens when on-line multi-speaker conversation over the Internet is added to text chat, creating what may become a widespread mode of communication in near future. The system used is called Magic Lounge. The paper presents rather clear-cut results on the respective communicative roles of speech and text chat from a series of user tests with the system in which different groups of users performed scenarios designed to explore the combined use of text chat and speech. The results reported may generalise to a wide range of applications which combine text and spoken information representation.

Volume 2, page 1189 Session C13

Real-Time Multiple Speaker Tracking by Multi-Modal Integration for Mobile Robots

Nakadai K[1], Hidai K-I[1], Okuno H G[2], Kitano H[3]
[1]Japan Science and Technology Corp., Japan, [2]Kyoto University, Japan, [3]Sony Computer Science Laboratories, Japan

In this paper, real-time multiple speaker tracking is addressed, because it is essential in robot perception and human-robot social interaction. The difficulty lies in treating a mixture of sounds, occlusion (some talkers are hidden) and real-time processing. Our approach consists of three components; (1) the extraction of the direction of each speaker by using interaural phase difference and interaural intensity difference, (2) the resolution of each speaker's direction by multi-modal integration of audition, vision and motion with canceling inevitable motor noises in motion in case of an unseen or silent speaker, and (3) the distributed implementation to three PCs connected by TCP/IP network to attain real-time processing. As a result, we attain robust real-time speaker tracking with 200 ms delay in a non-anechoic room, even when multiple speakers exist and the tracking person is visually occluded.

Volume 2, page 1193 Session C13

XISL: An Attempt to Separate Multimodal Interactions from XML Contents

Nitta T, Katsurada K, Yamada H, Nakamura Y, Kobayashi S
Toyohashi University of Technology, Japan

In this paper we outline a multimodal interaction description language XISL (Extensible Interaction-Sheet Language) that is developed to describe multimodal interactions (MMI), and to separate the description of interactions from XML contents. XISL makes an XML document independent of interactions that may differ between each terminal, and so enables such seamless services as web-browsing to be constructed easily. Since XISL also provides various combinatorial usage of modalities, a developer can describe a MMI scenario easily. We implemented an interpreter of XISL on prototype systems for multimedia lectures with different types of MMI and proved the viability of XISL by experiments using the systems.

Volume 2, page 1197 Session C13

Session C14 - Oral
Wednesday - 09.00 - 10.40

Speech Recognition and Understanding: Speaker Adaptation

Chair: Jean-Claude Junqua, Panasonic, USA

Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression

Gunawardana A, Byrne W
The Johns Hopkins University, USA.

We present a simplified derivation of the extended Baum-Welch procedure, which shows that it can be used for Maximum Mutual Information (MMI) of a large class of continuous emission density hidden Markov models (HMMs). We use the extended Baum-Welch procedure for discriminative estimation of MLLR-type speaker adaptation transformations. The resulting adaptation procedure, termed Conditional Maximum Likelihood Linear Regression (CMLLR), is used successfully for supervised and unsupervised adaptation tasks on the Switchboard corpus, yielding an improvement over MLLR. The interaction of unsupervised CMLLR with segmental minimum Bayes risk lattice voting procedures is also explored, showing that the two procedures are complimentary.

Volume 2, page 1203 Session C14

What is the Best Type of Prior Distribution for EMAP Speaker Adaptation?

Kenny P, Boulianne G, Dumouchel P
Centre de Recherche Informatique de Montreal, Canada

There are two types of prior distribution that can be viewed as natural for extended MAP (or EMAP) speaker adaptation. One arises from modeling the correlations between speakers (assumed to be constant across HMM Gaussians) and the other from modeling the correlations between HMM Gaussians (assumed to be constant across speakers). In this paper we present new results establishing the usefulness of correlations of the first type for speaker adaptation and we outline a tensor product construction which enables both types of correlation to be integrated in a common mathematical framework. We also present the results of some experiments which suggest that the two types of correlation are equally effective for speaker adaptation and that there is no incremental improvement to be gained by modeling both of them simultaneously.

Volume 2, page 1207 Session C14

Maximum-Likelihood Affine Cepstral Filtering (MLACF) Technique for Speaker Normalization

Kim Y
Stanford University, USA

We present a novel technique of minimizing the acoustic variability of speakers by transforming the features extracted from the speaker's data to better fit the recognition model. The concept of maximum-likelihood affine cepstral filtering (MLACF) will be introduced for feature transformation, along with solutions for the transformation parameters that maximize the likelihood of the test data with respect to a given recognition model. It is shown that for log-concave distributions, the solution of the MLACF problem can be obtained using convex programming. HMM-based digit recognition on the TIDIGITS database is presented to demonstrate the flexibility of the transformation in compensating for large acoustic mismatches between the speakers in the training and test database. In addition, it will be shown that the technique requires estimation of far fewer transformation parameters compared to existing techniques, thus allowing fast, real-time compensation.

Volume 2, page 1211 Session C14

A Novel Algorithm For Rapid Speaker Adaptation Based On Structural Maximum Likelihood Eigenspace Mapping

Zhou B, Hansen J
Univ. of Colorado at Boulder, USA

In this paper, we propose a novel algorithm for rapid speaker adaptation based on our Structural Maximum Likelihood Eigenspace Mapping (SMLEM). The proposed method constructs a binary-tree structured hierarchical Speaker Independent (SI) eigenspace at different levels from well-trained SI system models, and then dynamically constructs a new set of speaker dependent (SD) eigenspaces at corresponding levels, according to the availability of incoming adaptation data. By mapping the mixture Gaussian components from a SI eigenspace to SD eigenspaces in a maximum likelihood manner, the SI models are adapted towards SD models (EM algorithm is used to derive the eigenspace bias). Compared with conventional MLLR, the proposed algorithm is both computationally cheaper and more effective when only a very small amount (from 5 to 15 seconds) of adaptation data is available. In our simulations using the DARPA WSJ Spoke3 corpus, an average of 10.5% relative reduction in WER was achieved over MLLR adaptation when using 5 seconds data for adaptation.

Volume 2, page 1215 Session C14

Evaluation on Unsupervised Speaker Adaptation Based on Sufficient HMM Statictics of Selected Speakers

Yoshizawa S[1], Baba A[2], Matsunami K[3], Mera Y[3], Yamada M[3], Lee A[3], Shikano K[3]
[1]Matsushita Electric Industrial Co., Japan, [2]Laboratories of Image Information Science and Technology, Japan, [3]Nara Institute of Science and Technology, Japan

This paper describes an efficient method of unsupervised speaker adaptation. This method is based on (1) selecting a subset of speakers who are acoustically close to a test speaker, and (2) calculating adapted model parameters according to the previously stored sufficient statistics of the selected speakers' data. In this method, only a few unsupervised test speaker's data are necessary for the adaptation. Also, by using the sufficient HMM statistics of the selected speakers' data, a quick adaptation can be done. Compared with a pre-clustering method, the proposed method can obtain a more optimal cluster because the clustering result is determined according to test speaker's data on-line. Experimental results show that the proposed method attains better improvement than MLLR from the speaker-independent model. The proposed method is evaluated in details and discussed.

Volume 2, page 1219 Session C14

Session C15 - Poster
Wednesday - 09.00 - 10.40

Speech Recognition and Understanding: Adaptation

Chair: Christian Wellekens, Eurecom, France

A Novel Target-Driven MLLR Adaptation Algorithm with Multi-Layer Structure

Jia L, Xu B
Institute of Automation, Chinese Academy of Science, P. R. China

ABSTRACT: This paper presents a novel target-driven MLLR adaptation algorithm with multiply layer structure, which is based on the thorough analysis of MLLR using the generation of regression class trees. The new algorithm is constructed on the target-driven principal. It generates the regression class dynamically, basing on the outcome of the former MLLR transformation. The regression classes is defined in order to have the maximizing increase of the auxiliary function, which is in proportional to the likelihood of the occurrence of the adaptation data. Because of the new algorithm¡¯s special transformation structure, computation load in performing transformation is much reduced. In comparison with the conventional MLLR using the generation of regression class trees, the new algorithm give a further error reduction 10% and has only half computation time consuming.

Volume 2, page 1225 Session C15

Scaled Likelihood Linear Regression for Hidden Markov Model Adaptation

Wallhoff F, Willett D, Rigoll G
Gerhard-Mercator-University Duisburg, Germany

In the context of continuous Hidden Markov Model (HMM) based speech-recognition, linear regression approaches have become popular to adapt the acoustic models to the specific speaker's characteristics. The well known Maximum Likelihood Linear Regression (MLLR) and Maximum A Posteriori Linear Regression (MAPLR) are just two of them, which differ primarily in the training objective they are maximizing. However, besides the approaches mentioned above there exists another known training objective which is the Maximum Mutual Information (MMI). By combining this MMI-approach with the linear regression of the HMM's mean values, our research group developed a new adaptation technique that we call Scaled Likelihood Linear Regression (SLLR). In this approach, the distance of the correct model sequence against the wrong ones is discriminated framewise. Like all techniques using MMI objectives, this adaptation is computationally very expensive compared to techniques using ordinary ML based objectives. This paper therefore addresses the problem of an appropriate approximation technique to speed up this adaptation approach, by pruning the computation for tiny values in the discrimination objective. To further explore the potential of this adaptation technique and its approximation, the performance is measured on the LVCSR-system DUDeutsch developed by our research group at the Duisburg University and additionally on the 1993 WSJ adaptation tests of native and non-native speakers for the supervised case.

Volume 2, page 1229 Session C15

Fast Adaptation using Constrained Affine Transformations with Hierarchical Priors

Myrvoll T A[1], Paliwal K K[2], Svendsen T[1]
[1]NTNU, Norway, [2]Griffith University, Australia

In this paper we present an approach to transformation based model adaptation that combines a fast, closed form solution to the MAP estimation of our transforms with robust priors. The robust priors are

found using the technique of hierarchical priors, and a closed form solution is achieved by choosing diagonally constrained affine transformations and a suitable family of prior distributions for these transformations. We show that the method gives results comparable to other algorithms, but with significantly reduced computational complexity and memory demands. Experiments are conducted on the SI Recognition Outlier task from the Wall Street Journal corpus, where speaker independent models have to be adapted to handle speech from non-native speakers.

Volume 2, page 1233 Session C15

A Context Adaptation Approach for Building Context Dependent Models in LVCSR

Liu X, Yuan B, Yan Y
Intel China Research Center, P. R. China

This paper introduces a new context adaptation framework for building context dependent HMM models in LVCSR. In this new framework, all states of each center phone are clustered into groups by the decision tree algorithm. All the tied states of context dependent HMM models were then derived by adapting the parameters of the multiple-mixture context independent model via data dependent MAP (maximum a posteriori probability)method using the training vectors corresponding to the tied state. An advantage of this approach is that it can maintain a high prediction and classification power given limited training data therefore the model trained in this framework is more reliable than in conventional framework. Experimental results on Wall Street Journal corpora demonstrate that the proposed approach leads to a significant improvement in recognition performance.

Volume 2, page 1237 Session C15

Improving Genericity for Task-Independent Speech Recognition

Lefevre F, Gauvain J-L, Lamel L
LIMSI-CNRS, France

Although there have been regular improvements in speech recognition technology over the past decade, speech recognition is far from being a solved problem. Recognition systems are usually tuned to a particular task and porting the system to a new task (or language) is both time-consuming and expensive. In this paper, issues in speech recognizer portability are addressed through the development of generic core speech recognition technology. First, the genericity of wide domain models is assessed by evaluating performance on several tasks. Then, the use of transparent methods for adapting generic models to a specific taskis explored. Finally, further techniques are evaluated aiming at enhancing the genericity of the wide domain models. We show that unsupervised acoustic model adaptation and multi-source training can reduce the performance gap between task-independent and task-dependent acoustic models, and for some tasks even out-perform task-dependent acoustic models.

Volume 2, page 1241 Session C15

A Posteriori and a Priori Transformations for Speaker Adaptation in Large Vocabulary Speech Recognition Systems

Matrouf D, Bellot O, Nocera P, Linares G, Bonastre J-F
LIA, Avignon, France

The speaker-dependent HMM-based recognizers gives lower word error rates in comparison with the corresponding speaker-independent recognizers. The aim of speaker adaptation techniques is to enhance the speaker-independent acoustic models to bring their recognition accuracy as close as possible to the one obtained with speaker-dependent models. In this paper, we propose a method using test and training data for acoustic model adaptation. This method operates in two steps. The first one performs an a priori adaptation using the transcribed training data of the closest training speakers to the test speaker. This adaptation is done with MAP procedure allowing reduced variances in the acoustic models. The second one performs an a posteriori adaptation using the MLLR procedure on the test data, allowing mapping of Gaussians means to match the test speaker's acoustic space. This adaptation strategy was evaluated in a large vocabulary speech recognition task. Our method leads to a relative gain of 15% with respect to the baseline system and 10% with respect to the MLLR adaptation.

Volume 2, page 1245 Session C15

Bayesian methods for HMM speech recognition with limited training data

Purnell D W, Botha E C
University of Pretoria, South Africa

This paper presents a Bayesian approach to learning for HMMs in speech recognition. The implementation of Bayesian learning for HMMs in speech recognition is discussed, including the requirement of maintaining the original HMM constraints, choice of prior and utterance recognition. This work shows that the Bayesian learning approach can be successfully applied to complex models when the amount of training data is small.

Volume 2, page 1249 Session C15

Rapid Speaker Adaptation Using MLLR and Subspace Regression Classes

Wong K-M, Mak B
The Hong Kong University of Science and Technology, Hong Kong

In recent years, various adaptation techniques for hidden Markov modeling with mixture Gaussians have been proposed, most notably MAP estimation and MLLR transformation. When the amount of adaptation data is limited, adaptation can be done by grouping similar Gaussians together to form regression classes and then transforming the Gaussians in groups. The grouping of Gaussians is often determined at the full-space level. In this paper, we propose to group the Gaussians at a finer acoustic subspace level. The motivation is that clustering at subspaces of lower dimensions results in lower distortion. Besides, as the dimension of subspace Gaussians reduces, there are fewer parameters to estimate for the subsequent MLLR transformation matrix. This is particular attractive in fast adaptation. Speaker adaptation experiments on the Resource Management task with few seconds of speech show that the use of subspace regression classes is more effective than traditional full-space regression classes.

Volume 2, page 1253 Session C15

Speaker Adaptation of Output Probabilities and State Duration Distributions for Speech Recognition

Yoma N B, Silva J
University of Chile, Chile

This paper presents a comparison of maximum a posteriori (MAP) speaker adaptation of state duration distributions and output probabilities in HMM. Both adaptation procedures are compared and then combined in recognition experiments with clean and noisy signals. The results here shown suggest that the state duration distribution adaptation can lead to higher improvements than the adaptation of output probabilities, and the reduction in the error rate when both adaptations are combined is as high as 50% or 60% using only a few samples per word.

Volume 2, page 1257 Session C15

Cohorts Based Custom Models for Rapid Speaker and Dialect Adaptation

Wu J[1], Chang E[2]

[1]*The University of Hong Kong, Hong Kong, P. R. China*, [2]*Microsoft Research China, P. R. China*

It is well known that speaker dependent acoustic models can achieve an error rate that is up to a factor of two smaller compared to well trained speaker independent acoustic models. Thus, for improved accuracy, many modern dictation systems require the user to perform enrollment sessions to adapt the acoustic model of the system. In this paper, we present an approach that uses as few as three sentences from the test speaker to select N closest speakers (cohorts) from both the original training set and newly available training speakers to construct customized models. By using such an approach, our adaptation scheme can be updated online without re-configuring anything that has been calculated before. When applying this approach to address dialectal differences, the cohort based user specific models constructed with 3 user sentences can obtain a lower error rate even when compared to user-adapted models based on 170 user sentences.

Volume 2, page 1261 Session C15

Speaker Adaptation of Quantized Parameter HMMs

Vasilache M, Viikki O
Nokia Research Center, Finland

This paper extends the evaluation of Hidden Markov Models with quantized parameters (qHMM) presented in [5] to the case of speaker adaptive training. In speaker-independent speech recognition tasks, qHMMs were found to provide a similar performance as the original continuous density HMMs (CDHMM) with substantially reduced memory requirements. In this paper, we propose a Bayesian type of adaptation framework for qHMMs to improve the speaker-specific acoustic modeling accuracy. Experimental results indicate that the proposed qHMM adaptation scheme provides a comparable performance as obtained with the Bayesian adaptation of CDHMMs in a noise-free test environment. In the presence of noise, on the other hand, the performance improvement due to qHMM adaptation is lower than obtained in the CDHMM case. In general, the adaptation gains are on a similar scale fact that confers to qHMMs a great practical value.

Volume 2, page 1265 Session C15

Segmental Eigenvoice for Rapid Speaker Adaptation

Tsao Y, Lee S-M, Lee L-S
National Taiwan University, Taiwan, ROC

This paper presents a new approach to improve the conventional eigenvoice technique. In the conventional eigenvoice, an eigenspace is established by introducing a priori knowledge of training speakers via PCA. The adaptation data is then used to determine a group of coefficients with respect to the eigenspace and build the SD model for the testing speaker. In the proposed approach, the eigenspace in the conventional eigenvoice is segmented into N sub-eigenspaces. Each sub-eigenspace is established by those components in the training speaker SD models with similar properties to each other. With the adaptation data, N groups of coefficients corresponding to the N sub-eigenspaces can be determined to build SD model for the new testing speaker. Here, both mixture-based and feature-based segmentation of eigenspace were tested, and improved results compared to the conventional eigenvoice were obtained in both cases. Even better results were obtained when these approaches were properly combined.

Volume 2, page 1269 Session C15

Speaker adaptation in an ASR system based on nonlinear dynamical systems

Warakagoda N D, Johnsen M H
NTNU, Norway

The work presented here is centered around a speech production model called Chained Dynamical System Model (CDSM) which is motivated by the fundamental limitations of the mainstream ASR approaches. The CDSM is essentially a smoothly time varying continuous state nonlinear dynamical system, consisting of two sub dynamical systems coupled as a chain so that one system controls the parameters of the next system. The speech recognition problem is posed as inverting the CDSM, which is solved using the ideas borrowed from the theory of Embedding. The resulting architecture, which we call Inverted CDSM (ICDSM) is well suited for modeling variations of speaker and channel characteristics, by its nature. We have evaluated the ICDSM using a set of experiments involving speaker adaptation in a continuous speech recognition task on the TIMIT database. Results of these experiments confirm the feasibility and potential advantages of the approach.

Volume 2, page 1273 Session C15

Session C16 - Poster
Wednesday - 09.00 - 10.40

Dialogue Systems: Project Descriptions - I

Chair: Rolf Carlson, KTH, Stockholm, Sweden

An Interactive Directory Assistance Service for Spanish with Large-Vocabulary Recognition

Córdoba R, San-Segundo R, Montero J M, Colás J, Ferreiros J, Macías-Guarasa J, Pardo J M
Universidad Politécnica de Madrid, Spain

In the EU funded IDAS project (LE4-8315), demonstrators providing an automated interactive telephone-based directory assistance service have been developed by ten partners from Germany, Greece, Spain and Switzerland [6]. In this paper we will focus in the Spanish demonstrator. In particular, we will describe the following aspects: The general architecture of the system, paying special attention to the speech recognition module. We will present new alternatives for the estimation of continuous HMMs and the agglomerative clustering of context-dependent units. The most common problems encountered in the development of this kind of systems and their operation in a real environment. Impressions, opinions and scores from real-world users of the system. Keywords: large vocabulary recognition, telephone-based, directory assistance service, dialog.

Volume 2, page 1279 Session C16

A Multilingual-supporting Dialog System Using a Common Dialog Controller

Xu Y, Araki M, Niimi Y
Kyoto Institute of Technology, Japan

It is well known that a speech dialog system can be regarded as an integration of a speech interface which runs in the front end and a dialog controller which runs in the back end. The former is obviously language-dependent while the later could be language-independent relatively. This paper describes an approach to constructing a multilingual spoken dialog system. In this approach, we extended a dialog controller for Japanese to a language-independent one and combined it with a Chinese speech interface. Experimental result shows that the proposed approach is effective in constructing quickly a multilingual-supporting dialog system using a common dialog controller.

Volume 2, page 1283 Session C16

Graphic platform for designing and developing practical voice interaction systems

Nouza T, Nouza J
Technical University of Liberec, Czech Republic

A complete development environment for designing, building and running voice operated services has been created. It offers a system builder a graphic platform with several types of blocks, such as an ASR block, a TTS one, a switch block, a database query block, etc. Even a large dialogue scheme can be realized in very short time simply by placing blocks on the form, specifying their properties and aligning them into meaningful dialogue branches. Sequencing the blocks into branches is solved in a unique way without using any interconnecting lines, which makes the dialogue scheme easy for editing. The platform, named LOTOS, has been employed in building a large multi-domain information system with voice access via telephone.

Volume 2, page 1287 Session C16

Speech Translation for French in the NESPOLE! European Project

Besacier L, Blanchon H, Fouquet Y, Guilbaud J-P, Helme S, Mazenot S, Moraru D, Vaufreydaz D
CLIPS, France

This paper presents CLIPS laboratory activities in the context of the NESPOLE! European project, exploring future applications of automatic speech to speech translation in e-commerce and e-service sectors. The scientific and technological research issues particularly addressed in order to improve current experimental speech-to-speech translation systems, are: robustness, scalability, and cross-domain portability. The general architecture of the whole speech to speech translation demonstrator is first presented and the Interchange Format (IF) strategy for translation adopted in the project is quickly described. The French database recorded during the project and the French Human Language Technology (HLT) modules (recognition, synthesis and translation) are then fully detailed. First results obtained and future perspectives of the project are also discussed in this article.

Volume 2, page 1291 Session C16

Lessons from the Development of a Conversational Interface

Hickey M, St John Brittan P
HP Laboratories, UK

The design of an effective mixed initiative dialogue system still presents great challenges. This paper reports on the experiences gained in the design and implementation of an experimental spoken dialogue system, MIZIK, which revolves around a new domain, the music charts. It describes the processes we went through to: determine the development approach for a robust system; specify the scope of the domain; select an appropriate architecture and speech and language technology; collect training data specific to the domain and the target user population and, finally, to develop the experimental system. The paper concludes with a number of key lessons learnt during these processes, many of which are equally applicable to the design and development of any conversational speech interface.

Volume 2, page 1295 Session C16

SCANMail: Browsing and Searching Speech Data by Content

Hirschberg J[1], Bacchiani M[1], Hindle D[2], Isenhour P[3], Rosenberg A[1], Stark L[4], Stead L[1], Whittaker S[1], Zamchick G[1]
[1]AT&T Labs - Research, USA, [2]AnswerLogic, USA, [3]Virginia Tech, USA, [4]University of Delaware, USA

Increasing amounts of public, corporate, and private audio data are available for use, but limited in usefulness by the lack of tools to permit their browsing and search. In this paper, we describe SCANMail, a system that employs automatic speech recognition, information retrieval, information extraction, and human computer interaction technology to permit users to browse and search their voicemail messages by content through a graphical user interface interface. The SCANMail client also provides note-taking capabilities as well as browsing and querying features. A CallerId server also proposes caller names from existing caller acoustic models and is trained from user feedback. An Email server sends the original message plus its transcription to a mailing address specified in the user's profile.

Volume 2, page 1299 Session C16

Multi-Scale Retrieval in MEI: An English-Chinese Translingual Speech Retrieval System

Lo W-K[1], Schone P[2], Meng H[1]
[1]The Chinese University of Hong Kong, Hong Kong, [2]Department of Defense, USA

This paper presents a multi-scale retrieval approach in MEI (Mandarin-English Information), an English-Chinese cross-lingual spoken document retrieval (CL-SDR) system. It accepts an entire English news story (from newspaper text) as the input query, and automatically retrieves "relevant" Mandarin news stories (from broadcast audio). This allows the user to search for personally relevant content across the language and media barriers – a cross-lingual and cross-media retrieval task. MEI advocates a multi-scale paradigm for the retrieval task. Multi-scale refers to the use of both words and subwords (Chinese characters and syllables) for retrieval. Words offer lexical knowledge to enhance precision, and subwords can potentially alleviate some prevailing problems in CL-SDR, e.g. open vocabularies in translation and recognition, out-of-vocabulary words in audio indexing, and ambiguities in Chinese homophones and word tokenizaiton. We present techniques for word-subword fusion, which improved retrieval performance in our experiments with the Topic Detection and Tracking collection.

Volume 2, page 1303 Session C16

Compact Word Graph in Spoken Dialogue System

Chien S-C, Chang S-C
Industrial Technology Research Institute, Taiwan

In this paper, we introduce the multi-stage configuration for the interpretation of user¡¦s queries in our spoken dialogue system. In this configuration, a recovery mechanism is used to detect and recover the errors arising from speech recognition. To efficiently incorporate with this recovery mechanism, a recognition scheme that can provide a compact word graph is developed. The compact word graph is generated through a pruning method based on the N-best sentence score. Instead of setting threshold, we use the N-best sentence score to select word hypotheses of the compact word graph.

Volume 2, page 1307 Session C16

MINOS-II: A Prototype Car Navigation System with Mixed Initiative Turn Taking Dialogue

Sasajima M, Yano T, Shimomori T, Uehara T
TOSHIBA Corp, Japan

Spoken dialogue systems are classified into three types from the viewpoint of turn taking. Dialogue can be led by the system (system initiative), the user (user initiative), and their mixture (mixed initiative). In this paper, EUROPA, a framework for developing spoken dialogue systems, is introduced. EUROPA is applied to prototyping a car navigation system called MINOS-II. MINOS-II deals with a car navigation task of mixed initiative dialogue. First, the system takes the initiative to lead the user to set a route for the destination. Next, while driving along the route, the user takes the initiative and retrieves information about the route freely. MINOS-II is built on a portable PC, can process over 2 million sentence patterns, and is able to respond to a user's question within a few seconds.

Volume 2, page 1311 Session C16

Use of Topic Knowledge in Spoken Dialogue Information Retrieval System for Academic Documents

Kiriyama S, Hirose K, Minematsu N
University of Tokyo, Japan

An efficient search function based on topic estimation was integrated to our spoken dialogue system for academic document information retrieval. The following two points were mainly studied: 1) to properly categorize documents (to be retrieved) into related topics, and 2) to facilitate retrieval process using topic knowledge. For the first point, a method was developed to calculate recursively the relevance scores of retrieval words and documents for topics. Effects of the recursive process were proved through experimental results; better classification of retrieval words and documents into topics was realized. As for the second point, retrieval range was limited into topics estimated from retrieval words. It was shown through experiments of retrieval task solving that necessary number of dialogue turns (therefore, period of dialogue) could be largely reduced by the range limitation; a smooth retrieval process was proved to be realized using topic knowledge.

Volume 2, page 1315 Session C16

Domain-Independent Spoken Dialogue Platform using Key-Phrase Spotting based on Combined Language Model

Komatani K, Tanaka K, Kashima H, Kawahara T
Kyoto University, Japan

We present a portable platform for spoken dialogue systems. Conventional development of speech interfaces involves much labor cost in either describing a task grammar or collecting a task corpus. Our platform automatically generates a lexicon and a language model of key-phrases based on task description and the domain database. By spotting key-phrases using both the generated grammar and word 2-gram model trained with dialogue corpora of similar domains, we realize flexible speech understanding on a variety of utterances. Furthermore, adopting a GUI that explicitly displays acceptable utterance patterns is effective in guiding user utterances within the system's capability. We evaluate the generated system using 24 novice users. The number of unacceptable utterances are significantly reduced with the simple phrase grammar and GUI. And the phrase spotter using the combined language model improves the semantic accuracy by 15.5% compared with the conventional method decoding the whole sentence with a fixed grammar.

Volume 2, page 1319 Session C16

OASIS Natural Language Call Steering Trial

Durston P[1], Kuo H-K J[2], Farrell M[1], Afify M[2], Attwater D[1], Fosler-Lussier E[2], Allen J[1], Lee C-H[2]
[1]BTexaCT, UK, [2]Bell Labs, USA

A recent trial of natural language call steering on live UK calls to the operator is described along with its results. The characteristics of the problem are described along with the acoustic, language, semantic and dialogue modelling approaches employed. Natural language call steering is found to be viable, with recognition and semantic accuracy the current limiting factors.

Volume 2, page 1323 Session C16

First steps toward an adaptive spoken dialogue system in medical domain

Azzini I[1], Falavigna D[2], Gretter R[2], Lanzola G[1], Orlandi M[2]
[1]Universita` di Pavia, Italy, [2]ITC-Irst, Italy

Recently the spoken dialog group of ITC-irst and the Dipartimento di Informatica e Sistemistica of the University of Pavia are working together to realize an intelligent spoken dialog system with adaptation capabilities. In this framework, some telemedicine services able to handle multimodal interactions are going to be investigated and developed. In the paper dialog "adaptation" will be defined according to the medical domains in which the system is going to be used: i.e. to assist chronic patients. The present architecture of the system will be described and some ideas for its future development will be discussed. Although the system is placed in medical domains, the basic concepts can, in principle, be exported towards different applications. The work reported in the paper is part of a more wide-ranging project aimed at efficiently merging patient and domain specific knowledge, with statistical knowledge (e.g. n-grams and/or concept probabilities) derived from real user interactions.

Volume 2, page 1327 Session C16

Mokusei: A Telephone-based Japanese Conversational System in the Weather Domain

Nakano M[1], Minami Y[2], Seneff S[3], Hazen T J[3], Cyphers D S[3], Glass J[3], Polifroni J[3], Zue V[3]
[1]MIT Laboratory for Computer Science, USA / NTT, Japan, [2]NTT, Japan, [3]MIT Laboratory for Computer Science, USA

This paper describes Mokusei, an end-to-end Japanese version of our Jupiter weather information system. Mokusei delivers weather information over the phone through natural conversation with the user. For the most part, Mokusei uses the same components for recognition, understanding, and generation that Jupiter uses, and the database and the semantic frames for the weather information content are also shared. However, Mokusei motivated us to redesign our Genesis generation system, in order to improve the quality of translations of weather reports into Japanese. We also had to develop new ways to transcribe user utterances through morphological analysis. Mokusei is fully functional and has already been used for data collection with about 700 naive users. These data have been used for improvement and evaluation of Mokusei. This paper also presents the result of evaluating the current version of Mokusei.

Volume 2, page 1331 Session C16

SpeechBuilder: Facilitating Spoken Dialogue System Development

Glass J, Weinstein E
MIT Laboratory for Computer Science, USA

In this paper we report our attempts to facilitate the creation of mixed-initiative spoken dialogue systems for both novice and experienced developers of human language technology. Our efforts have resulted in the creation of a utility called SpeechBuilder, which allows developers to specify linguistic information about their domains, and rapidly create spoken dialogue interfaces to them. SpeechBuilder has been used to create domains providing access to structured information contained in a relational database, as well as to provide human language interfaces to control or transaction-based applications.

Volume 2, page 1335 Session C16

Voice-IF: A Mixed-Initiative Spoken Dialogue System for AT&T Conference Services

Rahim M, Di Fabbrizio G, Kamm C, Walker M, Pokrovsky A, Ruscitti P, Levin E, Lee S, Syrdal A, Schlosser K
AT&T, USA

This paper presents the Voice-IF system; a mixed-initiative spoken dialogue system for AT&T conference services. One objective for creating Voice-IF is to provide a vehicle for evaluating our technologies in speech synthesis, recognition, understanding, dialogue and user interfaces on a real application with relatively novice users. Another objective is to design, build and test a set of tools that allow us to rapidly prototype applications. In this paper, we describe the performance of Voice-IF during its 6-week deployment period. In particular, we report a) results of perceptual evaluations of the synthesized speech, b) system performance and user satisfaction ratings, c) PARADISE analysis of the data, and d) comparisons with other systems, including the W99 conference registration system used at the ASRU'99 workshop and the Travel Communicator system.

Volume 2, page 1339 Session C16

Session C21 - Oral
Wednesday - 11.10 - 12.30

ESE4 - SIGshow - Continued

Chair: Gerrit Bloothooft, University of Utrecht, The Netherlands

SIGdial - Special Interest Group on Discourse and Dialogue

Dybkjær L
Natural Interactive Systems Laboratory, Denmark

This paper describes the ACL and ISCA Special Interest Group on Discourse and Dialogue (SIGdial). Objective, activities, people, and plans are presented.

Volume 2, page 1345 Session C21

Integrating Speech Technology in Language Learning: An overview of the activities of InSTIL

Delcloque P
University of Abertay Dundee, UK

This presentation describes the activities of a Special Interest Group which focuses on the "Integration of Speech Technology in (Language) Learning". This SIG is a "bridge" between two essential, complementary traditions: Computer Assisted Language Learning (CALL) and Speech Science & Engineering. The history of the SIG is retraced since its foundation as CAPITAL (see later) in Edinburgh around 1994, past, present activities of the SIG are reviewed, its events, its publications, the origin and diversity of its membership, etc. Future activities are also mentioned. The success of the group has surprised its early members who thought that there were part of a very small minority group, this is a historic year for the SIG for two reasons, the first because it will write, present and publish the first "Illustrated History of Speech Technology in Language Learning" and second because it will be present for the first time at a EUROSPEECH conference.

Volume 2, page 1349 Session C21

ISCA SALTMIL SIG: Speech and Language Technology for Minority Languages

Nadeu C[1], Ó'Cróinín D[2], Petek B[3], Sarasola K[4], Williams B[5]
[1]Univ. Politècnica de Catalunya, Spain, [2]Linguistics Institute of Ireland, Ireland, [3]University of Ljubljana, Slovenia, [4]University of the Basque Country, Spain, [5]University of Edinburgh, UK

This paper presents International Speech Communication Association (ISCA) Special Interest Group (SIG, http://www.isca-speech.org/sig.html) on Speech And Language Technology for MInority Languages (SALTMIL). Overview of the group's mission, including its past and present activities are presented and discussed.

Volume 2, page 1353 Session C21

The Specificity of French Speech Processing (no proceedings paper)

Bimbot F[1], Bonastre J-F[2]
[1]IRISA, France, [2]LIA, Avignon, France

A presentation of the SIG Groupe Francophone de la Communication Parleé.

Volume 2, page 1344 Session C21

Session C22 - Oral
Wednesday - 11.10 - 12.30

Dialogue Systems: Resources

Chair: To be decided,

Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation

Shriberg E[1], Stolcke A[1], Baron D[2]
[1]SRI International / ICSI, USA, [2]ICSI / U.C. Berkeley, USA

We examine the distribution of overlapping speech in large multi-party conversations, including two different types of meetings, and two corpora of telephone conversations. Analyses are based on forced alignment and speech recognition using an identical recognizer across tasks. Three results are discussed. First, all corpora show high overall rates of overlap, with similar rates for meetings and telephone conversations. Second, speech recognition performance in nonoverlapped regions of meetings is no worse than that for single-channel telephone conversations, while recognition in overlap regions degrades considerably. Finally, interrupt locations are associated with endpoints of word-level events in a speaker's turn, including backchannels, discourse markers, and disfluencies. Results suggest that overlaps are an important inherent characteristic of conversational speech that should not be ignored; on the contrary, they should be jointly modeled with acoustic and language model information in machine processing of conversation.

Volume 2, page 1359 Session C22

Towards SMIL as a Foundation for Multimodal, Multimedia Applications

Di Fabbrizio G[1], Nils K[1], Beckham J L[2]
[1]AT&T - Labs Research, USA, [2]University of Wisconsin, USA

Rich and interactive multimedia applications, where audio, video, graphics and text are precisely synchronized under timing constraints are becoming ubiquitous. Multimodal applications further extend the concept of user interaction combining different modalities, like speech recognition, speech synthesis and gestures. However, authoring dialog-capable multimodal, multimedia services is a very difficult task. In this paper, we argue that SMIL is an ideal substrate for extending multimedia applications with multimodal facilities. SMIL as it stands is not a general notation for controlling media and input mode resources. We show that all what is needed are few natural extensions to SMIL along with the addition of a simple reactive programming language that we call ReX. Our language is designed to be maximally compatible with existing W3C recommendations through a generic event system based on DOM and an expression language based on XPATH.

Volume 2, page 1363 Session C22

Anvil - A Generic Annotation Tool for Multimodal Dialogue

Kipp M
University of the Saarland, Germany

Anvil is a tool for the annotation of audiovisual material containing multimodal dialogue. Annotation takes place on freely definable, multiple layers (tracks) by inserting time-anchored elements that hold a number of typed attribute-value pairs. Higher-level elements (suprasegmental) consist of a sequence of elements. Attributes contain symbols or cross-level links to arbitrary other elements. Anvil is highly generic (usable with different annotation schemes), platform-independent, XML-based and fitted with an intuitive graphical user interface. For project integration, Anvil offers the import of speech transcription and export of textand table data for further statistical processing.

Volume 2, page 1367 Session C22

DARPA Communicator Dialog Travel Planning Systems: The June 2000 Data Collection

Walker M
AT&T Shannon Labs, USA

This paper describes results of an experiment with 9 different DARPA Communicator systems who participated in the June 2000 Data collection. All Systems supported travel planning and utilized some form of mixed initiative interaction. However they varied in several critical dimensions: (1) They targeted different back-end databases for travel information; (2) They used different modules for ASR, NLU, TTS and dialog management. We describe the experimental design, the approach to data collection, the metrics collected, and results comparing the systems.

Volume 2, page 1371 Session C22

Session C23 - Oral
Wednesday - 11.10 - 12.30

Speaker Recognition: Features and Transforms

Chair: Frederic Bimbot, IRISA, Rennes, France

Analysis of Speaker Variability

Huang C[1], Chen T[2], Li S[1], Chang E[1], Zhou J[1]
[1]Microsoft Research China, P. R. China, [2]Tsinghua Univ., P. R. China

Analysis and modeling of speaker variability, such as gender, accent, age, speech rate, and phones realizations, are important issues in speech recognition. It is known that existing feature representations describing speaker variations can be of very high dimension. In this paper, we introduce two powerful multivariate statistical analysis methods, namely, principal component analysis (PCA) and independent component analysis (ICA), as tools for analysis of such variability and extraction of low dimensional feature representation. Our findings are the following: (1) the first two principal components correspond to the gender and accent, respectively. The result that the second component corresponding to the accent has never been reported before, to the best of our knowledge. (2) It is shown that ICA based features yield better classification performance than PCA ones. Using 2-dimensional ICA representation, we achieved about 6.1% and 13.3% error rate in gender and accent classification, respectively, for 980 speakers.

Volume 2, page 1377 Session C23

Speaker Recognition by Separating Phonetic Space and Speaker Space

Nishida M, Ariki Y
Ryukoku University, Japan

In speaker recognition, it is a problem that speech feature varies depending on sentences and time difference. This variation is mainly attributed to the variation of phonetic information and speaker information included in speech data. If these two kinds of information are separated each other, robust speaker recognition will be realized. In this study, we propose a speaker recognition method by separating the phonetic information and speaker information by a subspace method, under the assumption that a space with large within-speaker variance is a ``phonetic space'' and a space with small within-speaker variance is a ``speaker space''. We carried out comparative experiments of the proposed method with a conventional method based on GMM in an observation space as well as in a space transformed by LDA. As a result, we could construct a robust speaker model with a few model parameters using a few training data by the proposed method.

Volume 2, page 1381 Session C23

Eigen-MLLR Coefficients as New Feature Parameters for Speaker Identification

Wang N[1], Tsai W-H[2], Lee L-S[3]
[1]Philips Research East Asia-Taipei, Taiwan, [2]National Chiao Tung University, Taiwan, ROC, [3]National Taiwan University, Taiwan, ROC

Eigen-MLLR coefficients are proposed as new feature parameters for speaker-identification in this paper. By performing principle component analysis on MLLR parameters among training speakers, the eigen-MLLR coefficients (EMCs) are derived as the coefficients for the eigenvectors. The discriminating function of the new EMC features based on the Fisher criterion is found to be ten times larger than that of mel-frequency cepstral coefficient (MFCC) features, for distinguishing speakers. The speaker-identification accuracy using the EMC features are shown to be significantly better than that using MFCC features, especially when the quantity of enrollment data is limited. It is also shown that properly combining MFCC and EMC features can achieve a significant error rate reduction on the order of 50%-60% as compared to using MFCC features alone.

Volume 2, page 1385 Session C23

Speaker Verification Using Target and Background Dependent Linear Transforms and Multi-System Fusion

Navratil J, Chaudhari U V, Ramaswamy G N
IBM T.J. Watson Research Center, USA

This paper describes a GMM-based speaker verification system that uses speaker-dependent background models transformed by speaker-specific maximum likelihood linear transforms to achieve a sharper separation between the target and the nontarget acoustic region. The effect of tying, or coupling, Gaussian components between the target and the background model is studied and shown to be a relevant factor with respect to the desired operating point. A fusion of scores from multiple systems built on different acoustic features via a neural network with performance gains over linear combination is also presented. Results obtained on the 1999 speaker recognition evaluation set indicate reductions of the minimum detection cost of up to 13% and 25% for all tests and electret-only tests respectively, as compared to a baseline GMM system. The neural fusion of three systems gains further 5% cost reduction.

Volume 2, page 1389 Session C23

Session C24 - Oral
Wednesday - 11.10 - 12.30

Speech Perception: Prosody

Chair: Jacques Terken, Eindhoven Univ. of Technology

Testing the perceptual relevance of syntactic completion and melodic configuration for turn-taking in Dutch

Caspers J
Universiteit Leiden Centre for Linguistics, The Netherlands

The research presented in this paper focuses on the role of melodic configuration and syntactic completion in the turn-taking process in Dutch. Subjects were presented with fragments of task-oriented dialogue, in which syntactic completeness and four types of melodic configuration were systematically varied, asking them to indicate whether they expected the turn to change at a specific point or not. The number of expected speaker changes turns out to be very low when no possible syntactic completion point has been reached. A rising pitch accent followed by a level boundary tone (H* %) is generally interpreted as a signal that the speaker wishes to continue, while H* H%, H*L L% and H*L H% configurations at syntactic boundaries are expected to be followed by a speaker change in the majority of cases. The data support the view that syntactic and melodic completion play a major role in the projection of possible turn-transition places.

Volume 2, page 1395 Session C24

Cues for Perceived Pitch Register

Rietveld T[1], Vermillion P[2]
[1]University of Nijmegen, the Netherlands, [2]University of London, UK

The aim of this experiment was to assess empirically listeners' behaviours in characterising pitch contours with the label pitch register. The motivation for this assessment was initiated by the conflicting use of the term 'register' in speech science and intonology. The findings reported here indicate that pitch register would more appropriately be associated with position of the Low pitch targets and the mean value of the tonal contour. In addition, the importance of the F0-min and the distance between H and L targets have been found to be weaker than previously assumed.

Volume 2, page 1399 Session C24

Language-specific Effects of Pitch Range on the Perception of Universal Intonational Meaning

Chen A, Rietveld T, Gussenhoven C
University of Nijmegen, the Netherlands

Two groups of listeners, with Dutch and British English as their native language judged stimuli in Dutch and British English, respectively, on the scales CONFIDENT vs. NOT CONFIDENT and FRIENDLY vs. NOT FRIENDLY, two meanings derived from Ohala's universal Frequency Code. The stimuli, which were lexically equivalent, were varied in pitch contour and pitch range. In both languages, the perceived degree of confidence decreases and that of friendliness increases when the pitch range is raised, as predicted by the Frequency Code. However, at identical pitch ranges, British English is perceived as more confident and more friendly than Dutch. We argue that this difference in degree of the use of the Frequency Code is due to the difference in the standard pitch ranges of Dutch and British English.

Volume 2, page 1403 Session C24

Comparing Word-Level Intelligibility after Linear vs. Non-Linear Time-Compression

Janse E
Utrecht Institute of Linguistics OTS, the Netherlands

In this paper the question is addressed whether the word-level intelligibility of time-compressed speech can be improved over linear compression by using a type of non-linear compression. Two options are tested: one type of compression which takes into account the natural timing of fast speech; and one other type of compression that saves the segmental intelligibility of short unstressed syllables. This is tested at two rates of speech: fast and very fast. The results of the perception experiments (an articulation test and a speech-interference test) show that, at both rates of speech, neither of the two types of non-linear time-compression improves intelligibility over linear compression. This suggests that both the prosodic pattern and the segmental intelligibility of both syllables contribute to word recognition in fast speech.

Volume 2, page 1407 Session C24

Session C25 - Poster
Wednesday - 11.10 - 12.30

Speech Recognition and Understanding: Pronunciation and Subword Units - II

Chair: Jim Glass, MIT, Boston, USA

Dynamic Lexicon Using Phonetic Features

Lee K-T, Wellekens C J
Institut Eurécom, France

In order to better model pronunciation variations, we present in this paper a method to build a lexicon whose content changes dynamically with the input speech. To achieve this goal, we proceeded in two steps. In the first step, a static augmented lexicon is created by adding new phone transcriptions to a basic lexicon. These new variants are derived from phonetic features that are automatically extracted from some training speech. Then in the second step, phonetic features are extracted again during recognition and help to select entries in the augmented lexicon that best match the phonetic characteristics of a given speech. These selected transcriptions constitute the dynamic lexicon, which is specific to each input utterance. Experiments showed a 16.0% relative reduction in WER compared to the baseline and 16.7% compared to when a static augmented lexicon is used.

Volume 2, page 1413 Session C25

Triphone Tying Techniques combining A-Priori Rules and Data Driven Methods

Ziegenhain U, Bauer J G
Siemens, Germany

Tying of Hidden Markov Model states is an important issue for the use of triphones as modeling units in automatic speech recognition systems. This paper studies the application of a-priori rules for tying in combination with data driven methods. The baseline method features a combination of a-priori rules that reduce the theoretical number of units by an oder of magnitude and a simple back-off tying. Back-off tying is based on the frequency of units appearing in the training material. The use of the a-priori rules has practical advantages especially for the implementation of continuous phoneme recognition. This method is compared to the widely used decision tree based clustering that makes no use of a-priori rules. A third method is proposed that combines a-priori rules with decision tree based clustering. Experiments on telephone data show that the combined method outperforms both other methods preserving the advantages of applying a-priori rules.

Volume 2, page 1417 Session C25

Pronunciation modeling and lexical adaptation in mid-size vocabulary ASR

ten Bosch L, Cremelie N
Lernout & Hauspie Speech Products N.V., Belgium

A computational-phonological method is presented to automatically adapt the phone transcriptions in a lexicon to improve ASR performance in a number of mid-size recognition tasks. The lexical adaptation approach is based on supervised phoneme loops using cd-HMM segments to find alternatives for the transcriptions, and can be considered as a counterpart of the K-means algorithm but on symbolic level. The word error rate in a limited task (digit string recognition) with dialect speakers is shown to drop by 20-25 percent relative, starting from non-dialect digit transcriptions. Since the method is computationally involving, it is only feasible for relatively small tasks.

Volume 2, page 1421 Session C25

Estimating Pronunciation Variations from Acoustic Likelihood Score for HMM Reconstruction

Liu Y, Fung P
Hong Kong University of Science and Technology, Hong Kong

It is widely acknowledged that pronunciation modeling is an efficient way to improve recognition performance in spontaneous speech. In pronunciation modeling, almost all methods of generating variation probability are based on relative frequency counting from DP alignment. In this paper, we investigate the local model mismatching caused by pronunciation variations and propose to estimate variation probability from acoustic likelihood score. According to estimated probability, we present a method of reconstructing pre-trained HMM models to include alternate pronunciations by sharing optimal mixture components instead of distributions. Experimental results show that using reconstructed HMM set reduces syllable error rate by 2.03% absolutely compared to the baseline system, also the accuracy improvement gained from proposed method is almost double with respect to that from previous DP alignment.

Volume 2, page 1425 Session C25

Breadth-First Search for Finding the Optimal Phonetic Transcription from Multiple Utterances

Bisani M, Ney H
Lehrstuhl fuer Informatik VI, RWTH Aachen, Germany

Extending the vocabulary of a large vocabulary speech recognition system usually requires phonetic transcriptions for all words to be known. With automatic phonetic baseform determination acoustic samples of the words in question can substitute for the required expert knowledge. In this paper we follow a probabilitistic approach to this problem and present a novel breadth-first search algorithm which takes full advantage of multiple samples. An extension to the algorithm to genereate phone graphs as well as an EM based iteration scheme for estimating stochastic pronunciation models is presented. In preliminary experiments phoneme error rates below 5% with respect to the standard pronunciation are achieved without language or word specific prior knowledge.

Volume 2, page 1429 Session C25

Improved Data-Driven Generation of Pronunciation Dictionaries Using an Adapted Word List

Wolff M, Eichner M, Hoffmann R
Dresden University of Technology, Germany

Data-driven approaches to learning pronunciation variants for phonetic dictionaries have to deal with the problem of acquiring a sufficient amount of training data. The reason is not the size of the databases, but the unfavorable distribution of word frequencies in natural speech, which is known as Zipf's law. In this paper we suggest a method which reorganizes a phonetic dictionary according to a given speech database in order to maximize the number of word models for which pronunciation variants can be learned with this corpus. Reorganization takes place automatically by analyzing the orthographic and phonetic transcriptions of the corpus. The method produces an alternative word list consisting of units ranging from partial words to multi-words. The efficiency and the limits of the approach are discussed on the basis of experiments carried out on the German VERBMOBIL corpus.

Volume 2, page 1433 Session C25

Segment-Based Recognition on the PhoneBook Task: Initial Results and Observations on Duration Modeling

Livescu K, Glass J
MIT Laboratory for Computer Science, USA

This paper describes preliminary recognition experiments on PhoneBook, a corpus of isolated, telephone-bandwidth, read words from a large (almost 8,000-word) vocabulary. We have chosen this corpus as a testbed for experiments on the language model-independent parts of a segment-based recognizer. We present results showing that a segment-based recognizer performs well on this task, and that a simple Gaussian mixture phone duration model significantly reduces the error rate. We compare context-independent, stress-dependent, and word position-dependent duration models and obtain relative error rate reductions of up to 12% on the test set. Finally, we make some observations regarding the effects of stress and word position in this isolated-word task and discuss our plans for further research using PhoneBook.

Volume 2, page 1437 Session C25

Multilingual Text-To-Phoneme Mapping

Riis S K, Pedersen M W, Jensen K J
Nokia Mobile Phones, Denmark

This paper introduces a novel approach for generating multilingual text-to-phoneme mappings for use in multilingual speech recognition systems. The multilingual mappings are based on the weighted outputs from a neural network text-to-phoneme model, trained on data mixed from several languages. The multilingual mappings used together with a branched grammar decoding scheme is able to capture both inter- and intra-language pronunciation variations which is ideal for multilingual speaker independent speech recognition systems. A significant improvement in overall system performance was obtained for a multilingual speaker independent name dialing task when applying multilingual instead of language dependent text-to-phoneme mapping.

Volume 2, page 1441 Session C25

Pronunciation Variation Analysis with respect to Various Linguistic Levels and Contextual Conditions for Mandarin Chinese

Tsai M-Y[1], Chou F-C[2], Lee L-S[3]
[1]National Taiwan University / Applied Speech Technology, Taiwan, [2]Applied Speech Technology, Taiwan, [3]National Taiwan University, Taiwan

Chinese language has quite different characteristic structures from those of English. There are at least word, character, syllable, Initial-Final levels in Chinese, each carrying different levels of information with complicated correlations among them. In this paper, we investigate the dependency of pronunciation variation in conversational Mandarin speech on these different levels under various contextual conditions considering the structural features of the language. The influence of speaking rate and word frequency on such pronunciation variation is also analyzed. Different pruning methods, for including pronunciation variation in speech recognition were also evaluated, and the experimental results showed that improved accuracy is obtainable if the characteristics of the pronunciation variation found in the analysis can be properly taken into account. All discussions here are based on tests with the LDC Mandarin Call Home corpus.

Volume 2, page 1445 Session C25

Hypothesis-driven Accent Discrimination

Mayfield Tomokiyo L
Carnegie Mellon University, USA

Native and non-native use of language differs, depending on the proficiency of the speaker, in clear and quantifiable ways. It has been shown that customizing the acoustic and language models of a natural language understanding system can significantly improve handling of non-native input; in order to make such a switch, however, the nativeness status of the user must be known. In this paper, we show how the recognition hypothesis can be used to predict with very high accuracy whether the speaker is native. Effectiveness of both word-based and phone-based classification are evaluated, and a discussion of the primary discriminative features is presented. In an LVCSR system in which users are both native and non-native, we have achieved a 15.6% relative decrease in word error rate by integrating this classification method with speech recognition.

Volume 2, page 1449 Session C25

An Approach to Automatic Phonetic Baseform Generation Based on Bayesian Networks

Ma C, Randolph M
Motorola INC, USA

To improve the performance and the usability of the speech recognition devices, It is necessary for most applications to allow users to enter new words or personalize words to the system vocabulary. Voice-tagging technique is a simple example that use speaker dependent spoken sample to generate baseform transcriptions of the spoken words. More sophisticated techniques can use both spoken samples and texts of the new words to generate baseform transcriptions. In this paper, we propose a new approach to the problem. We use Bayesian networks to model the letter-to-sound rule probabilities. Compared to the common decision tree based method, This new approach shows a definite advantage.

Volume 2, page 1453 Session C25

Towards Discriminative Lexicon Optimization

Schramm H, Beyerlein P
Philips Research Laboratories Aachen, Germany

A lot of work has been done in deriving the pronunciation dictionary automatically from training data. These attempts focussed mainly on maximum likelihood or similar techniques. Due to the complexity and variability of the pronunciation process it is difficult to find an adequate pronunciation model. The model will deviate from the truth. Hence, the application of maximum likelihood techniques is likely to be suboptimal. For this reason we present an approach, where the pronunciation model is learned discriminatively from data. The corresponding theory utilizes (1) probabilistic weighting of pronunciation variants of words and (2) discriminative model combination (DMC) based on Viterbi-approximations. We will show that the derived theory adjusts the weighting of pronunciation variants with respect to the word error rate, to the frequency of occurence of the specific pronunciation in the training data, and to the likelihood of the acoustic observation sequence given the pronunciation.

Volume 2, page 1457 Session C25

Model Complexity Optimization for Nonnative English Speakers

He X, Zhao Y
University of Missouri, USA

In this paper, a study is made on selecting existing acoustic models that are trained from native English speech for improving recognition of nonnative English talkers' speech. The problem is addressed from the perspective that foreign accents prevent detailed tri-phone models that are commonly used in high-performance speech recognition systems to match well with these talkers' speech, and therefore an appropriate level of context-dependent acoustic modeling is needed for foreign accent speakers. In this work, model complexity selection is accomplished by empirically choosing a set of model tying thresholds and by using the principle of MDL. An experiment was performed on the Wall Street Journal task on three nonnative English talkers with Chinese accent (276 sentences). Compared to the result obtained from using the models optimized to native English speakers, the best model tying threshold and MDL yielded similar and significant reduction to recognition word errors by 23%.

Volume 2, page 1461 Session C25

Pronunciation Modeling in Hungarian Number Recognition

Fegyó T, Mihajlik P, Tatai P, Gordos G
Budapest University of Technology and Economics, Hungary

In Hungarian, as more or less in many other languages, a large percent of words and phrases can be pronounced in several, different, but correct ways. Introducing pronunciation alternatives for individual vocabulary elements may improve the efficiency of the recognition. But in connected word recognition tasks the modeling of inter-word phonetic changes has a greater significance. In this paper we introduce a rule-based method for the automatic generation of pronunciation alternatives used first for isolated words and later the method is extended to handle cross-word phonological changes in recognition networks, applying a special approach applicable for the Hungarian language. To evaluate the method it is tested in connected number recognition tests.

Volume 2, page 1465 Session C25

Session C26 - Poster
Wednesday - 11.10 - 12.30

Speech Production: Miscellaneous

Chair: To be decided,

AMSTIVOC (AMsterdam System for Transcription of Infant VOCalizations) Applied to Utterances of Deaf and Normally Hearing Infants

Koopmans-van Beinum F J, Clement C J, Van den Dikkenberg-Pot I
University of Amsterdam, the Netherlands

The need to transcribe infant sound productions from birth onwards by using universally applicable coding tools has been basic to the development of our AMSTIVOC classification system. In this system early infant vocalizations are described by means of a sensori-motor approach based on the source-filter model for speech production. We applied the AMSTIVOC classification system, among other things, to early vocalizations of 6 deaf and 6 hearing infants in order to answer the question whether and where the lack of auditory perception can be traced in the early sound productions of deaf infants. By using this classification system it can be demonstrated that auditory feedback is needed to coordinate the movements of the phonatory and the articulatory system. This coordination capacity is likely to be a prerequisite for the development of normal speech production.

Volume 2, page 1471 Session C26

Using Linguopalatal Contact Patterns to Tune a 3D Tongue Model

Engwall O
KTH, Sweden

The six articulatory parameters of a three-dimensional tongue model were adjusted to replicate linguopalatal contact patterns measured with Electropalatography (EPG). The tongue model is based on artificially sustained articulations measured with MRI and the EPG data provides one possibility to tune the parameters to dynamic speech. A 3D model was generated of the palate and the electrode distribution, allowing the synthetic contact patterns to be calculated. The tongue parameters were then adjusted to minimise the deviation from the natural contact patterns. Substantial reduction of the false and missing electrode contacts was made in the tuning and the synthetic linguopalatal contact pattern is shown to replicate the total characteristics of the natural patterns rather well. The remaining error is often due to lateral asymmetry or central-to-edge contact variations.

Volume 2, page 1475 Session C26

Electromagnetic articulograph (EMA) based on a non-parametric representation of the magnetic field

Kaburagi T[1], Honda M[2]
[1]Kyushu Institute of Design, Japan, [2]NTT Communication Science Laboratories, Japan

Electromagnetic articulograph (EMA) systems are useful to study the motor control of speech articulators and also to construct models of the speech production process. In the EMA system, the position of the receiver coil is predicted on the basis of a field function representing a spatial pattern of the magnetic field in relation to the relative position between the transmitter and receiver coils. This paper presents a new method of representing the magnetic field by using a multivariate spline function to overcome the problem of the field pattern having local fluctuations caused by interference between the transmitting signals. A procedure for determining the receiver position is also presented, and the piecewise property of the basis functions enables the spline function to flexibly approximate the field pattern and to attain a high measurement

accuracy: the mean error in estimating the receiver position was less than 0.1 mm for a 14x14-cm measurement area.

Volume 2, page 1479 Session C26

European Portuguese Nasal Vowels: An EMMA Study

Teixeira A J D S, Vaz F
Universidade de Aveiro, Portugal

In this paper new EMMA data regarding European Portuguese nasals is presented. Some details about corpus constitution, recording and annotation is given. First results from analysis are presented. Quantitative analysis of velum movement was done for nasal vowels between stops. For the other contexts representative examples are presented and qualitatively analysed. In all contexts nasal vowels are produced with an initial phase having an high velum position. This result supports our previous work conclusions, of nasal vowels viewed as dynamic sounds were beginning must have dominant lips radiation. Obtained knowledge has application in articulatory synthesis, our motivation for this study.

Volume 2, page 1483 Session C26

The role of the palate in tongue kinematics: an experimental assessment in VC sequences from EPG and EMMA data

Fuchs S[1], Perrier P[2], Mooshammer C[1]
[1]ZAS - Centre for General Linguistics, Germany, [2]INPG & Université Stendhal, France

The effect of palatal contact on tongue tip kinematics was investigated using simultaneous EMMA and EPG recordings. The material consisted of VC sequences, where C is a voiced or voiceless alveolar stop. The kinematic characteristics were studied by analyzing parameters of the velocity profile and the deceleration peaks of the closing gesture. No evidence could be found for a potential influence of lateral contacts. Central contacts, associated with the beginning of the consonantal closure, are strongly correlated in time with the velocity drop. It supports the hypothesis that for achieving a consonantal closure tongue tip kinematics is not controlled by a specific target on the palate, and that its deceleration phase is mostly influenced by the collision with the palate.

Volume 2, page 1487 Session C26

Modelling Care of Articulation with HMMs is Dangerous

Aylett M P
University of Edinburgh, UK

Changes in care of articulation (COA) affect both the spectral and durational characteristics of speech. This can have severe repercussions on both the success of speech recognition, and the quality of speech synthesis. Although auto-segmentation has proven useful for measuring the durational effects of COA, an automatic spectral measurement has proven more problematic. In this paper, we will explore the use of the acoustic log likelihoods generated by HMM autosegmentation as a measure of these changes in comparison with two phonetically motivated modeling systems based on vocalic F1/F2 values. When duration variation is controlled, the HMM output does not correlate with the human perception of vowel goodness, whereas, the phonetically motivated models do.

Volume 2, page 1491 Session C26

Spectral Tilt as a Perturbation-free Measurement of Noise Levels in Voice Signals

Murphy P
University of Limerick, Ireland

Acoustic analysis of voice quality proves useful in the objective assessment of voice disorders and for motivating new components for use in improving voice synthesis. A commonly used quantitative spectral index is the harmonics-to-noise ratio (HNR), which gives gross information regarding speech signal periodicity. However, as the measure is sensitive to all forms of waveform aperiodicities (not simply the additive random noise component of turbulent origin), it lacks specificity. Furthermore, the HNR of the radiated speech waveform has a fundamental frequency (f0) -dependence, increasing with fundamental frequency (for equal noise levels of the glottal source). Two spectral tilt measurements are applied to synthetically generated, aperiodic voice signals to investigate their sensitivity to the various forms of aperiodicity. The tilt measures are found to provide perturbation- (jitter and shimmer) free measures of noise levels in speech signals. However, for radiated speech waveforms the tilt measurements are strongly f0-dependent.

Volume 2, page 1495 Session C26

Estimation of the modulation frequency and modulation depth of the fundamental frequency owing to vocal micro-tremor of the voice source signal

Schoentgen J
Université Libre de Bruxelles, Belgium

The aim of the article is to present a method for estimating the modulation frequency and modulation level owing to micro-tremor of the vocal fundamental frequency. Vocal micro-tremor designates the modulation of the fundamental frequency of the voice source signal owing to physiological tremor of normal speakers. The analysis is based on the spectral density function of the time series of the glottal cycle lengths. In the spectrum the modulation owing to micro-tremor gives rise to prominent spectral peaks which are positioned at the modulation frequencies. We discuss the results obtained for 38 normal male and female speakers and compare the modulation levels and frequencies to those obtained by others by means of a demodulation of the speech signal.

Volume 2, page 1499 Session C26

The perceptual relevance of glottal-pulse parameter variations

van Dinther R, Veldhuis R, Kohlrausch A
IPO-Center for User-System Interaction, The Netherlands

The perceptual relevance of changes to glottal-pulse parameters is studied. First, it is demonstrated that a distance measure based on excitation patterns can predict audibility discrimination thresholds for small changes to the R parameters of the Liljencrants-Fant (LF) model. Next, by using this measure the perceptual relevance of the LF parameters is quantified. Results are presented for a number of sets of glottal-pulse parameters that were taken from literature, representing distinct voice qualities.

Volume 2, page 1503 Session C26

Speaker Normalization Based on Test to Reference Speaker Mapping

Ogner M, Kacic Z
University of Maribor, Slovenia

The paper presents the speaker normalization technique we implemented in a teach ing and training system for hearing handicapped children with the goal to reduce inter-speaker variability in time-frequency speech representation. In an effort to reduce variance caused by variation in vocal tract shape among speakers, a f ormant based nonlinear frequency warping approach to vocal tract normalization i s investigated. The proposed method can be efficiently realized in an Analysis by Synthesis framework. After the speech decomposition into the vocal tract envelope

and exc itation model, the vocal tract envelope is warped by the estimated frequency war ping function, while the excitation characteristics are mapped to the reference speaker excitation. The results have shown significant spectral distance decrease for correctly pronounced words between test and the reference speaker after the normalization has been applied, while for poor pronunciation by the test speaker the spectral distance remains relatively high.

Volume 2, page 1507 Session C26

A face-to-muscle inversion of a biomechanical face model for audiovisual and motor control research

Pitermann M[1], Munhall K[2]
[1]INRIA Lorraine, France, [2]Queen's University, Canada

Muscle-based models of the human face produce high quality animation but estimating modeled muscle activities has not been satisfying solved yet. In this paper we present a dynamic inversion of a muscle-based model that permits the animation to be created from kinematic recordings of facial movements. Using a nonlinear optimizer (Powell's algorithm) the inversion produces a muscle activity set for 16 muscle groups in the lower face that minimize the root mean square error between kinematic data recorded with OPTOTRAK and the corresponding nodes of the modeled facial mesh. This inverted muscle activity is then used to animate the facial model. The results of a first experiment showed that the inversion-synthesis method can accurately reproduce a synthetic facial animation, even for a partial sampling of the face. The results of a second experiment showed that the method is as successful for OPTOTRAK recording of a talker uttering a sentence.

Volume 2, page 1511 Session C26

A Model of Vowel Production under Positive Pressure Breathing

South A J
20/20 Speech Ltd, UK

Future combat aircraft using speech recognition in the cockpit interface will also use positive pressure breathing (PPB) to allow operation at high G levels. This paper describes work which extends the n-tube model of vowel production to include intra-oral pressure. The aim is to improve speech recogniser performance under these conditions. An 8-tube DRM model was used, with the assumption of uniform compliance in all regions of the vocal tract. A side branch was added to simulate the oesophagus. The model shows that as the pressure is increased, the vowel space in the F1/F2 plane shrinks towards the region of F1 = 400 Hz, F2 = 1200 Hz. Measurements made on real speech show a similar trend, but the reduction in the range of F2 is less than that predicted by the model, probably as a result of variation of compliance in different areas of the vocal tract.

Volume 2, page 1515 Session C26

Helium speech normalisation by codebook mapping

Podhorski A, Czepulonis M
Technical University of Szczecin, Poland

In this paper we present a non-parametric approach to solving the helium speech problem. Properties of helium speech are replaced by those pertaining to normal speech by means of codebook mapping of spectral envelopes. This method eliminates the drawbacks inherent in the previous procedures of helium speech unscrambling as it requires neither model of helium speech production nor estimation of formant parameters. The only assumption is the general source-filter model required for linear prediction analysis. In the traditional approach spectral transformations were computed based on the assumed helium speech production model. And in the non-model approach it was assumed that helium speech distortion is speaker dependent, so all spectral transformations were calculated from formant parameters and F0 extracted directly from speech signals. In all previous methods the resulting speech was still retaining a nasal quality due to inaccurate modelling and speech processing schemes that were unable to guarantee independent manipulation of formant parameters. On the contrary our system results in speech that is completely free of the hyperbaric helium quality however its technical quality is still unsatisfactory as the mapping introduces noise into the corrected speech.

Volume 2, page 1519 Session C26

Session D11 - Oral
Thursday - 08.50 - 10.40

ESE5 - Existing and Future Corpora: Next Generation Speech Resources

Chair: Christoph Draxler, LMU, München, Germany

Session Introduction (no proceedings paper)

Draxler C
University of Munich, Germany

N/A

Volume 3, page 1524 Session D11

Building a Corpus of Natural Speech - and tools for the processing of Expressive Speech

Campbell N
ATR JST/CREST, Japan

This paper details progress during the first year of the JST/CREST ESP Project, on the creation of natural-speech databases for the analysis and synthesis of ``Expressive Speech", and the development of software tools for parameter-extraction and speech database labeling. The research is still in its initial stages, but we now have a clearer understanding of the types of speech data that will be necessary, and of the software and speech processing tools that are available for analysis and treatment of the data. The testing of applications and prototyping in real-world situations is planned as future work, and our current task is the collection of a 1000-hour corpus of natural conversational speech upon which future research will be based.

Volume 3, page 1525 Session D11

Aspects of Modern Multi-modal/Multi-media Corpora Exploitation Environments

Broeder D, Brugman H, Wittenburg P
Max-Planck Institute for Psycholinguistics, the Netherlands

This paper wants to discuss several aspects of multimodal/multimedia language resources such as the use of metadata descriptions for easy location purposes, their collaborative annotation and exploitation via Internet, the generation of synchronized media and text streams in distributed environments, and general annotation formats. These aspects that although they may be discussed independently have to fit together seamlessly to offer users an adequate exploitation environment that is up to the huge amount of data that is available in modern multi-media corpora and is able to exploit fully the current technology advancements.

Volume 3, page 1529 Session D11

Emerging Requirements for Multi-Modal Annotation and Analysis Tools

Bigbee T, Loehr D, Harper L
The MITRE Corporation, USA

We review existing capabilities of multi-modal annotation and analysis tools by presenting a survey of seven representative tools, and providing a sample annotation using one system. We discuss emerging requirements including handling electronic ink, eye-gaze tracking, and other time-based considerations. We briefly review aspects of empirically evaluating tool effectiveness and suggest that multimodal interfaces in future analytical tools may be desirable. We conclude by providing a tentative list of desired features for next-generation tools.

Volume 3, page 1533 Session D11

Three-Dimensional Modelling of Speech Corpora: Added Value through Visualisation

Altosaar T[1], Karjalainen M[1], Vainio M[2]
[1]Helsinki University of Technology, Finland, [2]University of Helsinki, Finland

Collections of annotated spoken language have formed an important basis for the development of speech technology. Their existence has promoted speech analysis research as well as enabled robust synthesis and recognition methods to be developed. However, many complex relationships remain unspecified within a corpus due to a lack of meta-data that describes the raw information in sufficient detail as well as the inter-relationships between signals, recording conditions, talkers, etc. A deficit of standards and formats, needed to express complex relationships, has also hindered the potential use and value of available corpora. This paper presents a novel three-dimensional model for exploring temporal as well as atemporal information existing in speech corpora. Examined are the potential benefits that are gained through corpus visualisation during the phases of creation, editing, verification, use, and exploration. The paper suggests that by providing a three-dimensional model of speech data, more of the inherent and potential value of a corpus can be utilised.

Volume 3, page 1537 Session D11

The Technical Processing in SmartKom Data Collection: a Case Study

Türk U
Ludwig-Maximilians-University, Germany

This paper dicusses the specific technical features and processing steps of the multimodal data collection in the SmartKom project. It gives an overview of the goals of the project and the requirements to the multimodal corpus. The processing steps from the data recording to the final distribution of the data are detailed. We focus on the problem of recording temporal synchronous data from different sources and present our manual synchronization process based on standard software and hardware. In addition, we describe shortly the logistic system for organizing the working teams and managing the processing of the data.

Volume 3, page 1541 Session D11

Session D12 - Oral
Thursday - 09.00 - 10.40

Dialogue Systems: Project Descriptions - II

Chair: Peter Heeman, OGI, Oregon, USA

SmartKom: Multimodal Communication with a Life-Like Character

Wahlster W, Reithinger N, Blocher A
DFKI, Germany

SmartKom is a multimodal dialog system that combines speech, gesture, and mimics input and output. Spontaneous speech understanding is combined with the video-based recognition of natural gestures. One of the major scientific goals of SmartKom is to design new computational methods for the seamless integration and mutual disambiguation of multimodal input and output on a semantic and pragmatic level. SmartKom is based on the situated delegation-oriented dialog paradigm, in which the user delegates a task to a virtual communication assistant, visualized as a life-like character on a graphical display. We describe the SmartKom architecture, the use of an XML-based mark-up language for multimodal content, and some of the distinguishing features of the first fully operational SmartKom demonstrator.

Volume 3, page 1547 Session D12

ISIS: A Learning System with Combined Interaction and Delegation Dialogs

Meng H[1], Chan S F[1], Wong Y F[1], Chan C C[1], Wong Y W[1], Fung T Y[1], Tsui W C[1], Chen K[2], Wang L[2], Wu T Y[2], Li X[2], Lee T[1], Choi W N[1], Ching P C[1], Chi H S[2]
[1]The Chinese University of Hong Kong, Hong Kong, [2]Peking University, P. R. China

This paper presents a progress update of our ISIS trilingual spoken dialog system. ISIS is a conversational system for the stocks domain, and supports interactions in the languages of our region - English and two dialects of Chinese (Mandarin and Cantonese). ISIS provides a system test-bed for our initial explorations with the CORBA architecture, and delegation to KQML agents. CORBA offers the advantages of interoperability, scalability and location transparency in client/server systems development. Users can delegate tasks to software agents to help monitor information (e.g. a drop in the price of a pre-specified stock), and generate user alert messages. Our current work presents new research directions in the context of ISIS: (i) automatic incorporation of newly listed stocks into our system's knowledge base; (ii) switching between on-line interaction and off-line delegation in a single dialog thread. We will also report on enhancements in the system's architecture and features.

Volume 3, page 1551 Session D12

Robust Language Understanding in Mipad

Wang Y-Y
Microsoft Research, USA

MiPad is an application prototype for the study of conversational, multi-modal interface in Microsoft Research. It has a Tap and Talk interface that allows users to effectively interact with a PDA device. The major Spoken Language Understanding (SLU) engine component behind MiPad is a robust chart parser. This paper discusses some novel features of the parser that enable it to take full advantage of the Tap and Talk interface and better support semantic based analysis. It also describes some implementation issues so that these new features can be accommodated without slowing down the parser. The new implementation speeds up the parser by a factor of three, making it more suitable for a SLU server.

Volume 3, page 1555 Session D12

The WITAS Multi-Modal Dialogue System I

Lemon O, Bracy A, Alexander G, Peters S
Stanford University, USA

We present the first demonstration version of the WITAS dialogue system for multi-modal conversations with autonomous mobile robots, and motivate several innovations currently in development for version II. The human-robot interaction setting is argued to present new challenges for dialogue system engineers, in comparison to previous work in dialogue systems under the travel-planning paradigm, in that dialogues must be asynchronous, mixed-initiative, open-ended, and involve a dynamic environment. We approached these general problems in a dialogue interface to the WITAS robot helicopter, or UAV (`Unmanned Aerial Vehicle'). We present this system and the modelling ideas behind it, and then motivate changes being made for version II of the system, involving more richly structured dialogue states and the use of automated reasoning systems over task, ability, and world-state models. We argue that these sorts of enhancement are vital to the future development of conversational systems.

Volume 3, page 1559 Session D12

Universalizing Speech: Notes from the USI Project

Shriver S, Rosenfeld R, Zhu X, Toth A, Rudnicky A, Flueckiger M
Carnegie Mellon University, USA

This paper discusses progress in designing a standardized interface for speech interaction with simple machines – the Universal Speech Interface (USI) project. We discuss the motivation for such a design and issues that must be addressed by such an interface. We present our current proposals for handling these issues, and comment on the usability of these approaches based on user interactions with the system. Finally, we discuss future work and plans for the USI project.

Volume 3, page 1563 Session D12

Session D13 - Oral
Thursday - 09.00 - 10.40

Signal Analysis: Speech Processing in Car Environments

Chair: Børge Lindberg, CPK, Denmark

Use of Real and Contaminated Speech for Training of a Hands-Free In-Car Speech Recognizer

Matassoni M, Omologo M, Svaizer P
ITC-Irst, Italy

A database of in-car speech for the Italian language was collected under the European projects SpeechDatCar and VODIS II. It consists of 600 sessions recorded under various noise and driving conditions and includes close-talk signals and far microphone signals for hands-free interaction. This paper describes some recognition experiments on two tasks conceived on a portion of this database: connected digit sequences and isolated command words. Recognition rate achieved by means of HMMs trained on real in-car speech is compared with that accomplished by a speech contamination approach, which aims at simulating in-car data starting from a clean speech corpus. Recognition performance is also analyzed as a function of the different noise conditions and of the consequent SNR at the far microphones. Finally, the effect of HMM adaptation is investigated in order to tune the recognizer on the conditions of the various sessions.

Volume 3, page 1569 Session D13

Combined Front-End Signal Processing for In-Vehicle Speech Systems

Plucienkowski J P, Hansen J H L, Angkititrakul P
Univ. of Colorado Boulder, USA

In this paper, we investigate the integration of two processing methods to improve speech quality for in-vehicle speech systems: multi-sensor beamforming and constrained iterative (Auto-LSP) speech enhancement. The intent is to establish an intelligent microphone array processing scheme in high noise environments by considering the effectiveness of a multi-sensor beamformer method and the Auto-LSP single channel speech enhancement method. The goal therefore is to design a system where the strengths of one method help compensate any potential weaknesses of the other. The noise cancellation method is an acoustic beamformer designed and constructed using a linear microphone array. The speech enhancement method is the constrained iterative Auto-LSP approach, previously considered for single channel enhancement. After establishing the combined processing scheme, evaluations are performed using speech and acoustic noise data collected in vehicles. Noise suppression levels by the beamformer is established for different road noise conditions. Quality improvement from the enhancement scheme is assessed using objective speech quality measures over a test speech corpus using TIMIT data. The results show that while beamforming alone can suppress background noise levels, the combination of beamforming and constrained enhancement can provide as much as a 63% improvement in objective quality, suggesting a potential single comprehensive solution for in-vehicle speech systems.

Volume 3, page 1573 Session D13

Robust Automatic Speech Recognition in Low-SNR Car Environments by the Application of a Connectionist Subspace-Based Approach to the Mel-based Cepstral Coefficients

Selouani S-A, Tolba H, O'Shaughnessy D
INRS-Telecommunications, Canada

In this paper, the problem of robust continuous-speech recognition (CSR) in the presence of highly interfering car noise has been considered. Our approach is based on the noise reduction of the parameters that we use for recognition, that is, the Mel-based cepstral coefficients. This is achieved by the use of a Multilayer Perceptron (MLP) network for noise reduction in the cepstral domain in order to get less-variant parameters. Then, the obtained enhanced features are {it refined} via the Karhunen-Loève Transform (KLT) implemented using the Principal Component Analysis (PCA). Experiments show that the use of the enhanced parameters using such an approach increases the recognition rate of the CSR process in highly interfering car noise environments. Results show that the proposed hybrid technique when included in the front-end of an HTK-based CSR system, outperforms that of the conventional recognition process based on either a KLT- or an MLP-based preprocessing recognition in severe interfering car noise environments for a wide range of SNRs varying from 16 dB to -4 dB using a noisy version of the TIMIT database.

Volume 3, page 1577 Session D13

Recognition of Spelled City Names in Automotive Environments

Korthauer A
Robert Bosch, Germany

This contribution presents the development and evaluation of a spelled letter recognizer for automotive environments. Specifically, the spoken language dialog for the navigation system requires reliable recognition of thousands of city names. In this context the recognition of spelling sequences is needed as fall-back strategy and for the disambiguation of similar sounding names. For that purpose we have developed a speaker-independent spelled letter recognizer on the basis of hidden Markov models using the HTK toolkit. Speech data which have been collected in real-world driving situations are used for the training of the hidden Markov models. Several feature extraction schemes were investigated and compared with regard to the recognition performance of the system. The best results for both arbitrary spelling sequences and constrained city name recognition are achieved by a system with two-channel LDA and integrated noise reduction.

Volume 3, page 1581 Session D13

Acoustic Echo Control and Noise Reduction for Cabin Car Communication

Lleida E, Masgrau E, Ortega A
University of Zaragoza, Spain

A Cabin Car Communication System (CCCS) has the goal of improving the communication among passengers inside the car. Wind, road and engine noise, the distance between passengers and other factors make difficult the communication inside vehicles. The driver must often look away from the road and passengers move out of normal seating positions. The CCCS makes use of a set of microphones to pick up the speech and the car-audio loudspeakers to reinforce the sound level. This scenario presents a great challenger for acoustic echo control and noise reduction. Acoustic echo control must prevent the overall system from howling and becoming unstable with the additional problem that the system must always work with double talk. The noise reduction must clean the microphone signal to avoid the reinforce of the noise inside the car. In this paper, we describe a combined acoustic echo control and noise reduction algorithm suitable for cabin car communication systems. A real-time system has been developed working together with the European Technological Center of Lear Corporation.

Volume 3, page 1585 Session D13

Session D14 - Oral
Thursday - 09.00 - 10.40

Speech Recognition and Understanding: Finite State Transducers for ASR

Chair: Renato De Mori, Univ. of Avignon, France

FST-Based Recognition Techniques for Multi-Lingual and Multi-Domain Spontaneous Speech

Hazen T J, Hetherington I L, Park A
MIT Laboratory for Computer Science, USA

In this paper we present techniques for building multi-domain and multi-lingual recognizers within a finite-state transducer (FST) framework. The flexibility of the FST approach is also demonstrated on the task of incorporating networks modeling different types of non-speech events into an existing word lattice network. The ability to create robust multi-domain and/or multi-lingual recognizers for spontaneous speech will enable a conversational system to switch seamlessly and automatically among different domains and/or languages. Preliminary results using a bi-domain recognizer exhibit only small recognition accuracy degradation in comparison to domain-dependent recognition. Similarly promising results were observed using a bi-lingual recognizer which performs simultaneous language identification and recognition. When using the FST techniques to add non-speech models to the recognizer, experiments show a 10% reduction in word error rate across all utterances and a 30% reduction on utterances containing non-speech events.

Volume 3, page 1591 Session D14

A Transducer Approach to Word Graph Generation

Boulianne G, Ouellet P, Dumouchel P
Centre de Recherche Informatique de Montréal, Canada

We describe word graph generation in terms of transducer composition, and show that a simple modification to a Viterbi search avoids the usual assumptions of word-pair or phone-pair approximations when the search space is represented with a transducer detailed down to the level of HMM transitions. On a 20,000-word French language dictation task, this graph generation method increases recognition time by only 20%. The word graphs produced can be further reduced in size by applying automata minimization, and this operation can be done faster than real-time. When the resulting graphs are rescored using larger acoustic and language models, recognition rate remains near-optimal for word graph densities as low as 8 words per spoken word.

Volume 3, page 1595 Session D14

An Efficient Implementation of Phonological Rules using Finite-State Transducers

Hetherington I L
MIT Laboratory for Computer Science, USA

Context-dependent phonological rules are used to model the mapping from phonemes to their varied phonetic surface realizations. Others, most notably Kaplan and Kay, have described how to compile general context-dependent phonological rewrite rules into finite-state transducers. Such rules are very powerful, but their compilation is complex and can result in very large nondeterministic automata. In this paper we present a simplified rewrite rule system and a technique to efficiently compile such a system into finite-state transducers.

Volume 3, page 1599 Session D14

A Weight Pushing Algorithm for Large Vocabulary Speech Recognition

Mohri M, Riley M
AT&T Labs - Research, USA

Weighted finite-state transducers provide a general framework for the representation of the components of speech recognition systems; language models, pronunciation dictionaries, context-dependent models, HMM-level acoustic models, and the output word or phone lattices can all be represented by weighted automata and transducers. In general, a representation is not unique and there may be different weighted transducers realizing the same mapping. In particular, even when they have exactly the same topology with the same input and output labels, two equivalent transducers may differ by the way the weights are distributed along each path. We present a "weight pushing" algorithm that modifies the weights of a given weighted transducer in a way such that the transition probabilities form a stochastic distribution. This results in an equivalent transducer whose weight distribution is more suitable for pruning and speech recognition. We demonstrate substantial improvements of the speed of our recognition system in several tasks based on the use of this algorithm. We report a 45% speedup at 83% word accuracy with a simple single-pass 40,000-word vocabulary North American Business News (NAB) recognition system on the DARPA Eval '95 test set. With the same technique, we report a 550% speedup at 88% word accuracy in rescoring NAB word lattices with more accurate 2nd-pass models. We finally report a 280% speedup at 68% word accuracy for 100,000 first name-last name pairs recognition.

Volume 3, page 1603 Session D14

Transducer optimizations for tight-coupled decoding

Seward A
KTH, Sweden

In this paper we apply a framework of finite-state transducers (FST) to uniformly represent various information sources and data-structures used in speech recognition. These source models include context-free language models, phonology models, acoustic model information (Hidden Markov Models), and pronunciation dictionaries. We will describe how this unified representation can serve as a single input model for the recognizer. We will demonstrate how the application of various levels of optimizations can lead to a more compact representation of these transducers and evaluate the effects on recognition performance, in terms of accuracy and computational complexity.

Volume 3, page 1607 Session D14

Session D15 - Poster
Thursday - 09.00 - 10.40

Speech Recognition and Understanding: Acoustic Modelling - II

Chair: Gerard Chollet, ENST, Paris, France

Distinctive Features For Use in an Automatic Speech Recognition System

Eide E
IBM, USA

In this paper we develop a method of representing the speech waveform in terms of a set of abstract, linguistic distinctions in order to derive a set of discriminative features for use in a speech recognizer. By combining the distinctive feature representation with our original waveform representation we are able to achieve a reduction in word error rate of 33 percent on an automatic speech recognition task.

Volume 3, page 1613 Session D15

Improved Context-Dependent Acoustic Modeling for Continuous Chinese Speech Recognition

Zhang J, Zheng F, Li J, Luo C, Zhang G
Tsinghua Univ., P. R. China

This paper describes the new framework of context-dependent (CD) Initial/Final (IF) acoustic modeling using the decision tree based state tying for continuous Chinese speech recognition. The Extended Initial/Final (XIF) set is chosen as the basic speech recognition unit (SRU) set according to the Chinese language characteristics, which outperforms the standard IF set. An adaptive mixture increasing strategy is applied when splitting the single Gaussian into mixed Gaussians in each tied state after the decision tree has been constructed. Our experimental results show that these two improvements are helpful to the acoustic modeling of Chinese speech recognition and that the CD XIF model outperforms the baseline syllable model over 30%.

Volume 3, page 1617 Session D15

Class Definition in Discriminant Feature Analysis

Duchateau J, Demuynck K, Van Compernolle D, Wambacq P
Katholieke Universiteit Leuven - ESAT, Belgium

The aim of discriminant feature analysis techniques in the signal processing of speech recognition systems is to find a feature vector transformation which maps a high dimensional input vector onto a low dimensional vector while retaining a maximum amount of information in the feature vector to discriminate between predefined classes. This paper points out the significance of the definition of the classes in the discriminant feature analysis technique. Three choices for the definition of the classes are investigated: the phonemes, the states in context independent acoustic models and the tied states in context dependent acoustic models. These choices for the classes were applied to (1) standard LDA (linear discriminant analysis) for reference and to (2) MIDA, an improved, mutual information based discriminant analysis technique. Evaluation of the resulting linear feature transforms on a large vocabulary continuous speech recognition task shows, depending on the technique, the best choice for the classes.

Volume 3, page 1621 Session D15

Feature Extraction from Time-Frequency matrices for Robust Speech Recognition

Segura J C[1], Benítez M C[2], de la Torre Á[1], Rubio A J[1]
[1]Universidad de Granada, Spain, [2]Universidad de Granada , Spain / International Computer Science Institute, USA

In this paper we present a study about time-frequency distribution of acoustic-phonetic information for the Spanish language. This is based on a large Spanish database automatically labeled, and we conclude that results are similar to those obtained for hand-labeled English databases. We use bidimensional LDA to extract discriminant features in time-frequency domain (TF) that are more robust in noise than the standard ones based on MFCC and time derivatives. We show that TF domain and its corresponding transformed domain (CTM) are equivalent from the point of view of LDA analysis and use this fact to reduce the dimensionality of the problem. Finally, cascade unidimensional LDA (CLDA) is applied first in frequency and then in time. This gives better estimates of projection vectors and better recognition performance. The proposed techniques are evaluated in a connected digit recognition task. Utterances have been artificially corrupted with additive real noises.

Volume 3, page 1625 Session D15

Using Spatial Correlation Information in Speech Recognition

Yu P, Wang Z
Tsinghua University, P. R. China

Acoustic model training is very important in speech recognition. But in traditional training algorithm, we take each state separately, and the relationship between different states is not considered. In this paper we bring forward a novel idea of using the correlation information between states, which is called ~{!0~}spatial correlation~{!1~}. We describe this correlation information as linear constraints. According to phonetic knowledge, we firstly divide states into small groups named ~{!0~}correlation sub-space~{!1~}. In every sub-space, we use eigen value decomposition to get linear constraints. The constraints are then used in a new training algorithm. Experiments of the new training algorithm show significant improvement over traditional training algorithm.

Volume 3, page 1629 Session D15

On the Choice of Classes in MCE based discriminative HMM-Training for Speech Recognizers used in the Telephone Environment

Bauer J G
Siemens, Germany

One of the most commonly used discriminative approaches in parameter estimation for Hidden Markov Models is the Minimum Classification Error (MCE) method . This paper studies possible choices for the classes (i.e. basic speech units) in MCE training and their application for several tasks suitable for speech driven dialog systems in the telephone environment. The considered choices of classes are HMM states, phonemes, words and sequences of words. The theoretical suitability and practical considerations for the different criteria are discussed. Using the different training criteria consistent experimental results are given for four tasks: non-task-specific training, training for small vocabulary isolated word recognition, training for connected digit recognition and for letter recognition. In all experiments not only the objective of the optimization but also the resulting word recognition performance is investigated. It shows that for the given setup only word and word string based criteria are capable to reduce the word error rate.

Volume 3, page 1633 Session D15

Plosive Spotting with Margin Classifiers

Keshet J[1], Chazan D[2], Bobrovsky B-Z[1]
[1]Tel Aviv University, Israel, [2]IBM Israel - Science and Technology, Israel

This paper presents a novel algorithm for precise spotting of plosives. The algorithm is based on a pattern matching technique implemented with margin classifiers, such as support vector machines (SVM). A

special hierarchical treatment to overcome the problem of fricative and false silence detection is presented. It uses the loss-based multi-class decisions. Furthermore, a method for smoothing the overall decisions by sequential linear programming is described. The proposed algorithm was tested on the TIMIT corpus, which produced a very high spotting accuracy. The algorithm presented here is applied to plosives detection, but can easily be adapted to any class of phonemes.

Volume 3, page 1637 Session D15

Model Agglomeration for Context-Dependent Acoustic Modeling

Brugnara F
ITC-Irst - Centro per la Ricerca Scientifica e Tecnologica, Italy

This work describes a method for generating back-off models for context-dependent unit modeling. The main characteristic of the approach is that of building generic models by gathering statistics of detailed models, collected during Baum-Welch reestimation. The construction of back-off models does not require additional processing of the training data, allowing to quickly build different models sets with different back-off criteria starting from the same set of trained models and their statistics. Experiments are reported on the TIMIT and Wall Street Journal corpora, that show the consistency of the approach and compare it with state tying based on Phonetic Decision Trees.

Volume 3, page 1641 Session D15

Multipass algorithm for acquisition of salient acoustic morphemes

Levit M, Gorin A L, Wright J H
AT&T Laboratories-Research, USA

We are interested in spoken language understanding within the domain of automated telecommunication services. Our current methodology involves training statistical language models from large annotated corpora for recognition and understanding. Since the transcribing of large speech corpora is a resource consuming task, we are motivated to exploit speech without transcriptions. In particular, we learn the semantic associations for a task exploiting only phone-based sequences from the output of a task-independent ASR-system. In this paper we present a new multipass algorithm for acquiring salient phone sequences from untranscribed speech corpora and evaluate their utility for the HMIHY task. Compared to our previous strategy, this algorithm is shown to produce improved call-classification results while reducing up to 7-fold the number of salient phone-sequences selected for training.

Volume 3, page 1645 Session D15

Rapid Vocal Tract Length Normalization using Maximum Likelihood Estimation

Emori T, Shinoda K
NEC Corporation, Japan

Recently, vocal tract length normalization (VTLN) techniques have been developed for speaker normalization in speech recognition. This paper proposes a new VTLN method, in which the vocal tract length is normalized in the cepstrum space by means of linear mapping whose parameter is derived using maximum-likelihood estimation. The computational costs of this method are much lower than that of such conventional methods as ML-VTLN, in which the parameter for mapping is selected from among several parameters. Further, the new method offers greater precision in determining parameters for individual speakers. Experimental use of the method resulted in an error reduction rate of 7.1%. A combination of the proposed method with cepstrum mean normalization (CMN) method was also examined and found to reduce the error rate even more, by 14.6%.

Volume 3, page 1649 Session D15

Towards the creation of acoustic models for stressed Japanese speech

Okuda K, Matsui T, Nakamura S
ATR Spoken Language Translation Research Laboratories, Japan

In error recovery utterance, the user using the speech recognition system changes his or her speaking style to aid the system in recognizing the speech. However, this change leads the mismatch between the acoustic models and reduces the performance of the system. This degradation causes a serious problem of speech recognition for a dialog system or a speech translation system. In error recovery utterance in Japanese, the occurrence of syllable-stressed speech increases. In syllable-stressed speech, each syllable is uttered slowly and emphasized. The characteristics of each syllable are strongly altered by this modification and the speech recognition performance is reduced. This paper investigates how to create acoustic models robust in recognizing error recovery utterances, especially syllable-stressed speech. In this paper, we propose an acoustic modeling method for syllable-stressed speech by combining existing acoustic models. Our results indicate that the proposed method improves the system performance. Furthermore, the method does not need any expansion of the recognition dictionary or explicit model selection.

Volume 3, page 1653 Session D15

Elderly Acoustic Model for Large Vocabulary Continuous Speech Recognition

Baba A[1], Yoshizawa S[2], Yamada M[3], Lee A[3], Shikano K[3]
[1]Matsushita Electric Works, Japan, [2]Matsushita Electric Industrial Co., Japan, [3]Nara Institute of Science and Technology, Japan

In this paper, we evaluate elderly speaker acoustic models in LVCSR, which are trained by the 301 elderly speakers' database from the age of 60 to 90. Each speaker utters 200 sentences. The elderly speaker PTM (Phonetic Tied Mixture) acoustic model attains 88.9% word recognition rate, which is better than 86.0% word recognition rate by the usual adult (an average age of 28.6) PTM acoustic model. To achieve higher recognition rates, we use two types of speaker adaptation methods, which are a supervised MLLR and an unsupervised adaptation method based on the sufficient HMM statistics. In our experimental results, the elderly acoustic model is better as the adaptation baseline HMM model than the usual adult model for elderly speakers.

Volume 3, page 1657 Session D15

A Hybrid Approach to Enhance Task Portability of Acoustic Models in Chinese Speech Recognition

Zhang J-S, Zhang S-W, Sagisaka Y, Nakamura S
ATR Spoken Language Translation Research Laboratories, Japan

This paper presents our approach to enhance the portability of acoustic models by mitigating the phonetic mismatch arising from a new testing task which is rather different from the training data. The approach is a hybrid one which combines knowledge-based context categorization to generate a context rich set of subword units, and data-driven-based acoustic model clustering on the level of context category. Compared with the conventional approach of only phonetic decision tree based model clustering and unseen model generation, the new approach improved greatly the desired subword coverage for the new testing domain, and achieved an error rate reduction by 10.8% for Chinese character accuracy in the recognition experiments. Together with the effect of the newly adopted basic units of 9 glottal stops, we achieved a total 23.5% error rate reduction in the testing compared to the baseline system.

Volume 3, page 1661 Session D15

Evaluation of sublexical and lexical models of acoustic disfluencies for spontaneous speech recognition in Spanish

Rodriguez L J, Torres I, Varona A
Universidad del Pais Vasco, Spain

Spontaneous speech is full of acoustic disfluencies that rarely appear in read or laboratory speech. A very simple and straightforward approach is presented, in which acoustic disfluences are modelled by augmenting the inventory of sublexical units, which originally consisted of 23 context independent phones plus a special unit for silent pauses. This set was augmented with 12 additional units accounting for lengthenings of sounds, filled pauses and noises. Two speech databases, both in Spanish, were used in the experiments. A phonetically balanced database was used for initializing the acoustic models. A spontaneous speech database consisting of 227 dialogues was used both for training and testing purposes. Recognition rates, in terms of acoustic-phonetic accuracy and word accuracy, with and without filtering acoustic disfluencies prior to alignments, were obtained to evaluate the contribution of these models to speech recognition. Also, some specific but significant examples were explored and discussed. Experimental results showed that using explicit models of acoustic disfluencies clearly improved the performance of a spontaneous speech recognition system.

Volume 3, page 1665 Session D15

Structural Learning of Dynamic Bayesian Networks in Speech Recognition

Deviren M, Daoudi K
INRIA-LORIA, France

We present a speech modeling methodolgy where no a priori assumption is made on the dependencies between the observed and hidden speech processes. Rather, dependencies are learned from data. This methodology guarantees improvement in modeling fidelity compared to HMMs. In addition, it gives the user a control on the trade-off between modeling accuracy and model complexity. Furthermore, the approach is technically very attractive because all the computational effort is made in the training phase.

Volume 3, page 1669 Session D15

Session D16 - Poster
Thursday - 09.00 - 10.40

Resources, Assessment and Standards: Assessment Tools & Methodology

Chair: David Pallett, NIST, US

A New Method for Testing Communication Efficiency and User Acceptability of Speech Communication Channels

Van Wijngaarden S, Smeele P, Steeneken H
TNO Human Factors, the Netherlands

The performance of speech communication channels featuring long delay times is usually subjectively experienced as lower than similar channels without delay. Yet most conventional speech intelligibility and speech quality tests are not sensitive to the effects of delay. Moreover, these conventional test do not take the effects of human compensating strategies into account, which help cope with adverse communication conditions by adapting our speech. Test types that do incorporate such effects are sometimes known as 'speech communicability' tests. Based on the lessons learned from literature on speech communicability testing, a list of requirements for the design of a good communicability test method was composed, followed by the actual design of a new test method combining attractive features of existing communicability tests. The suitability of the test design was verified by conducting a pilot experiment. The results of this experiment show that the new method is capable of measuring efficiency and acceptability, and is sufficiently sensitive to delay and background noise.

Volume 3, page 1675 Session D16

Phonetic Transcriptions in the Spoken Dutch Corpus: how to Combine Efficiency and Good Transcription Quality

Cucchiarini C, Binnenpoorte D, Goddijn S
University of Nijmegen, The Netherlands

This paper reports on an experiment aimed at establishing how phonetic transcriptions for the large CGN corpus can be obtained most efficiently. This experiment explores the po-tential of an automatically generated transcription (AGT) by comparing an AGT with a reference transcription (Tref) of the same material, to determine whether and how the AGT can be improved to make it more similar to Tref. The results indicate that the AGT can be optimized through pronunciation variation modelling so as to make human corrections more efficient or even superfluous, at least for some speech styles.

Volume 3, page 1679 Session D16

A Functional Approach to Speech Recognition Evaluation

Hutchinson B
Syrinx Speech Systems, Australia

The paper describes a new evaluation measure for speech recognition in spoken language dialogue systems. The measure is based on the usefulness of the recognition for the system, and the usefulness is measured at the level of meaning representation. It is argued that the new measure is more useful than word error rate, and is more accurate than simpler functional measures.

Volume 3, page 1683 Session D16

Instrumental Derivation of Equipment Impairment Factors for Describing Telephone Speech Codec Degradations

Möller S[1], Berger J[2]
[1]Ruhr-University Bochum, Germany, [2]T-Nova Deutsche Telekom Innovationsges. mbH Berkom, Germany

The impairment factor methodology has been adopted by the ITU-T for describing the relative impact of telephone transmission degradations on the overall quality of transmitted speech. Input parameters to this methodology are mainly instrumentally measurable characteristics of the transmission path, with the exception of low bit-rate codecs, whose perceptual characteristics still have to be determined in auditory tests. In this paper, we describe a new approach for deriving impairment factors for low bit-rate codecs in a purely instrumental way. Using instrumental quality prediction models like PESQ or TOSQA, quality estimations are obtained which can be combined with other degradations in order to obtain an overall quality estimation for the whole transmission channel. A comparison with defined values for well-known codecs shows a high correlation of instrumentally derived impairment factors with the corresponding defined values, as well as with auditory test data.

Volume 3, page 1687 Session D16

Julius --- an Open Source Real-Time Large Vocabulary Recognition Engine

Lee A[1], Kawahara T[2], Shikano K[1]
[1]Nara Institute of Science and Technology, Japan, [2]Kyoto University, Japan

Julius is a high-performance, two-pass LVCSR decoder for researchers and developers. Based on word 3-gram and context-dependent HMM, it can perform almost real-time decoding on most current PCs in 20k word dictation task. Major search techniques are fully incorporated such as tree lexicon, N-gram factoring, cross-word context dependency handling, enveloped beam search, Gaussian pruning, Gaussian selection, etc. Besides search efficiency, it is also modularized carefully to be independent from model structures, and various HMM types are supported such as shared-state triphones and tied-mixture models, with any number of mixtures, states, or phones. Standard formats are adopted to cope with other free modeling toolkit. The main platform is Linux and other Unix workstations, and partially works on Windows. Julius is distributed with open license together with source codes, and has been used by many researchers and developers in Japan.

Volume 3, page 1691 Session D16

Local Refinement of Phonetic Boundaries: A General Framework and Its Application Using Different Transition Models

Torre Toledano D[1], Hernández Gómez L A[2]
[1]Telefónica I+D, Spain, [2]Universidad Politécnica de Madrid, Spain

In the last few years we have been experimenting with an automatic phonetic segmentation and labeling system based on a modified HMM phonetic recognizer followed by a local phonetic boundary refinement system. During this period we have used different approaches for the local refinement, including fuzzy rules and neural networks. In this paper we present a unified framework for the local refinement of phonetic boundaries that has allowed us to thoroughly evaluate and compare these approaches and yet another one based on gaussian mixture models. Results show that neural networks outperform the rest of the approaches in speaker dependent mode, achieving a precision almost equal to a manual segmentation. In speaker independent mode, however, neural networks and fuzzy rules achieve almost the same performance, a bit worse than a manual segmentation.

Volume 3, page 1695 Session D16

Detection of Digital Transmission Systems for Voice Quality Measurements

Ludwig T, Heute U
Christian-Albrechts University of Kiel, Germany

In-service, Non-intrusive Measurement Devices (INMD) estimate the perceived quality of the telephone link by extracting quality-defining criteria like echo attenuation, echo delay, active speech level, noise level, frame losses and transient failures from a telephone call. In addition, the quality depends on the used digital transmission systems (codec systems). This paper proposes a method to distinguish between two codec classes. With the help of features determined from the speech signal, a classifier decides about the class affiliation of the signal. The recognition rate for signals with 16 seconds of active speech is about 97%.

Volume 3, page 1699 Session D16

Automatic Segmentation of Recorded Speech into Syllables for Speech Synthesis

Lewis E[1], Tatham M[2]
[1]University of Bristol, UK, [2]University of Essex, UK

Concatenated waveform text-to-speech synthesis systems require an inventory of stored waveforms from which units of speech can be extracted for subsequent rearrangement and concatenation as needed. In previous papers [1], [2] we have argued that for natural sounding speech the syllable should be the preferred unit. The mark-up of the stored waveforms for segmentation into syllables must be precise and for our MeteoSPRUCE limited domain system the mark-up has been done by manual editing. In this paper we describe how most of the segmentation can be done automatically, leaving only those waveforms which would be prone to error to be segmented manually. With automatic labelling of both the pitch periods and the syllables the task of generating different synthetic voices to order becomes feasible.

Volume 3, page 1703 Session D16

Phonetic Events from the Labeling the European Portuguese DataBase for Speech Synthesis, FEUP/IPB-DB

Teixeira J P[1], Freitas D[2], Braga D[2], Barros M J[2], Latsch V[2]
[1]Instituto Politécnico de Bragança, Portugal, [2]Faculdade de Engenharia da Universidade do Porto, Portugal

In this paper a labeled new speech signal database (FEUP/IPB-DB) in Standard European Portuguese is presented. The objective of this work is, on one hand, to provide phonetic material for TTS systems construction, either from the start or to improve the quality of existing ones, and, on the other hand, to place at service of the European Portuguese scientific community a phonetically and prosodically valuable speech corpus, essential for Speech Synthesis or Phonetics research. The main features of the database will be described as well as some basic statistical aspects. A discussion of some methodological problems and some observed phenomena in experimental phonetics deriving from the speech signal labeling is also done. The approach in our work is to produce a resource that can be further improved in subsequent steps with minimal re-work. The phonetic, linguistic and technical consistency are guaranteed through the involvement of a multidisciplinary team.

Volume 3, page 1707 Session D16

Acoustical and topological experiments fo an HMM-based speech segmentation system

Nefti S[1], Boeffard O[2]
[1]France Télécom R&D, DIH/IPS/VMI, France, [2]IRISA, Université de Rennes 1, ENSSAT, France

Several specific tasks in the field of text-to-speech synthesis requires a huge amount of labeled speech corpora. Mostly, these labels correspond to phone marks aligned on the speech waveform. Different kind of solutions have been applied to this problem from rule-based systems to stochastic-based ones. We validate here a solution based on Hidden Makov Models. Various test configurations are proposed. At the acoustic level, we compare LSP to MFCC coefficients and the fitness of multigaussians for this segmentation task. At the topological level, we compare standard left-to-right models to phonological dependent topologies. The best configuration we found is related to an MFCC analysis with standard left-to-right models and with diagonal multi-gaussians per state. For this configuration the overall root mean squared error on the test database is 18 +/- 0.3 ms within a 99% confidence interval.

Volume 3, page 1711 Session D16

TclBLASR: An Automatic Speech Recognition Extension for Tcl

Zhou Q, Zheng J, Lee C-H
Bell Labs, Lucent Technologies, USA

We present TclBLASR, a framework to integrate a proprietary speech recognition engine, an open source script language, such as Tcl/Tk and an open source sound analysis toolkit, such as Snack from KTH, into a user friendly platform that a user can write a Tcl/Tk script application quickly for speech recognition evaluation, speech data collection and automatic annotation, and speech technology demonstration. This framework is extremely useful for third party customer evaluation of speech technologies that do not involve heavy C/C++ program development and extensive knowledge on low-level speech engine APIs. Using the Bell Labs Automatic Speech Recognition (BLASR) engine, coupled with the real-time audio I/O and visualization provided by Snack and the flexible graphical user interface tools embedded in Tcl/Tk, the TclBLASR platform proves to be a useful framework for quick packaging of ASR engines for customer evaluation of the technology without extensive customization of interfaces to meet different needs from a wide range of customers.

Volume 3, page 1715 Session D16

Session D21 - Oral
Thursday - 11.10 - 12.30

ESE5 - Existing and Future Corpora: Automated Analysis of Speech Resources

Chair: Florian Schiel, LMU, München, Germany

Lower WERs do not guarantee better transcriptions

Kessens J M, Strik H
University of Nijmegen, The Netherlands

The goal of this paper is to investigate the effect of various properties of the CSR on automatic transcription. To this end, we used various versions of a continuous speech recognizer (CSR) to make automatic transcriptions. Our results show that changing certain properties of the CSR affects the resulting automatic transcriptions. The best results were obtained when 'short' hidden Markov models (HMMs), and context-independent HMMs were used. Furthermore, we found that minimizing the amount of contamination in the HMMs improves the quality of the automatic transcriptions. Another important result is that there does not appear to be a straightforward relation between word error rate (WER) and the transcription quality. In other words: A CSR with a lower WER does not always guarantee better transcriptions.

Volume 3, page 1721 Session D21

An Elitist Approach to Articulatory-Acoustic Feature Classification

Chang S[1], Greenberg S[1], Wester M[2]
[1]International Computer Science Institute, USA, [2]Nijmegen University, The Netherlands

A novel framework for automatic articulatory-acoustic feature extraction has been developed for enhancing the accuracy of place- and manner-of-articulation classification in spoken language. The “elitist” approach focuses on frames for which neural network (MLP) classifiers are highly confident, and discards the rest. Using this method, it is possible to achieve a frame-level accuracy of 93% for manner information on a corpus of American English sentences passed through a telephone network (NTIMIT). Place information is extracted for each manner class independently, resulting in an appreciable gain in place-feature classification relative to performance for a manner-independent system. The elitist framework provides a potential means of automatically annotating a corpus at the phonetic level without recourse to a word-level transcript and could thus be of utility for developing training materials for automatic speech recognition and speech synthesis applications, as well as aid the empirical study of spoken language.

Volume 3, page 1725 Session D21

A Dutch Treatment of an Elitist Approach to Articulatory-Acoustic Feature Classification

Wester M[1], Greenberg S[2], Chang S[2]
[1]Nijmegen University, The Netherlands, [2]International Computer Science Institute, USA

A novel approach to articulatory-acoustic feature extraction has been developed for enhancing the accuracy of classification associated with place and manner of articulation information. This “elitist” approach is tested on a corpus of spontaneous Dutch using two different systems, one trained on a subset of the same corpus, the other trained on a corpus from a different language (American English). The feature dimensions, voicing and manner of articulation transfer relatively well between the two languages. However, place information transfers less well. Manner-specific training can be used to improve classification of articulatory place information.

Volume 3, page 1729 Session D21

Session D22 - Oral
Thursday - 11.10 - 12.30

Dialogue Systems: Dialogue Systems and Generation

Chair: Keikichi Hirose, Uni. of Tokyo

Hybrid Natural Language Generation for Spoken Dialogue Systems

Galley M, Fosler-Lussier E, Potamianos A
Lucent Technologies, USA

The natural language generation component of most dialogue systems is based on templates. Template-based generators are hard to maintain and reuse, and the sentences they produce lack the variability and robustness needed by conversational systems. In this paper, a flexible and domain-independent natural language generator for spoken dialogue systems is proposed which combines fixed surface expressions with freely generated text. The generation algorithm follows a hybrid approach, combining finite state machine (FSM) grammars and corpus-based language models. In this approach, the FSM grammar (a reversible parser grammar) is constrained by a word and concept n-gram that takes terminals and non-terminal co-occurrences into account. The n-gram grammar helps prevent inappropriate derivations, therefore improving the quality of the generated texts. The proposed algorithm achieves faster than real-time performance because of the limited number of derivations.

Volume 3, page 1735 Session D22

The Generation of Speech for a Search Guide

Cook N, Benest I
University of York, UK

A major problem with any interface to a hierarchical information system, however shallow the hierarchy might be, is that information below the current level is hidden from view. To determine whether there is useful information at any level below the current one, requires an inference-based look-and-ponder process followed by a tedious point-click-wait-read-back process of manipulation. This equally applies to the results obtained from a search engine. An alternative is to provide a search interface that offers oral cues to buried information and relies on the intelligence of the user to recognise the usefulness behind the cues. The result will be a conversational search guide and this paper addresses the production of speech utterances, using pre-recorded speech, so that the guide remains almost as fresh as a human guide.

Volume 3, page 1739 Session D22

An Automatic Dialogue System Generator from the Internet Information Contents

Araki M, Ono T, Ueda K, Nishimoto T, Niimi Y
Kyoto Institute of Technology, Japan

We propose a semi-automatic dialogue system generator from the Internet information contents. We classify the practical Web site into three classes of task: slot-filling, database search, and explanation. Using three levels of dialogue library for each task, our generator translates XML based Web site into VoiceXML, which controls a conversation between a user and a computer system. In this paper, we explain an outline of our project and report implementation examples.

Volume 3, page 1743 Session D22

Training a Sentence Planner for Spoken Dialog: The Impact of Syntactic and Planning Features

Rogati M[1], Walker M[2], Rambow O[2]
[1]Carnegie Mellon University, USA, [2]AT&T Shannon Labs, USA

The dialog manager of a spoken dialog system often performs domain dependent functions as well as general dialog tasks. It is possible to separate the domain specific knowledge from knowledge about language using techniques from natural language generation. However a natural language generator often has to be tuned for particular applications. In this work, we describe a new method for automatically training the natural language generator and examine the role that domain specific and domain independent features have on performance. We show that although the general features have the largest impact, the use of domain specific features improves performance, while still retaining the benefits of automatic domain customization through training.

Volume 3, page 1747 Session D22

Session D23 - Oral
Thursday - 11.10 - 12.30

Speaker Recognition: Alternative Trends in Verification - I

Chair: George Doddington, NIST, USA

A Comparative Study of MLP-based Artificial Neural Networks in Text-Independent Speaker Verification Against GMM-based Systems

Vivaracho C E[1], Ortega-García J[2], Alonso L[3], Moro Q I[1]
[1]Universidad de Valladolid, Spain, [2]Universidad Politécnica de Madrid, Spain, [3]Universidad de Salamanca, Spain

Text-independent speaker verification is an interesting task where the use of Gaussian Mixture Models is almost a must. Nevertheless, some preliminar encouraging results obtained in previous works using ANN in speaker verification have led us to consider to perform a direct comparison between these different methods. In this sense, this paper is only focused on the classification stage of both GMM-based and ANN-based speaker verification systems. Experiments are accomplish making use of the AHUMADA/GAUDI spanish speech database, specially oriented for speaker-recognition tasks as it contains multisession and multichannel data of about 500 speakers. Results confirm a better performance when using GMM-based system and microphonic speech but, on the other hand, when testing in specific conditions and with real telephone speech ANN outperforms GMM results.

Volume 3, page 1753 Session D23

Enhancing GMM Scores using SVM "Hints"

Fine S, Navratil J, Gopinath R
IBM T.J. Watson Research Center, USA

This paper proposes a classification scheme that combines statistical models and support vector machines. It exploits the fact that GMM and SVM classifiers with roughly the same level of performance produce uncorrelated errors. We describe a novel scheme which employs an SVM classifier as an ``advisor'' to the GMM classifier in uncertain cases. The utility of the combined generative/discriminative approach is demonstrated on standard text-independent speaker verification and speaker identification tasks in matched and mismatched training and test conditions. Results indicate significant improvements in performance without much computational overhead.

Volume 3, page 1757 Session D23

Combining GMM's with Suport Vector Machines for Text-independent Speaker Verification

Kharroubi J, Petrovska-Delacretaz D, Chollet G
ENST, CNRS-LTCI, France

Current best performing speaker recognition algorithms are based on Gaussian Mixture Models (GMM). Their results are not satisfactory for all experimental conditions, especially for the mismatched between train and test conditions. Support Vector Machine is a new and very promissing technique in statistical learning theory. Recently, this technique produced very interesting results in image processing and for the fusion of experts in biometric authentification. In this paper we address the issue of using the Support Vector Learning technique in combination with the currently well performing GMM models, in order to improve speaker verification results.

Volume 3, page 1761 Session D23

A Text-Independent Speaker Verification System Using Support Vector Machines Classifier

Gu Y, Thomas T
Vocalis Ltd., UK

In the recent years the technology for speaker verification or call authentication has received an increasing amount of attention in IVR industry. However due to the complexity of speaker information embedded in the speech signals the current technology still can not produce the verification accuracy to meet the requirement for some applications. In this paper we introduce a new pattern classification approach, support vector machines (SVM) for the text-independent speaker verification. The SVM is a new way of statistical learning based on a principle of structural risk minimisation. In the paper various evaluation results for the SVM verification system are presented and a comparison with a baseline GMM approach is also given. The results demonstrate that the SVM approach perform much better than the GMM approach. On the same training and testing data set the SVM approach gives an EER 1.2% versus 3.9% EER from the GMM approach.

Volume 3, page 1765 Session D23

Session D24 - Oral
Thursday - 11.10 - 12.30

Speech Recognition and Understanding: Speech Understanding

Chair: Hermann Ney, Aachen University of Technology, Germany

Advances in Automatic Speech Summarization

Hori C, Furui S
Tokyo Institute of Technology, Japan

This paper reports recent advances in automatic speech summarization method. In our proposed method, a set of words maximizing a summarization score is extracted from automatically transcribed speech. This extraction is performed according to a target compression ratio using a dynamic programming technique. The extracted set of words is then connected to build a summarized sentence. The summarization score consists of a word significance measure, a confidence measure, linguistic likelihood, and a word concatenation probability which is determined by a dependency structure in the original speech given by Stochastic Dependency Context Free Grammar. Japanese broadcast news speech transcribed using a large vocabulary continuous speech recognition system is summarized and evaluated in comparison with manual summarization by human subjects. The manual summarization results are combined to build a word network, and word accuracy of each automatic summarization result is calculated comparing with the most similar word string in the network.

Volume 3, page 1771 Session D24

A Word Graph Interface for a Flexible Concept Based Speech Understanding Framework

Hacioglu K, Ward W
University of Colorado, USA

In this paper, we introduce a word graph interface between speech and natural language processing systems within a flexible speech understanding framework based on stochastic concept modeling augmented with background "filler" models. Each concept represents a set of phrases (written as a context free grammar (CFG)) with the same meaning, and is compiled into a stochastic recursive transition network (SRTN). The arcs (or rules) are tagged with probabilities after training. The filler models are used for phrases that are not covered by the concept networks. The structure in concept+filler sequences is captured by n-grams. The interface is implemented within the context of CU Communicator spoken dialog system. We investigate the effect of several different filler models and interpolation of complementary language models on the system performance. We report notable performance improvements compared to the baseline system. The gain in performance along with the efficiency and flexibility of the method motivates future work on the implementation of a tighter interface.

Volume 3, page 1775 Session D24

Comparing grammar-based and robust approaches to speech understanding: a case study

Knight S[1], Gorrell G[2], Rayner M[2], Milward D[1], Koeling R[2], Lewin I[2]
[1]SRI International, UK, [2]Netdecisions, UK

Previous work has demonstrated the success of statistical language models when enough training data is available, but despite that, grammar-based systems are proving the preferred choice in successful commercial systems such as HeyAnita, BeVocal and Tellme, largely due to the difficulty involved in obtaining a corpus of training data. Here we trained an SLM on data obtained using a grammar-based system and compared the performance of the two systems with regards to recognition. We also parsed the output of the SLM using a robust parser

and compared the accuracy of the semantic output of the systems. The SLM/robust parser showed considerable improvement on unconstrained input, and similar precision/recall (per slot value) on utterances provided by trained users.

Volume 3, page 1779 Session D24

Integrating Multiple Knowledge Sources For Improved Speech Understanding

Abdou S, Scordilis M
University of Miami, USA

In spoken dialog systems it is often the case that the sentence produced by the decoder with the highest recognition probability may not be the best choice for extracting the intended concepts. Lower ranking hypotheses may present better alternatives. In this paper, we show how to integrate multiple knowledge sources for the decision of selecting one of these hypotheses. A scoring schema combining information from the recognizer output, the parser, an utterance type classifier and dialog context is used. The scaling weights of the combined scores are determined automatically by an optimization procedure. Finally, we show the results of testing this approach and its performance compared to the approach of selecting the best recognition hypothesis.

Volume 3, page 1783 Session D24

Session D25 - Poster
Thursday - 11.10 - 12.30

Speech Recognition and Understanding: Algorithms and Architectures

Chair: Kuldip Paliwal, Griffith University, Australia

Classification of Transition sounds with application to Automatic Speech Recognition

Litichever Z[1], Chazan D[2]
[1]Technion Technology Institute, Israel, [2]IBM Israel - Science and Technology, Israel

This paper addresses the problem of classification of speech transition sounds. A number of non parametric classifiers are compared, and it is shown that some non-parametric classifiers have considerable advantages over traditional hidden Markov models. Among the non parametric classifiers, support vector machines were found the most suitable and the easiest to tune. Some of the reasons for the superiority of non parametric classifiers will be discussed. The algorithm was tested on the voiced stop consonant phones extracted from the TIMIT corpus and resulted in very low error rates.

Volume 3, page 1789 Session D25

Gaussian Subtraction (GS) Algorithms for Word Spotting in Continuous Speech

Faizakov A[1], Cohen A[1], Vaich T[2]
[1]Ben-Gurion University, Israel, [2]Speech Recognition R&D Center, Israel

In this paper, a novel approach for the design of cohort models for word spotting in continuous speech is presented. This new approach is based on modifying the probability density function of a conventional filler so that regions in the feature space that are related to the keyword will be reduced or removed. By modifying these regions, the filler and keyword models become more orthogonal in the sense that they represent different areas in the feature space, making the filler appropriate to be used as a cohort model. The algorithms, named Gaussian Subtraction (GS) and Gaussian Removal (GR), may be considered discriminative training algorithms.

Volume 3, page 1793 Session D25

Relating Frame Accuracy with Word Error in Hybrid ANN-HMM ASR

Shire M L
International Computer Science Institute, USA

Frame accuracy is a common and natural summary statistic to use in neural-network-based ASR. It is often used as an indication of the performance of the neural network probability estimator and in the stopping criterion during its training. Though considered an important factor for word recognition, the frame accuracy presents an incomplete and sometimes deficient indicator of performance for the overall task of word recognition, as with many such summary statistics. Many in the ASR community have seen instances where an improvement in the acoustic posterior probability estimation yielded a disappointing effect on word recognition. We conducted experiments in an effort to illustrate some of the variability in word-recognition performance associated with frame accuracy. Our experiments attempt to shed light on some of the factors that might give rise to instances where frame accuracy and word error correlate. Some of the results are confirmation of intuitive or commonly known trends.

Volume 3, page 1797 Session D25

A Two-Layer Lexical Tree Based Beam Search in Continuous Chinese Speech Recognition

Zhang G, Zheng F, Wu W
Tsinghua Univ., P. R. China

In this paper, an approach to continuous speech recognition based on a two-layer lexical tree is proposed. The search network is maintained by the two-layer lexical tree, in which the first layer reflects the word net and the phone net while the second layer the dynamic programming (DP). Because the acoustic information is tied in the second layer, the memory cost is so small that it has the ability to process some complicated applications, such as the use of cross-word context-dependent (CD) triphone models, the Chinese fuzzy syllable mapping and the pronunciation modeling. The search algorithm based on the two-layer lexical tree is also proposed, which is derived from the token-passing algorithm. Finally, an implementation of the two-layer lexical tree using the cross-word context-dependent triphone models is presented, and the experimental results show that the highly efficient decoding can be achieved without too much memory cost.

Volume 3, page 1801 Session D25

Automatic Labeling and Digesting for Lecture Speech Utilizing Repeated Speech by Shift CDP

Itoh Y[1], Tanaka K[2]
[1]Iwate Prefectural University, Japan, [2]Electrotechnical Laboratory, Japan

This paper proposes an automatic labeling and digesting method for lecture speech. The method utilizes same sections, such as same words or same phrases that are thought to be important and are repeated in the speech. To extract the same sections, we have proposed a new efficient algorithm, called Shift Continuous DP, because it is an extension of Continuous DP and realizes fast matching between arbitrary sections in two speech data sets frame-synchronously. Shift CDP is extended to extract same sections in single long speech data in this paper. This paper describes ways to apply the algorithm to labeling and digesting for a lecture speech. We conduct some preliminary experiments to show the method can extract same sections and a sequence of extracted sections can be regarded as a digest of the speech.

Volume 3, page 1805 Session D25

Improved Phoneme-History-Dependent Search for Large-Vocabulary Continuous-Speech Recognition

Hori T[1], Noda Y[2], Matsunaga S[1]
[1]NTT Cyber Space Laboratories, Japan, [2]NTT Communications Corporation, Japan

This paper describes an improved phoneme-history-dependent (PHD) search algorithm. This method is an optimum algorithm under the assumption that the starting time of a word depends on only a few preceding phonemes (phoneme history). The computational cost and number of recognition errors made by a multi-pass-based recognizer can be reduced if the PHD search of the first decoding pass uses re-selection of the preceding word and the optimum length of phoneme histories. These improvements increase the speed of the first decoding pass and help that the word lattice has the correct word sequence. Consequently search errors can be reduced in the second decoding pass. In 65k-word domain-independent Japanese read-speech dictation task and 1000-word spontaneous-speech airplane-reservation task, the improved PHD search was 1.2-2.0 times faster than a traditional word-dependent search under the condition of equal word accuracy.

Volume 3, page 1809 Session D25

Comparison of MFCC and PLP Parameterizations in the Speaker Independent Continuous Speech Recognition Task

Psutka J, Muller L, Psutka J V
University of West Bohemia, Czech Republic

The authors of this paper wish to contribute to the discussion about an optimal parameterization of speech signals in speech recognition systems. Our experiments deal with a telephone-based speaker independent continuous speech recognition task in which the MFCC and PLP parameterizations were tested and compared. The benefit of an adjustment of the filters used in the MFCC and PLP parameterizations to the critical bandwidth of hearing was explored and the impact of the number of filters and enumerated parameters to the recognition accuracy was tested. The results of these experiments showed that the MFCC parameterization is less sensitive to satisfying the theory of the critical bandwidth of hearing than the PLP parameterization. Experiments also proved that 5 PLP-cepstral (including derived 5 delta + 5 delta-delta) coefficients do not afford the best results as could be deduced from recent work. However, after optimal setting both parameterization techniques provided almost comparable results

Volume 3, page 1813 Session D25

N-best List Generation using Word and Phoneme Recognition Fusion

Pusateri E[1], Van Thong J-M[2]
[1]MIT Spoken Language Group, USA, [2]Compaq Cambridge Research Laboratory, USA

This paper describes an approach for combining phoneme and word recognition to produce an accurate N-best list of hypotheses. We run two decoding threads in parallel. The first performs phoneme recognition, while the other performs word recognition on the same recorded utterance. The output of the word recognition thread is returned as the most likely hypothesis, and the result of the phoneme recognition thread is used to lookup a list of words for the rest of the N-best list. The algorithm is simple to implement and efficient. In our evaluation, we found that this approach has similar performance to the classical lattice-based N-best search methods on isolated word recognition. This method has the potential to improve existing ASR systems or can be used in interactive multi-modal applications.

Volume 3, page 1817 Session D25

A One Pass Semi-dynamic Network Decoder based on Language Model Network

Ahn D-H, Chung M
Sogang University, Korea

Decoding in a precompiled static network, compared with one in a dynamically managed network, is easier to implement and faster enough to yield a near real time response. However, when the recognition system handles a complex task, it has a problem of intensive memory usage. To overcome this weakness, we present a new decoding strategy that combines the advantages of static and dynamic network architectures. In this strategy, we first define a language model (LM) network that can represent an arbitrary back-off N-gram in a finite state network (FSN). The LM network enables constructing a precompiled static network and partitioning the whole network into subnetworks using LM histories. Then the recognition network can be dynamically created and destroyed on the subnetwork¡¯s basis. To make dynamic management of networks as simple as possible, we also devise a data structure for network representation that self-structures its nodes and arcs. The final decoder maintains subnetworks as needed, but does not need to maintain nodes and arcs. Experimental results show that this semi-dynamic management of networks dramatically reduces memory usage at the cost of less than 10% increase of recognition time.

Volume 3, page 1821 Session D25

Improving Automatic Speech Recognition Using Tangent Distance

Macherey W, Keysers D, Dahmen J, Ney H
RWTH Aachen, University of Technology, Germany

In this paper we present a new approach to variance modelling in automatic speech recognition (ASR) that is based on tangent distance (TD). Using TD, classifiers can be made invariant w.r.t. small transformations of the data. Such transformations generate a manifold in a high dimensional feature space when applied to an observation vector. While conventional classifiers determine the distance between an observation and a prototype vector, TD approximates the minimum distance between their manifolds, resulting in classification that is invariant w.r.t. the underlying transformation. Recently, this approach was successfully applied in image object recognition. In this paper we describe how TD can be incorporated into ASR systems based on Gaussian mixture densities (GMD). The proposed method is embedded into a probabilistic framework. Experiments on the SieTill corpus for telephone line recorded digit strings show a significant improvement in comparison with a conventional GMD approach using comparable amounts of model parameters.

Volume 3, page 1825 Session D25

N-best Speech Hypotheses Reordering Using Linear Regression

Chotimongkol A, Rudnicky A
Carnegie Mellon University, USA

We propose a hypothesis reordering technique to improve speech recognition accuracy in a dialog system. For such systems, additional information external to the decoding process itself is available, in particular features derived from the parse and the dialog. Such features can be combined with recognizer features by means of a linear regression model to predict the most likely entry in the hypothesis list. We introduce the use of concept error rate as an alternative accuracy measurement and compare it withy the use of word error rate. The proposed model performs better than human subjects performing the same hypothesis reordering task.

Volume 3, page 1829 Session D25

Low-Resource Hidden Markov Model Speech Recognition

Deligne S, Eide E, Gopinath R, Kanevsky D, Maison B, Olsen P, Printz H, Sedivy J
IBM, USA

We describe techniques for enhancing the accuracy, efficiency and features of a low-resource, medium-vocabulary, grammar-based speech recognition system, which uses hidden Markov models. Among the issues and techniques we explore are reducing computation via silence detection, applying the Bayesian information criterion (BIC) to build smaller and better acoustic models, minimizing finite state grammars, using hybrid maximum likelihood and discriminative models, and automatically generating baseforms from single new-word utterances. We report WER figures where appropriate.

Volume 3, page 1833 Session D25

Speech recognition at multiple sampling rates

Hirsch H-G[1], Hellwig K[2], Dobler S[2]
[1]University of Applied Sciences Niederrhein, Germany, [2]Ericsson Eurolab, Spain

A feature extraction scheme is presented that analyzes speech signals sampled at different sampling rates. This will be needed in the future because of terminals in the telecom network that will transmit speech information also in the frequency region above 4 kHz. A cepstral analysis scheme is applied in the frequency range up to 4 kHz to create a common set of acoustic parameters for all sampling rates. Additional parameters are determined describing the subband energy in the frequency region above 4 kHz. As the major advantage of this feature extraction no individual recognizer has to be trained for each sampling frequency. It is shown with a recognition experiment that terminals and recognition systems can be combined without a remarkable loss in recognition performance with the terminal operating at a different sampling frequency than the recognizer has been trained on.

Volume 3, page 1837 Session D25

Support Vector Machine with Dynamic Time-Alignment Kernel for Speech Recognition

Shimodaira H[1], Noma K-I[1], Nakai M[1], Sagayama S[2]
[1]Japan Advanced Institute of Science and Technology, Japan, [2]University of Tokyo, Japan

A new class of Support Vector Machine (SVM) which is applicable to sequential-pattern recognition is developed by incorporating an idea of non-linear time alignment into the kernel. Since time-alignment operation of sequential pattern is embedded in the kernel evaluation, same algorithms with the original SVM for training and classification can be employed without modifications. Furthermore, frame-wise evaluation of kernel in the proposed SVM (DTAK-SVM) enables frame-synchronous recognition of sequential pattern, which is suitable for continuous speech recognition. Preliminary experiments of speaker-dependent 6 voiced-consonants recognition demonstrated excellent recognition performance of more than 98% in correct classification rate, whereas 93% by hidden Markov models (HMMs).

Volume 3, page 1841 Session D25

Efficient Scalable Speech Compression for Scalable Speech Recognition

Srinivasamurthy N, Ortega A, Narayanan S
University of Southern California, USA

We propose a scalable recognition system for reducing recognition complexity. Scalable recognition can be combined with scalable compression in a distributed speech recognition (DSR) application to reduce both the computational load and the bandwidth requirement at the server. A low complexity pre-processor is used to eliminate the unlikely classes so that the complex recognizer can use the reduced subset of classes to recognize the unknown utterance. It is shown that by using our system it is fairly straightforward to trade-off reductions in complexity for performance degradation. Results of preliminary experiments using the TI-46 word digit database show that the proposed scalable approach can provide a 40% speed up, while operating under 1.05 kbps, compared to the baseline recognition using uncompressed speech.

Volume 3, page 1845 Session D25

Session D26 - Poster
Thursday - 11.10 - 12.30

Signal Analysis: Speech Enhancement and Noise Processing

Chair: Unto K. Laine, HUT, Finland

Voice Activity Detection in Noisy Environments

Stadermann J[1], Stahl V[2], Rose G[2]
[1]University of Duisburg, Germany, [2]Philips Research Laboratories, Germany

The subject of this paper is robust voice activity detection (VAD) in noisy environments, especially in car environments. We present a comparison between several frame based VAD feature extraction algorithms in combination with different classifiers. Experiments are carried out under equal test conditions using clean speech, clean speech with added car noise and speech recorded in car environments. The lowest error rate is achieved applying features based on a likelihood ratio test which assumes normal distribution of speech and noise and a perceptron classifier. We propose modifications of this algorithm which reduce the frame error rate by approximately 30% relative in our experiments compared to the original algorithm.

Volume 3, page 1851 Session D26

An Improved Wavelet-Based Speech Enhancement System

Sheikhzadeh H[1], Abutalebi H R[2]
[1]Amirkabir University of Technology, Iran / DspFactory Ltd., Canada, [2]Amirkabir University of Technology, Iran

The problem of speech enhancement using wavelet thresholding algorithm is considered. Major problems in applying the basic algorithm are discussed and modifications are proposed to improve the method. First, we propose the use of different thresholds for different wavelet bands. Next, by employing a pause detection algorithm, noise profile is estimated and the thresholds are adapted. This enables the modified enhancement system to handle colored and non-stationary noises. Finally, a wavelet-based voiced/unvoiced classification is proposed and implemented that can further improve the performance of the enhancement system. To evaluate the system performance, we have used real-life noise types such as multi-talker babble and low-pass noises. Subjective and objective evaluations show that the proposed system improves the performance the wavelet thresholding algorithm.

Volume 3, page 1855 Session D26

Enhancing Distributed Speech Recognition with Back-End Speech Reconstruction

Ramabadran T, Meunier J, Jasiuk M, Kushner B
Motorola, USA

In this paper, we present a method to enhance the usefulness of a Distributed Speech Recognition (DSR) system by providing it the capability to reconstruct speech at the back-end. Speech reconstruction is achieved using the standard DSR parameters, viz., Mel-Frequency Cepstral Coefficients (MFCC) and log-energy, and some additional parameters, viz., voicing class, pitch period, and (optionally) higher-resolution energy information. From the MFCC parameters and energy information, the spectral magnitudes at the harmonics of the pitch frequency are estimated. Based on the class information, the harmonic phases are appropriately modeled. The harmonic magnitudes and phases are used to reconstruct speech according to the well-known sinusoidal model for speech synthesis. Transmission of the additional parameters for speech reconstruction increases the DSR bit rate by less than 20%. Evaluation by Mean-Opinion-Score (MOS) test and Diagnostic Rhyme Test (DRT) show that speech reconstructed as above is of reasonable quality and quite intelligible.

Volume 3, page 1859 Session D26

Implementation Effective One-Channel Noise Reduction System

Tihelka J, Sovka P
FEE CTU in Prague, Czech Republic

This contribution addresses the problem of additive noise reduction using one-channel noise suppression system. A new implementation effective method is suggested and evaluated. The method consists of two independent parts. The noise estimation part is based on the noise matched filter producing an estimation of background noise without the need of a voice activity detector. The noise reduction part uses short-time spectral attenuation technique. The main idea reducing computational costs lies in the use of reduced number of frequency bands for computation of attenuation factors. Except of reduced number of operations this approach decreases fluctuations of estimated spectral gains, and therefore the speech distortion is low. Thus the suggested system eliminates the need of any enhanced speech postprocessing. A new effective approach is used for the inverse frequency transformation. In spite of the simplicity of the suggested method its performance is comparable with other existing one-channel noise reduction methods.

Volume 3, page 1863 Session D26

Efficient Speech Enhancement by Diffusive Gain Factors (DGF)

Kim H-G[1], Obermayer K[1], Bode M[2], Ruwisch D[2]
[1]Technical University of Berlin, Germany, [2]Cortologic, Germany

In this paper we propose a very simple but highly effective algorithm for single channel noise reduction of speech signals. One of the main objectives is to find a balanced tradeoff between noise reduction and speech distortion in the processed signal. This is accomplished by a system based on spectral minimum detection and diffusive gain factors. Our approach to speech enhancement is capable of distinguishing between speech and noise interference in the microphone signal, even when they are located in the same frequency band.

Volume 3, page 1867 Session D26

Correction of the Voice Timbre Distortions on Telephone Network

Mahé G, Gilloire A
France Télécom R&D / DIH / IPS, France

In a telephone link, the voice timbre is affected by the loss of low frequencies components and distortions due to the analog lines. We analyze first how the quantization noise limits the restoration of the timbre. Within this limitation, a method of equalization, inspired by the cepstral subtraction, is then proposed to correct the timbre and is validated by experimental results.

Volume 3, page 1871 Session D26

Speech Enhancement based on IMM with NPHMM

Lee Y[1], Lee J[2], Lee K Y[1], Shirai K[2]
[1]Soongsil University, Korea, [2]Waseda Univ., Japan

The nonlinear speech enhancement method with interactive parallel-extended Kalman filter is applied to speech contaminated by additive white noise. To represent the nonlinear and nonstationary nature of speech, we assume that speech is the output of a nonlinear prediction HMM (NPHMM) combining both neural network and HMM. The NPHMM is a nonlinear autoregressive process whose time-varying parameters are controlled by a hidden Markov chain. The simulation

results shows that the proposed method offers better performance gains relative to the previous results [6] with slightly increased complexity.

Volume 3, page 1875 Session D26

Speech Recognition under Musical Environments Using Kalman Filter and Iterative MLLR Adaptation

Fujimoto M, Ariki Y
Ryukoku University, Japan

In this paper, we propose a speech recognition method under non-stationary musical environments using Kalman filtering speech signal estimation method and iterative unsupervised MLLR adaptation. Our proposing method estimates the speech signal under non-stationary noisy environments such as musical background by applying speech state transition model to Kalman filtering estimation. The speech state transition model represents the state transition of speech component in non-stationary noisy speech and is modeled by using Taylor expansion. In this model, the state transition of noise is estimated by using linear predictive estimation. Furthermore, to obtain higher recognition accuracy, we consider to adapt the acoustic models by using iterative unsupervised MLLR adaptation to speech spectra distorted by Kalman filtering residual noise. In order to evaluate the proposed method, we carried out large vocabulary continuous speech recognition experiments under 3 types of music. As a result, the proposed method obtained the significant improvement in word accuracy.

Volume 3, page 1879 Session D26

Dual Channel Speech Enhancement using Coherence Function and MDL-based Subspace Approach in Bark Domain

Vetter R, Renevey P, Krauss J
CSEM, Switzerland

A novel algorithm for dual channel speech enhancement is presented. It combines the coherence function and a subspace approach in the Bark domain together with an optimal subspace selection through the minimum description length (MDL) criterion. The coherence function allows one to exploit the spatial diversity of the sound field. The processing in the Bark domain permits to take into account of masking properties of the human auditory system while the MDL-based subspace approach ensures statistical robustness. Performance evaluation in real sound fields has highlighted the ability of the algorithm to enhance noisy signals and improve intelligibility for various experimental conditions.

Volume 3, page 1883 Session D26

Entropy Based Voice Activity Detection in Very Noisy Conditions

Renevey P[1], Drygajlo A[2]
[1]CSEM, Switzerland, [2]EPFL, Switzerland

This paper addresses the problem of robust voice activity detection (VAD) capable for working at very low signal-to-noise ratios (SNR<10dB). A new algorithm that we propose is based on entropy estimation measures of the time-frequency magnitude spectrum. The problem of the estimation of the distribution of noise in detected non-speech segments of analysed signal is also presented. It is shown that the new entropy based VAD significantly outperforms the commonly used energy-based algorithms in all (stationary, non-stationary, white and coloured) noise conditions at SNRs from 10 dB down to -10 dB and below. One of the main advantages of the method proposed in this paper is that it is not very sensitive to the changing level of noise.

Volume 3, page 1887 Session D26

Discrimination between speech and music based on a low frequency modulation feature

Karnebäck S
KTH, Sweden

The possibility to discriminate between speech and music signals by using a feature based on low frequency modulation has been investigated. Three different low frequency modulation parameters have been extracted and tested concerning the ability of discrimination. The low frequency modulation amplitudes calculated over 20 critical bands and their standard deviations were found to be good features for this discrimination task even with VQ models. They were also found to be less sensitive to channel quality and model size than MFCC features.

Volume 3, page 1891 Session D26

Credibility Proof for Speech Content and Speaker Verification by Fragile Watermarking with Consecutive Frame-Based Processing

Cheng Y-W, Lee L-S
National Taiwan University, Taiwan

With rapid growth in real-world speech-based transactions via communication networks, the need for a reliable mechanism to prove the credibility of speech content is highly desired. This paper presents a fragile watermarking technique for such purposes. The proposed approach is also useful for speaker verification. It breaks the speech signal into a series of nonoverlapping frames and encodes the watermark into those frames consecutively. The watermark sequence for each frame is dependent on the statistical characteristics of the previous frame, therefore any signal discontinuity caused by malicious purposes will be detected and the speaker can be verified as well by the watermark. Experiments showed very encouraging results, including reasonable detection rate even under signal compression and filter attacks.

Volume 3, page 1895 Session D26

Map Estimation for On-line Noise Compensation of Time Trajectories of Spectral Coefficients

Potamitis I, Fakotakis N, Kokkinakis G
University of Patras, Greece

This paper presents a novel data driven compensation technique that modifies on-line the incoming spectral representation of degraded speech in order to approximate the features of high quality speech used to train a classifier. We apply the Bayesian inference framework to the degraded spectral coefficients based on the modeling of clean speech linear-spectrum with appropriate non-Gaussian distributions that allow maximum a-posteriori (MAP) closed form solution. The MAP solution leads to spectral magnitude estimation adapted to the spectral characteristics and noise variance of each spectral band. We perform extensive evaluation of our algorithm using white and coloured Gaussian noise on the task of improving the quality of speech perception as well as Automatic Speech Recognition (ASR), and demonstrate its robustness at very low SNRs. The enhancement process comes at little to no extra computational overhead for ASR systems, thus achieving real time performance.

Volume 3, page 1899 Session D26

A New Method for Speech Denoising and Robust Speech Recognition Using Probabilistic Models for Clean Speech and for Noise

Attias H, Deng L, Acero A, Platt J C
Microsoft Research, USA

We present a new method for speech denoising and robust speech recognition. Using the framework of probabilistic models allows us to integrate detailed speech models and models of realistic non-stationary noise signals in a principled manner. The framework transforms the denoising problem into a problem of Bayes-optimal signal estimation,

producing minimum mean square error estimators of desired features of clean speech from noisy data. We describe a fast and efficient implementation of an algorithm that computes these estimators. The effectiveness of this algorithm is demonstrated in robust speech recognition experiments, using the Wall Street Journal speech corpus and Microsoft Whisper large-vocabulary continuous speech recognizer. Results show significantly lower word error rates than those under noisy-matched condition. In particular, when the denoising algorithm is applied to the noisy training data and subsequently the recognizer is retrained, very low error rates are obtained.

Volume 3, page 1903 Session D26

Session D32 - Oral
Thursday - 14.00 - 15.20

Speech Synthesis: Grapheme-to-Phoneme Conversion

Chair: To be decided,

Designing very compact decision trees for grapheme-to-phoneme transcription

Kienappel A K[1], Kneser R[2]
[1]Philips Research Laboratories, Germany, [2]Philips Speech Processing, Germany

Decision trees are a popular technique for automatic generation of a phonetic transcription for a given word spelling. We investigate different methods of decision tree design to obtain more compact trees and at the same time better grapheme-to-phoneme transcription quality. We evaluate different approaches to decision tree question selection and pruning using one English and two German grapheme-to-phoneme transcription tasks. In particular, we present a method of automatic generation of decision tree questions from the training data that significantly improves decision tree design.

Volume 3, page 1911 Session D32

Using Machine Learning Techniques for Grapheme to Phoneme Transcription

Mana F, Massimino P, Pacchiotti A
Loquendo, Vocal Technology and Services, Italy

The renewed interest in grapheme to phoneme conversion (G2P), due to the need of developing multilingual speech synthesizers and recognizers, suggests new approaches more efficient than the traditional rule&exception ones. A number of studies have been performed to investigate the possible use of machine learning techniques to extract phonetic knowledge in a automatic way starting from a lexicon. In this paper, we present the results of our experiments in this research field. Starting from the state of art, our contribution is in the development of a language-independent learning scheme for G2P based on Classification and Regression Trees (CART). To validate our approach, we realized G2P converters for the following languages: British English, American English, French and Brazilian Portuguese.

Session D31 - Demonstrations
Thursday - 14.00 - 15.20

ESE6 - Education Arena

Chair: Gerrit Bloothooft, University of Utrecht, The Netherlands

The Education Arena session during Eurospeech 2001- Scandinavia will showcase Computer-based Educational Materials in phonetics, speech technology and allied areas. It follows the success of a similar event mounted by the Socrates Thematic Network in Phonetics and Speech Communication at Eurospeech 1999 in Budapest. The Education Arena is organised by the ISCA Education Special Interest Group (EduSIG), with the collaboration of ELSNET. As in 1999, a CD-ROM of materials and demonstrations will be produced and handed out to attendees for free.

Volume 3, page 1908 Session D31

Volume 3, page 1915 Session D32

Knowledge of Language Origin Improves Pronunciation Accuracy of Proper Names

Llitjós A F, Black A W
Carnegie Mellon Univ., USA

As it is impossible to have a lexicon with complete coverage, and a high proportion of unknown words are proper names, this paper addresses the issue of automatically finding pronunciations of unseen proper names in US English. Proper names, especially in the US, may come from a large range of ethnic backgrounds. We present a model and results showing that including ethnic origin of words in a statistical model can improve pronunciation results. We used a lexicon of 56,000 proper names from CMUDICT, and gathered data (text and proper names) from 26 languages to built statistical models that provide an estimate of word origin. Tests against held out data showed a 7.6% absolute improvement from a baseline of 54.8% when language based features were added to our CART-based model. Our user studies show a 17% preference for the model with language features compared to the baseline.

Volume 3, page 1919 Session D32

On the pronunciation of acronyms in French and in Italian

Boula de Mareüil P[1], Floricic F[2]
[1]Elan TTS, France, [2]ERSS, France

Acronyms are used more and more nowadays: this article describes their pronunciation in French and in Italian. Rules are proposed, so that a text-to-speech converter may know whether it must spell or read acronyms, and which way. Our analysis, which gave rise to a computational realisation, leans on the number of letters which constitute acronyms, on the vowel/consonant opposition, as well as on the distribution between continuous and momentary consonants.

Volume 3, page 1923 Session D32

Session D33 - Oral
Thursday - 14.00 - 15.20

Signal Analysis: Speech Enhancement

Chair: Chang D. Yoo, KAIST, Korea

Enhancement of Noisy Speech by Using Improved Global Soft Decision

Shin V, Kim D-S, Kim M Y, Kim J
Samsung AIT, Korea

We propose a novel speech enhancement algorithm, termed improved global soft decision (IGSD). IGSD is a unified framework for global soft decision on speech absence/presence, noise spectrum estimation, spectral gain modification based on Ephraim-Malah noise suppression. In IGSD, speech absence probability (SAP) is the most important factor, and we propose an efficient and novel SAP estimation in which the SAP is derived based on the general hypothesis for speech absence/presence. In IGSD, the global SAP based on the global hypothesis for speech absence/presence is used to prevent from the problem caused by insufficient amount of data, but more general hypothesis is utilized in the derivation of global SAP estimation. The performance of IGSD is evaluated both subjectively and objectively, and the quality of speech is improved significantly, compared with conventional GSD speech enhancement algorithm.

Volume 3, page 1929 Session D33

Enhancement of Speech Using Bark-Scaled Wavelet Packet Decomposition

Cohen I
Lamar Signal Processing Ltd., Israel

In this paper, we propose a speech enhancement system, which integrates a bark-scaled wavelet packet decomposition (BS-WPD), a soft-decision gain modification and a "magnitude" decision-directed estimation technique. The BS-WPD provides an overcomplete auditory representation, having a higher frequency resolution than the critical band decomposition. Speech is estimated by Wiener filtering in the wavelet packet domain, modified by the signal presence probability. We introduce a "magnitude" decision-directed estimator for the variance of speech, which is closely related to the decision-directed estimator of Ephraim and Malah. This estimator achieves, in the established process, a better tradeoff between noise reduction and signal distortion. The proposed enhancement algorithm is tested with various noise types, and compared to a conventional log-spectral amplitude estimator. We show that noise can be further suppressed, while preserving its natural structure and the intelligibility and quality of the speech components.

Volume 3, page 1933 Session D33

A New Approach for Wavelet Speech Enhancement

Bahoura M, Rouat J
Université du Québec à Chicoutimi, Canada

We propose a new approach to improve the performance of speech enhancement techniques based on wavelet thresholding. First, space-adaptation of the threshold is obtained by extending the principle of the level-dependent threshold to the Wavelet Packet Transform (WPT). Next, the time-adaptation is introduced using the Teager Energy Operator (TEO) of the wavelets coefficients. Finally, the time-space adapted threshold is proposed. Comparisons with the Ephraim and Malah Filter are reported.

Volume 3, page 1937 Session D33

Speech/Noise-Dominant Decision for Speech Enhancement

Yoon S, Yoo C D
Korea Advanced Institute of Science and Technology (KAIST), Korea

A novel method to reduce additive non-stationary noise is proposed. The proposed method requires neither the statistical assumption about noise nor the estimate of the noise statistics from any pause regions. The enhancement is performed on a band-by-band basis for each time frame. Based on both the decision on whether a particular band in a frame is speech or noise dominant and the masking property of the human auditory system, an appropriate amount of noise is reduced using modified spectral subtraction. The proposed method was tested on various noisy conditions - car noise, F16 noise, white Gaussian noise, pink noise, tank noise and babble noise. On the basis of comparing segmental SNR with spectral subtraction proposed by Boll with pause detection for estimating noise, and visually inspecting the enhanced spectrograms and listening to the enhanced speech, the proposed method was found to effectively reduce various noise while minimizing distortion to speech.

Volume 3, page 1941 Session D33

Session D34 - Oral
Thursday - 14.00 - 15.20

Speech Recognition and Understanding: Discriminative Training

Chair: Phil Woodland, Cambridge Univ., UK

An MCE based Classification Tree Using Hierarchical Feature-Weighting in Speech Recognition

Wang F, Zheng F, Wu W
Tsinghua University, P. R. China

In this paper a hierarchical classification framework using the feature-weighting tree for the objective of applying diverse weighting to acoustic features is proposed for speech recognition. The hierarchical feature-weighting tree with a flexible structure complexity can be constructed optimally with the optimal splitting for the recognition confusion graph. Based on the minimum classification error principle, the subset-dependent training and the multi-level recognition method are proposed, where the feature weighting can be automatically trained without normalization in recognition. Both the mathematical analysis and the experimental results show that such a supervised hierarchical classification tree based on the feature weighting is efficient to reduce the speech recognition error.

Volume 3, page 1947 Session D34

Selective MCE Training Strategy in Mandarin Speech Recognition

Zhou J, Chang E, Chao H
Microsoft Research China, P. R. China

In this paper, selective strategy about MCE based discriminative training method,in particular for mandarin syllable loop recognition,is introduced. The basic idea is that since the decoding errors occur in parts of the models in whole decoded sentence, it is reasonable to adjust the parameters of the "wrong models". As a result, weighted MCE formulation is derived, which can provide more effective convergence property and about 10% error rate reduction for a large training set is achieved. On the other hand, from our experiments, we observed that although the whole performance of recognition system is improved, some original correct recognition results are misrecognized after discriminative training, divide and conquer strategy is proposed to solve it. Combining above two methods, we got more than 14.5% error reduction in syllable loop recognition experiments.

Volume 3, page 1951 Session D34

Discriminative Disfluency Modeling for Spontaneous Speech Recognition

Wu C-H, Yan G-L
National Cheng Kung University, Taiwan, ROC

Most automatic speech recognizers (ASRs) have concentrated on read speech, which is different from speech with the presence of disfluencies. These ASRs cannot handle the speech with a high rate of disfluencies such as filled pauses, repetition, repairs, false starts, and silence pauses in actual spontaneous speech or dialogues. In this paper, we focus on the modeling of the filled pauses ¡§uh¡¨ and ¡§um.¡¨ The filled pauses contain the characteristics of nasal and lengthening, and the acoustic parameters for these characteristics are analyzed and adopted for disfluency modeling. A Gaussian mixture model (GMM), trained by a discriminative training algorithm that minimizes the recognition error, is proposed. A transition probability density function is defined from the GMM and used to weight the transition probability between the boundaries of fluency and disfluency models in the one-stage algorithm. Experimental result shows that the proposed method yields an improvement rate of 27.3% for disfluency compared to the baseline system.

Volume 3, page 1955 Session D34

Comparative Analysis for Data-Driven Temporal Filters Obtained Via Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) In Speech Recognition

Hung J-W[1], Wang H-M[1], Lee L-S[2]
[1]Institute of Information Science, Academia Sinica, Taiwan, [2]National Taiwan University,Taiwan

The Linear Discriminant Analysis (LDA) has been widely used to derive the data-driven temporal filtering of speech feature vectors. In this paper, we proposed that the Principal Component Analysis (PCA) can also be used in the optimization process just as LDA to obtain the temporal filters, and detailed comparative analysis between these two approaches are presented and discussed. It's found that the PCA-derived temporal filters significantly improve the recognition performance of the original MFCC features as LDA-derived filters do. Also, while PCA/LDA filters are combined with the conventional temporal filters, RASTA or CMS, the recognition performance will be further improved regardless the training and testing environments are matched or mismatched, compressed or noise corrupted.

Volume 3, page 1959 Session D34

Session D35 - Poster
Thursday - 14.00 - 15.20

Speech Coding: Advances in Speech Coding

Chair: Gernot Kubin, Graz University of Technology, Austria

Coding method for successive pitch periods

Heikkinen A, Ruoppila V T, Pietilä S
Nokia Research Center, Finland

This paper presents a coding method for successive pitch periods of a speech signal. In the proposed method, a priori knowledge of the statistical properties of successive pitch periods is used by designing a shaped lattice structure that covers the most probable points in the pitch space. We also present briefly a pitch search algorithm for the shaped lattice. An example implementation of shaped lattice coding is given for a modified IS-641 speech coder. Based on the simulation results, the proposed scheme achieves capacity savings compared to the conventional methods

Volume 3, page 1965 Session D35

Objective Evaluation of Methods for Quantization of Variable-Dimension Spectral Vectors in WI Speech Coding

Nurminen J[1], Heikkinen A[2], Saarinen J[1]
[1]Tampere University of Technology, Finland, [2]Nokia Research Center, Finland

In this paper, we present a comprehensive evaluation of five quantization techniques for variable-dimension spectral vectors in a waveform interpolation speech coder. Each technique included in the evaluation is based on dimension conversions. The conversions are performed using zero-pad and truncation, frequency bins, band-limited interpolation, discrete cosine transform, and polynomial approximation. In addition to assessing quantization accuracy, this study considers the complexity of the techniques. The evaluation indicates that the selection of the optimal quantization technique is a trade-off between coding accuracy, complexity, and memory requirements. According to our results, the technique based on discrete cosine transform appears to be a strong candidate for many applications.

Volume 3, page 1969 Session D35

Squared Error as a Measure of Phase Distortion

Pobloth H, Kleijn W B
KTH, Sweden

In this article, we investigate how accurately the squared error captures perceptual errors introduced by Fourier phase spectrum changes. We measure the perceptual error using the Auditory Image Model by Patterson et al.. The squared error is found to represent the perceptual error well for low squared errors but it saturates. Thus, a further increase in squared error does on average not lead to any further increase in perceptual error. This suggests that encoding phase using squared-error trained codebooks only improves perceived quality when operating at high bit rates. To verify this, phase was encoded with codebooks of different sizes. As expected, increasing the codebook size has very little influence on the average perceptual error for low rates, which is confirmed by listening tests. Our results suggest that a direct phase codebook is an inefficient representation of the relevant information contained in phase.

Volume 3, page 1973 Session D35

Non-Linear Predictive Vector Quantization of speech

Faundez-Zanuy M
Escola Universitaria Politecnica de Mataro, Spain

In this paper we propose a Non-Linear Predictive Vector quantizer (PVQ) for speech coding, based on Multi-Layer Perceptrons. We also propose a method to evaluate if a quantizer is well designed, and if it exploits the correlation between consecutive outputs. Although the results of the Non-linear PVQ do not improve the results of the non-linear scalar predictor, we check that there is some room for the PVQ improvement.

Volume 3, page 1977 Session D35

A Variable Rate Hybrid Coder Based on a Synchronized Harmonic Excitation

Katugampala N, Kondoz A
University of Surrey, UK

A novel synchronization technique is proposed for hybrid coders employing harmonic and waveform coding. A new classification technique based on analysis by synthesis to distinguish between stationary and transitional segments is also proposed. Harmonic excitation is synchronized with the LPC residual by transmitting the location of the pitch pulse closest to the frame boundary and a phase value that represents the shape of the corresponding pitch pulse. A hybrid coder is designed to demonstrate the new techniques, which has three modes: scaled white noise excitation colored by LPC for unvoiced, ACELP for transitions, and harmonic excitation for stationary segments. Subjective listening tests show that the speech quality of the variable rate hybrid coder outperforms the quality of ITU G.723.1 coders, at maximum bit rates lower than those of G.723.1 coders.

Volume 3, page 1981 Session D35

A Hybrid Sub-Band Sinusoidal Coding Scheme

Ho M-S, Molyneux D J, Cheetham B M G
University of Manchester, UK

This paper describes a hybrid sub-band speech coding scheme based on sinusoidal coding and CELP. Purely voiced speech is encoded using sinusoidal coding techniques and phase information is selectively transmitted. For mixed and unvoiced speech, the lower band is processed by sinusoidal coding algorithms while the upper band is encoded using CELP. To accommodate the extra bandwidth required by the encoded CELP parameters, the phase information is disregarded. The proposed coder is enhanced by sub-band discrete all-pole modeling and a voicing detection technique based on an analysis-by-synthesis approach. An efficient adaptive spectral shaping technique based on bandwidth widening in the LSP domain is employed. The proposed technique is capable of producing high quality speech at 4.1 kbit/s.

Volume 3, page 1985 Session D35

Low rate speech coding incorporating Simultaneously Masked Spectrally Weighted Linear Prediction

Lukasiak J, Burnett I S, Ritz C H
University of Wollonggong, Australia

Linear prediction (LP) is the cornerstone of most modern speech compression algorithms. Previously it has been shown that incorporating a weighting function based on the simultaneous masking property of the ear into the calculation of the LP coefficients (SMWLPC) allows the filter to better model the unmasked sections of the input spectrum. This paper conducts a detailed analysis of the implementation of SMWLPC in low rate speech codecs. The analysis allows the cause of inconsistencies in the technique to be identified and solutions formulated. Experimental results show that when combined with the proposed changes, the SMWLPC technique is suitable for implementation in any low rate LP based speech codec and the net result

is an improvement in the perceptual quality of synthesised speech for all speakers.

Volume 3, page 1989 Session D35

Narrowband Perceptual Audio Coding: Enhancements for Speech

Najaf-Zadeh H, Kabal P
McGill University, Canada

This paper presents a bi-modal coding paradigm to compress narrowband audio signals at 8 kbit/s. In the general mode, the Enhanced Narrowband Audio Coder (ENPAC) exploits the characteristics of the human hearing system to adaptively code the perceptually important spectral components of the input audio. The other mode is employed to handle audio inputs with a strong harmonic structure. In that mode, the input block is represented by its audible harmonics. The spectral magnitude is modeled by the linear prediction analysis in the time domain. The phase of each harmonic is predicted and the phase residues are quantized using an adaptive bit allocation algorithm. This paper introduces a perceptually-based upper bound for phase errors of spectral components. The ENPAC encoder delivers good quality for narrowband speech and non-speech inputs.

Volume 3, page 1993 Session D35

Techniques for high-quality ACELP coding of wideband speech

Bessette B[1], Lefebvre R[1], Salami R[2], Jelinek M[1], Vainio J[3], Rotola-Pukkila J[3], Mikkola H[3], Jarvinen K[3]
[1]University of Sherbrooke, Canada, [2]VoiceAge Corporation, Canada, [3]Nokia Research Center, Finland

We present in this paper new methods for achieving high-quality wideband speech at low rates using the ACELP algorithm. Several innovations are introduced to optimize the quality and minimize the complexity of the coder. A multi-rate wideband speech encoding algorithm based on these techniques was recently selected by 3GPP as the standard for AMR-WB, and is currently one of the candidates for the ITU-T wideband speech coder standard at around 16 kbit/sec. This standard was jointly developed by VoiceAge and Nokia.

Volume 3, page 1997 Session D35

Wideband ACELP at 16 kb/s with Multi-band Excitation

Pujalte S, Moreno A
Universitat Politècnica de Catalunya, Spain

This paper describes two wideband CELP coders at 16 kb/s. Their main feature is fast searching, achieving quality comparable to G.722 at 56 and 48 kb/s. Both the excitations derive from a multi-band algebraic structure in order to reduce computational complexity and bit allocation. The first one, the faster, has a full band excitation with two gains. The second one, which provides better quality, has a full band excitation with emphasized high frequencies. Analysis and error minimization are done over the full band.

Volume 3, page 2001 Session D35

Wideband Speech Coding Algorithm with Application of Discrete Wavelet Transform to Upper Band

Lee S W, Bae K S
Kyungpook National University, Korea

In this paper, we propose a new wideband speech coder which combines the European standard of a narrowband speech coder, i.e., GSM-EFR, and a transform coder using the discrete wavelet transform. Input speech is first split into two bands with equal bandwidth. A subband coder with wavelet transformed speech is designed for a upper band coder, and a GSM-EFR coder is adopted as a lower band coder. The total bit rate of the proposed coder is 18.9 kbps, and informal listening test results have shown that the proposed coder has comparable speech quality to that of G.722 with 56 kbps.

Volume 3, page 2005 Session D35

A Switched DPCM/Subband Coder for Pre-echo Reduction

Satheesh S, Sreenivas T V
Indian Institute of Science, India

Recently, adaptive subband coders based on wavelet packet decomposition and psychoacoustic modelling have been proposed to achieve transparent quality compression of audio signals. While these coders perform well for stationary signals, there is no special mechanism in the coder to prevent the pre-echo artifact when transient signals are encoded. In this paper, we propose a switched DPCM/Subband structure to remove the pre-echo problem. This is achieved through a novel temporally varying bit allocation scheme which is based on the temporal masking properties of the human auditory system. The proposed coder/decoder output is found to be free from pre-echo artifact even at a lower bitrate than the adaptive subband coder.

Volume 3, page 2009 Session D35

A Generalized Multistage VQ Approach for Spectral Magnitude Quantization

Etemoglu C O, Cuperman V
University of California, Santa Barbara, USA

This paper presents a novel vector quantization (VQ) technique in which the quantized vector is formed by adding the transformed outputs of a multistage codebook rather than just adding the outputs of the stages as in regular multistage vector quantization (MSVQ). The transformations are selected from a family of linear transformations represented by a codebook of matrices. This technique can be viewed as a generalized form of MSVQ. If the transformations are constrained to be the identity transformation, this technique becomes identical to the regular MSVQ. The design algorithm is based on joint optimization of the linear transformations and the stage codebooks. It is shown that the proposed technique yields high quality spectral magnitude quantization with performance exceeding that of multistage vector quantization (MSVQ) of similar complexity and bit rate.

Volume 3, page 2013 Session D35

Efficient Implementation of ITU-T G.723.1 Speech Coder for Multichannel Voice Transmission and Storage

Jung S-K, Park Y-C, Yoon S-W, Kim K-T, Youn D-H
Yonsei University, Korea

This paper presents an efficient implementation of G.723.1 speech coder. To simplify the excitation quantization procedure which is the most computationally demanding, we propose fast algorithms for adaptive codebook and fixed codebook search. In the fast adaptive codebook search, pitch delay and pitch gains are computed sequentially. In the fast fixed codebook search, the codebook structure is redesigned based on the interleaved single-pulse permutation (ISPP) design at high rate mode and the depth-first tree search is applied instead of nested-loop search at low rate mode. A real-time implementation is achieved using a 16-bit fixed-point TMS320C62x DSP. The implemented G.723.1 speech coder requires 8.70 and 10.29 MHz clock cycles at low and high rate, respectively, 57.8 kByte of program memory and 55 kByte of data memory. Thus, more than 16 channels of G.723.1 coder can be operated in real-time using a single TMS320C62x DSP.

Volume 3, page 2017 Session D35

Session D36 - Poster
Thursday - 14.00 - 15.20

Resources, Assessment and Standards: Corpora

Chair: Harald Höge, Siemens, Germany

"CU-Move" : Analysis & Corpus Development for Interactive In-Vehicle Speech Systems

Hansen J H L, Angkititrakul P, Plucienkowski J, Gallant S, Yapanel U, Pellom B, Ward W, Cole R
Univ. of Colorado - Boulder, USA

In this paper, we present our recent work in the analysis and formulation of a new acoustic speech corpus for developing in-vehicle interactive systems for route planning and navigation. The CU-Move Corpus development is partitioned into two phases: [I] acoustic noise collection and analysis across vehicles, and [II] data collection consisting of +1000 speakers from across the United States. We present results from Phase I acoustic noise data analysis across vehicles to determine guidelines for Phase II large-scale data collection using a single vehicle type. A total of 14 noise conditions are identified for analysis across 6 vehicles. We also present our plan for Phase II collection including speakers, dialect regions, data collection hardware, prompts and dialog domains. Since previous studies in speech recognition have shown significant loses in performance when speakers are under task stress, it is important to develop conversational systems that minimize operator stress for the driver. This will be the first U.S. based corpus of its kind consisting of multi-channel data, intended for use in developing mixed-initiative dialog speech systems; the initial application being route planning and navigation through a wireless information retrieval sub-system connected to the WWW.

Volume 3, page 2023 Session D36

Multimedia Data Collection of In-Car Speech Communication

Kawaguchi N, Matsubara S, Takeda K, Itakura F
Nagoya University, Japan

This paper reports the details of the collection of @the multimedia data such as audio, video and auxiliary @information of the vehicle during a spoken dialogue in a @moving car. The system specially built in a Data CollectionVehicle (DCV) supports synchronous recording of @multi-channel audio data from 16 microphones, 3-channel video data and the vehicle related data. Multimedia data has been collected for three sessions of spoken dialogue in about a 60-minute drive by each of 200 subjects. Data has been collected for two dialogue modes:(1) prompted dialogue between the driver and an accompanying @operator and (2) natural dialogue between the @driver and a telephone operator for information access @over a cellular phone while driving a car. The corpus can @be used for analysis of multimedia data in a moving car @environment and also for modeling spoken dialogue in @scenarios such as information access while driving a car.

Volume 3, page 2027 Session D36

The U.S. SpeechDat-Car Data Collection

Heeman P A, Cole D, Cronk A
Oregon Graduate Institute, USA

The SpeechDat-Car data collection effort is an ambitious effort to collect data from multiple languages in an in-car setting. This paper describes the U.S. data collection effort. We discuss problems we had implementing the collection procedure; and changes we made to improve the procedure. This paper should benefit future in-car data collections.

Volume 3, page 2031 Session D36

Word Unit Based Multilingual Comparative Analysis of Text Corpora

Németh G, Zainkó C
Budapest University of Technology and Economics, Hungary

Parallel study of three linguistically different languages - Hungarian. German and English - using text corpora of a similar size gives a possibility for the exploration of both similarities and differences. Corpora of publicly available Internet sources was used. Besides traditional corpus coverage, word length and occurence statistics, some new features about prosodic boundaries (sentence beginning and final positions, preceding and following a comma) were also computed. Among others, it was found, that the coverage of corpora by the most frequent words follows a parallel logarithmic rule for all languages in the 40-85% coverage range. The functions are much nearer for English and German than for Hungarian. The results can be applied in such diverse domains as predictive text input, word hyphenation, language modeling in speech recognition, corpus-based speech synthesis, etc. Keywords: text corpora, corpus analysis, multilinguality, word length, sentence length, unit based analysis, language modeling, corpus-based speech synthesis

Volume 3, page 2035 Session D36

Creating a European English Broadcast News Transcription Corpus and System

Backfried G, Hecht R, Loots S, Pfannerer N, Riedler J, Schiefer C
Sail Labs, Austria

Based on BBN's Rough'n'Ready suite of technologies used in the DARPA Hub-4 evaluations we describe the Sail-Labs Media Indexer system aiming at processing European English television broadcasts. We discuss the development of a European English broadcast news corpus, suitable for measuring performance of system components, such as speaker identification and speech recognition. We further report evaluation results on our multi-purpose test set, and outline the integration of real-time indexing into a spoken document retrieval system.

Volume 3, page 2039 Session D36

The Nespole! VoIP Dialogue Database

Burger S[1], Besacier L[2], Coletti P[3], Metze F[4], Morel C[1]
[1]Carnegie Mellon University, USA, [2]Université Joseph Fourier, France, [3]Istituto Trentino di Cultura - Centro per la Ricerca Scientifica e Tecnologica, Italy, [4]University of Karlsruhe, Germany

This paper presents the status of the NESPOLE! data collection as of end of February, 2001. A multilingual VoIP (Voice over Internet Protocol networks) database consisting of 200 dialogues in 4 languages (English, German, Italian and French) was recorded and transcribed. Dialogue speakers were connected via a H323 video-conferencing terminal. We describe the task, the technical architecture, the recording procedure and the transcription process of the NESPOLE! data collection. We provide some statistics concerning the data and, finally, we address problems that arose during the collection and annotation process.

Volume 3, page 2043 Session D36

Design of Speech Corpus for Text-to-Speech Synthesis

Matousek J, Psutka J, Kruta J
University of West Bohemia in Pilsen, Czech Republic

This paper deals with the design of a speech corpus for a concatenation-based text-to-speech (TTS) synthesis. Several aspects of the design process are discussed here. We propose a sentence selection algorithm to choose sentences (from a large text corpus) which will be read and stored in a speech corpus. The selected sentences should include all possible triphones in a sufficient number of occurrences. Some notes on recording the speech are also discussed to ensure a quality speech corpus. As some popular speech synthesis techniques require knowing the moments of principal excitation of vocal tract during the speech, pitch-mark detection is also a subject of our attention. Several automatic pitch-mark detection methods are discussed here and a comparison test is performed to find out the best method.

Volume 3, page 2047 Session D36

The IFA Corpus: a Phonemically Segmented Dutch "Open Source" Speech Database

Van Son R J J H[1], Binnenpoorte D[2], Van den Heuvel H[2], Pols L C W[1]
[1]University of Amsterdam, the Netherlands, [2]Nijmegen University, the Netherlands

An open source database of hand-segmented Dutch speech was constructed with off-the-shelf software using speech from 8 speakers in a variety of speaking styles. For a total of 50,000 words, speech acquisition and preparation took around 3 person-weeks per speaker. Hand segmentation took 1,000 hours of labeling altogether. The asymptotic segmentation speed was about one word, or four boundaries, per minute. An evaluation showed that the Median Absolute Difference of the segment boundaries was 6 ms between labelers, and 4 ms within labelers. Label differences (substitutions, insertions, and deletions) were found in 8% of the segments between labelers and 5% within labelers. Compiled data are available in relational database format for querying with SQL.

Volume 3, page 2051 Session D36

African Speech Technology (AST) Telephone Speech Databases: Corpus Design and Contents

Louw P H[1], Roux J C[1], Botha E[2]
[1]University of Stellenbosch, South Africa, [2]University of Pretoria, South Africa

The African Speech Technology project is developing telephone speech databases for five of South Africa's eleven official languages, i.e. South African English, Afrikaans, Zulu, Xhosa, and Southern Sotho. These databases will be fully transcribed – orthographically and phonetically – and will be used for the training and testing of phoneme-based, speaker-independent speech recognition systems. The project aims to deliver a telephone speech application developer's software toolkit. A prototype multilingual enquiry and booking system for the hotel industry will be developed as a first application. This paper describes the design and contents of the speech corpus that is currently being collected over both mobile and fixed networks. In particular language coverage is discussed within the framework of the multilingual character of the South African population. Some language specific differences with regards to the contents of the different databases are noted. Methods and tools applied in the acquisition of phonetic information are discussed.

Volume 3, page 2055 Session D36

SpeechDat-E: Five Eastern European Speech Databases for Voice-Operated Teleservices Completed

Van den Heuvel H[1], Boudy J[2], Bakcsi Z[3], Cernocky J[4], Galunov V[5], Kochanina J[5], Majewski W[6], Pollak P[7], Rusko M[8], Sadowski J[9], Staroniewicz P[9], Tropf H S[10]
[1]University of Nijmegen, the Netherlands, [2]Lernout & Hauspie France, France, [3]Budapest University of Technology and Economics, Hungary, [4]Brno University of Technology, Czech Republic, [5]AudiTech, Ltd, St.Petersburg, Russia, [6]ITA, Wroclaw University of Technology, Poland, [7]Czech Technical University in Prague, Czech Republic, [8]Institute of Control Theory and Robotics, Slovak Academy of Sciences, Slovakia, [9]Wroclaw University of Technology, Poland, [10]Siemens, Germany

In the Speechdat-E project five medium large telephone speech databases have been collected for Czech, Hungarian, Polish, Russian, and Slovak. The project was recently concluded. This paper reports briefly on the contents of the databases, elaborates on experiences gained from the data recordings and from the validation of the databases. The availability of the databases to the public is addressed, too.

Volume 3, page 2059 Session D36

Concordancing for Parallel Spoken Language Corpora

Gibbon D[1], Trippel T[1], Sharoff S[2]
[1]Universität Bielefeld, Germany, [2]Russian Research Institut for Artificial Intelligence / Universität Bielefeld, Germany

Concordancing is one of the oldest corpus analysis tools, especially for written corpora. In NLP concordancing appears in training of speech-recognition system. Additionally, comparative studies of different languages result in parallel corpora. Concordancing for these corpora in a NLP context is a new approach. We propose to combine these fields of interest for a multi-purpose concordance for Spoken Language Data, opening the opportunity of combining corpus-linguistic and NLP methods resulting in a broader empirical basis for NLP research. Theoretic models for audio-concordances are discussed. Principles of the structure and design of a parallel audio concordance are given, coding by means of XML to ensure reusability and flexibility, using time stamps for referencing from annotations to the signal.

Volume 3, page 2063 Session D36

Large Broadcast News and Read Speech Corpora of Spoken Czech

Psutka J[1], Radova V[1], Muller L[1], Matousek J[1], Ircing P[1], Graff D[2]
[1]University of West Bohemia in Pilsen, Czech Republic, [2]University of Pennsylvania, USA

This paper presents the first annotated and phonetically transcribed large speech corpora developed for spoken Czech. All corpora were collected during the last two years at the Department of Cybernetics, University of West Bohemia (UWB) in Pilsen. The first two collections are broadcast news, the third corpus is a high-quality read-speech database. This paper describes the collection conditions, annotation and phonetic transcription process related to each corpus. The basic phonetic and lexical characteristics of all corpora will be given and compared mutually.

Volume 3, page 2067 Session D36

Development of Russian Lexical Databases, Corpora and Supporting Tools for Speech Products

Yablonsky S
St.-Petersburg State Transport University, Russia

The situation with regard to Russian language resources is fragmented and disorganized. For this reason, it is important to promote for Russian the development of its basic resources in one package that could be used for development of speech products. The paper presents a design of the Russian lexical databases, corpora and supporting tools (system for construction and support of lexical databases, system for transcription, morphological analyzer and normalyzer) developed for wide usage in speech engineering.

Volume 3, page 2071 Session D36

Constructing a segment database for Greek time domain speech synthesis

Fotinea S-E, Tambouratzis G, Carayannis G
Institute for Language and Speech Processing, Greece

In this article, a methodology is presented regarding the design of a segment database for use with a time-domain speech synthesis system for the Greek language. The main issue of this process is the systematic generation of a corpus containing all possible instances of the segments for the specific language. Particular issues such as the phonetic coverage, the sentence selection as well as iterative evaluation techniques employing custom-built tools are discussed. The resulting corpus is characterised by a near-minimal size, provides a complete coverage of the Greek language and its distribution of phonemes is similar to that of natural corpora. A typical spoken acquisition procedure may then be performed, resulting in a segment database for use with a time-domain Greek synthesizer. The corpus creation procedure allows for the fine-tuning of the segment database's language-dependent characteristics and thus assists in the generation of high-quality text-to-speech synthesis.

Volume 3, page 2075 Session D36

Session D41 - Demonstrations
Thursday - 15.50 - 17.30

ESE6 - Education Arena - Continued

Chair: Gerrit Bloothooft, University of Utrecht, The Netherlands

The Education Arena session during Eurospeech 2001- Scandinavia will showcase Computer-based Educational Materials in phonetics, speech technology and allied areas. It follows the success of a similar event mounted by the Socrates Thematic Network in Phonetics and Speech Communication at Eurospeech 1999 in Budapest. The Education Arena is organised by the ISCA Education Special Interest Group (EduSIG), with the collaboration of ELSNET. As in 1999, a CD-ROM of materials and demonstrations will be produced and handed out to attendees for free.

Volume 3, page 2080 Session D41

Session D42 - Oral
Thursday - 15.50 - 17.30

Resources, Assessment and Standards: Assessment Methodology

Chair: Herman Steeneken, TNO, The Netherlands

Subjective Assessment of Speech-System Interface Usability

Hone K[1], Graham R[2]
[1]Brunel University, UK, [2]Motorola, UK

Methods for the evaluation of the efficiency and effectiveness of speech input / output systems are well established. However, user subjective reactions to speech interfaces may well be a more important predictor of real world success. A review of existing subjective measures of speech system usability reveals a number of limitations in the approaches taken. This paper then summarises the work we have conducted in developing a new measure for the subjective assessment of speech system interfaces (SASSI).

Volume 3, page 2083 Session D42

An Objective Measure for Estimating MOS of Synthesized Speech

Chu M, Peng H
Microsoft Research China, P. R. China

This paper proposes an average concatenative cost function as the objective measure for naturalness of synthesized speech. All its seven component-costs can be derived directly from the input text and the scripts of speech database. A formal Mean Opinion Score (MOS) experiment shows that the average concatenative cost and its seven components are all highly correlated with MOS obtained subjectively. The correlation coefficient between the objective measure and subjective measure is –0.872. The mean of errors in MOS estimation for individual waveforms is 0.32 with 0.40 RMSE. When estimating the overall MOS for TTS systems, the mean error is smaller than 0.05. With the proposed objective measure, it becomes possible and easy for us to track the performance in naturalness regularly. The proposed cost function could also serve as criteria for optimizing the algorithms for unit selecting and speech database pruning.

Volume 3, page 2087 Session D42

Comparing the performance of two CSRs: How to determine the significance level of the differences

Strik H, Cucchiarini C, Kessens J
Univ. of Nijmegen, the Netherlands

When two CSRs are compared, it is important to test what the significance level of the difference is. For this purpose a metric and a statistical test are needed. In this paper we compare several combinations of a metric with a statistical test, in order to find a combination which is suitable for this task. Four combinations which are introduced in this paper appear to be suitable for this task.

Volume 3, page 2091 Session D42

Prediction of Low Recognition Rate Words for Isolated Word Recognition System

Terashima R, Hoshino H, Wakita T
Toyota Central R&D Labs., Japan

This paper describes an efficient method to predict words whose recognition rates are low. This method is based on the idea that the minimum value of the word pair recognition rates corresponds to the

word recognition rate. The word pair recognition rates can be calculated by the measured distributions of phoneme log likelihood difference. The proposed method was evaluated by recognition experiments using about 3000 word pairs. The correlation coefficient between the predicted and the measured recognition rates was 0.87 when the phoneme lengths of both words of the word pair were equal. Furthermore, we also estimated a 95% confidence interval for the measured recognition rates, and the percentage of the predicted words that were contained in the confidence interval was 94.8%. The results showed the effectiveness of the proposed method for predicting the word pair recognition rates.

Volume 3, page 2095 Session D42

An Objective Measure for Assessment of the Concatenative TTS Segment Inventories

Batusek R
Masaryk University, Czech Republic

In the paper we present a method for assessment of the segment inventories for concatenative text-to-speech synthesis. We argue that the overall comprehensibility of the synthesized speech depends on the length of the segments - longer segments imply more intelligible speech. The problem of minimum text cover by the given segment set is formulated in the paper as well as an algorithm finding the solution. Some improvements speeding up the algorithm are discussed in the rest of the paper.

Volume 3, page 2099 Session D42

Session D43 - Oral
Thursday - 15.50 - 17.30

Speech Recognition and Understanding: Confidence Measures

Chair: Richard Rose, AT&T Labs - Research, USA

Word Level Confidence Annotation using Combinations of Features

Zhang R, Rudnicky A
Carnegie Mellon University, USA

This paper describes the development of a word-level confidence metric suitable for use in a dialog system. Two aspects of the problems are investigated: the identification of useful features and the selection of an effective classifier. We find that two parse-level features, Parsing-Mode and Slot-Backoff-Mode, provide annotation accuracy comparable to that observed for decoder-level features. However, both decoder-level and parse-level features independently contribute to confidence annotation accuracy. In comparing different classification techniques, we found that Support Vector Machines (SVMs) appear to provide the best accuracy. Overall we achieve 39.7% reduction in annotation uncertainty for a binary confidence decision in a travel-planning domain.

Volume 3, page 2105 Session D43

A Boosting Approach for Confidence Scoring

Moreno P J, Logan B, Raj B
Compaq Cambridge Research Lab., USA

In this paper we present the application of a boosting classification algorithm to confidence scoring. We derive feature vectors from speech recognition lattices and feed them into a boosting classifier. This classifier combines hundreds of very simple `weak learners' and derives classification rules that can reduce the confidence error rate by up to 34%. We compare our results to those obtained using two other standard classification techniques, Support Vector Machines (SVMs) and Classification and Regression Trees (CART), and show significant improvements. Furthermore, the nature of the boosting algorithm allows us to combine the best single classifier and improve its performance. We present experimental results on real world corpora derived from our SpeechBot Web index http://www.speechbot.com) and from the HUB4 DARPA evaluation sets. We believe these results have wide applicability to audio indexing and to acoustic and language modeling adaptation where word confidence scores can be used in iterative adaptation schemes.

Volume 3, page 2109 Session D43

On Combining Confidence Measures for Improved Rejection of Incorrect Data

Charlet D, Mercier G, Jouvet D
France Télécom R&D, France

In this paper, techniques for combining confidence measures are proposed and evaluated. Confidence measures are useful for rejecting incorrect data, which is an important issue in speech recognition based interactive systems. Many ways of computing individual confidence measures have already been investigated. A detailed analysis of various confidence measures shows that they behave differently for what concerns rejection of incorrect data on various field data subsets (substitution errors, out-of-vocabulary data & noise tokens) collected from a vocal directory task. Two combination methods are then presented. One combines confidence measures by means of a neural network and the other through logistic regression. Evaluations shows that both combination techniques are efficient, and both take the best of the various individual confidence measures involved on each data subset.

Volume 3, page 2113 Session D43

Improved Word Confidence Estimation using Long Range Features

Palmer D[1], Ostendorf M[2]
[1]The MITRE Corporation, USA, [2]University of Washington, USA

This paper describes experiments in improving word confidence estimation using document- and task-level features of the hypothesized word sequence from a recognizer. The improved confidence estimates are shown to improve information extraction performance, specifically named entity (NE) recognition. The detected names can then be used to further improve confidence estimation in a multi-pass NE recognition framework.

Volume 3, page 2117 Session D43

Is This Conversation on Track?

Carpenter P, Jin C, Wilson D, Zhang R, Bohus D, Rudnicky A
Carnegie Mellon University, USA

Confidence annotation allows a spoken dialog system to accurately assess the likelihood of misunderstanding at the utterance level and to avoid breakdowns in interaction. We describe experiments that assess the utility of features from the decoder, parser and dialog levels of processing. We also investigate the effectiveness of various classifiers, including Bayesian Networks, Neural Networks, SVMs, Decision Trees, AdaBoost and Naive Bayes, to combine this information into an utterance-level confidence metric. We found that a combination of a subset of the features considered produced promising results with several of the classification algorithms considered, e.g., our Bayesian Network classifier produced a 45.7% relative reduction in confidence assessment error and a 29.6% reduction relative to a handcrafted rule.

Volume 3, page 2121 Session D43

Session D44 - Oral
Thursday - 15.50 - 17.30

Speech Recognition and Understanding: Language Modelling

Chair: To be decided,

Automatic N-gram Language Model Creation from Web Resources

Nisimura R[1], Komatsu K[2], Kuroda Y[3], Nagatomo K[1], Lee A[1], Saruwatari H[1], Shikano K[1]
[1]Nara Institute of Science and Technology, Japan, [2]Laboratories of Image Information Science and Technology, Japan, [3]TIS Inc., Japan

This paper describes an automatic building of N-gram language models from Web texts for large vocabulary continuous speech recognition. Although a huge amount of well-formed texts are needed to train a model, collecting and organizing such text corpus for every task by hand needs a great labor. We need the language model to update frequently to cover the current topics. To deal with this problem, we propose an automatic language model creation method by collecting Web texts via keyword-based Web search engines. We can build a task-dependent language model by selecting suitable keywords for the task. A text filtering algorithm based on character perplexity is developed to extract proper Japanese texts from Web texts. A language model for a medical consulting task created by the proposed method shows the higher word recognition rate by 11.4% than that of a conventional newspaper language model.

Volume 3, page 2127 Session D44

On Integrating the Lexicon with the Language Model

Caseiro D A, Trancoso I
INESC-ID/IST, Portugal

The goal of this work was to develop an algorithm for the integration of the lexicon with the language model which would be computationally efficient in terms of memory requirements, even in the case of large trigram models. Two specialized versions of the algorithm for transducer composition were implemented. The first one is basically a composition algorithm that uses the precomputed set of the output labels that can be reached from a particular epsilon edge of the lexicon; the second includes an "on the fly" implementation of the pushing of weights and output labels. Very significant memory savings were obtained with the proposed algorithms compared with the general determinization algorithm for weighted transducers.

Volume 3, page 2131 Session D44

Back-off smoothing evaluation over syntactic language models

Varona A, Torres I
Facultad de Ciencias. UPV-EHU., Spain

Continuous Speech Recognition systems require a Language Model (LM) to represent the syntactic constraints of the language. In LMs development a smoothing technique needs to be applied to also consider events not represented in the training corpus. In this work, several back-off smoothing approaches have been compared: classical discounting-distribution schema including Witten-Bell, Absolute and Linear discounting and a new proposal, the Delimited discounting. Delimited discounting deals with the Turing discounting problems while keeping the Katz´s smoothing scheme. The experimental evaluation was carried out over a Spanish speech application task, showing that an increase of the test set perplexity of a LM does not always mean a degradation in the model performance when integrated into a CSR system. Besides, there is a strong dependence between the amount of probability reserved by the

smoothing technique to be assigned to unseen events and the value of the balance parameter applied to the LM probabilities in the Bayes´s rule needed to get the best system performance.

Volume 3, page 2135 Session D44

An Online Incremental Language Model Adaptation Method

Wu G, Zheng F, Jin L, Wu W
Tsinghua University, P. R. China

In this paper, an online incremental language model adaptation method is proposed, which is different from the traditional offline language model adaptation method. There are some problems in the online incremental adaptation. The first one is how to adjust the model parameters online and modify the model incrementally. The second one is how to induce new words and assign initial probabilities to the n-grams related to them. In our application for Chinese character input method editor, the language model is divided into two parts, corresponding to the background (general-purpose) model and the user model, respectively. A modified maximum a posterior method is proposed for adapting the user model dynamically. Experiments are done to test the proposed method on an Chinese sentence input system and the results show that a satisfying word error rate reduction is obtained when the input articles are of similar topics.

Volume 3, page 2139 Session D44

Using Boosting and POS Word Graph Tagging to Improve Speech Recognition

Samuelsson C[1], Hieronymus J[2]
[1]Xerox Research Centre Europe, France, [2]Research Institute for Advanced Computer Science, USA

The word graphs produced by a large vocabulary speech recognition system usually contain a path labelled with the correct utterance, but this is not always the highest scoring path. Boosting increases the probability of words which occur often in the word graph, which are in some sense robust. Adding syntactic information allows rescoring of arc probabilities with the possibility that more grammarical word sequences will also be the correct ones. A theory is developed which allows general probabilistic syntactic models to be used to rescore word lattices. Experiments conducted on the Wall Street Journal (WSJ) corpus with a version of the AT&T 1995 FST LVSR system with part of speech (POS) trigram sequences show that using only POS leads to a loss in performance. Boosting alone provides an improvement in performance which is not statistically significant. Cascading the two methods, boosting first and then using syntactic information improves performance 4.5 % relative on a large portion of the 1995 DARPA test set.

Volume 3, page 2143 Session D44

Session D45 - Poster
Thursday - 15.50 - 17.30

Dialogue Systems: Techniques and Strategies

Chair: Julia Hirschberg, AT&T, Florham Park, USA

Robust Parsing in Spoken Dialogue Systems

Yan P, Zheng F, Xu M
Tsinghua University, P. R. China

The rule-based parsing is a prevalent method for the natural language understanding (NLU) and has been introduced in dialogue systems for spoken language processing (SLP). However, additional measures must be taken to cope with the severe spoken linguistic phenomena, such as garbage, repetition, ellipsis, word disordering, fragment and ill form, which frequently occur in the spoken language. We propose in this paper a robust parsing scheme, which integrates the following methods. Keywords are used as terminal symbols; hence the symbol set of the grammar is purely within the semantical category. The definition of the grammar is extended to accommodate four types of rules, called up-tying, by-passing, up-messing, and over-crossing respectively. An improved chart parser, named marionette, is designed to parse the semantic grammar instance. The robust parsing scheme has been adopted in an air traveling information service system, called EasyFlight, and has achieved a high performance when dealing with the spontaneous speech.

Volume 3, page 2149 Session D45

A Theme Structure Method for the Ellipsis Resolution

Huang Y, Zheng F, Su Y, Li F, Wu W
Tsinghua University, P. R. China

The purpose of this paper is to solve the contextual ellipsis problem that is popular in our Chinese spoken dialogue system named EasyNav. A Theme Structure is proposed to describe the attentional state. Its dynamic generation feature makes it suitable to model the topic transition in user-initiative dialogues. By studying the differences and the similarities between the ellipsis and the anaphora phenomena, we extend the resolution procedure and the theory from anaphora to ellipsis. The ellipsis resolution is now based on the semantic knowledge and the discourse factor other than the syntactic information. A Theme Structure Method proposed in this paper for the ellipsis resolution is uniform to not only all kinds of elliptical elements but also some particular ellipsis types such as the fragmental ellipsis and the default ellipsis.

Volume 3, page 2153 Session D45

Deriving Document Structure from Prosodic Cues

Haase M[1], Kriechbaum W[2], Möhler G[1], Stenzel G[2]
[1]Universität Stuttgart, Germany, [2]IBM Deutschland Entwicklung, Germany

This study presents an approach for prosody-driven segmentation of speech data. The model is based solely on F0 contours and RMS envelopes. Phoneme or word information from a speech recognizer is unneccesary. Using data from German broadcast news, we show how this prosodic information can be exploited to retrieve structural information of the spoken text. The suitability of the CART-like algorithm for utterance boundary prediction has been evaluated on 7 five-minutes-news-reports, using 28 reports as training material for the classification tree. Sentence boundaries were predicted with a precision of 93%, at a recall of 88%.

Volume 3, page 2157 Session D45

Design of a Semantic Parser with Support to Ellipsis Resolution in a Chinese Spoken Language Dialogue System

Su Y, Zheng F, Huang Y
Tsinghua University, P. R. China

In this paper, a semantic parser with support to ellipsis resolution in a Chinese spoken language dialogue system is proposed. The grammar and parsing strategy of this parser is designed to address the characteristics of spoken language and to support the ellipsis resolution. Namely, it parses the user utterance with a domain-specific semantic grammar based on a template-filling approach. Syntactic constraints extracted by a Generalized LR parser are also used in the parsing process. With a paradigm of two-state bottom-up parsing and a scoring scheme, the ellipsis resolution module is integrated into the parser seamlessly. The parsing result is represented by a linked structure of semantic frames, which is convenient to both the parser and its successive components of the dialogue system.

Volume 3, page 2161 Session D45

Methodology for Dialogue Design in Telephone-Based Spoken Dialogue Systems: a Spanish Train Information System

San-Segundo R, Montero J M, Colás J, Gutiérrez J M, Ramos J M, Pardo J M
UPM., Spain

In this paper, we propose a new methodology for designing dialogue managers in telephone-based spoken dialogue systems. This methodology comprises five steps: database analysis, design by intuition, design by observation, simulation and iterative improvement. At each step, several measures to evaluate the designing alternatives are presented. We introduce confidence measures in recognition to define an efficient confirmation strategy in each case. The use of user-modeling techniques adapts the system to the user ability. This methodology is applied for designing a telephone-based system that provides rail travel information for the main Spanish intercity connections, including timetables, simulated fares and reservations. With 30 users completing 4 scenarios, the average duration for a fully automatic call is 204 seconds. The users validated the applicability and usability of the system with a global score of 3.9 (out of 5).

Volume 3, page 2165 Session D45

Spoken Dialogue Management as Planning and Acting under Uncertainty

Zhang B[1], Cai Q[1], Mao J[2], Chang E[3], Guo B[3]
[1]Univ. of Sci. & Tech. of China, P. R. China, [2]Tsinghua University, P. R. China, [3]Microsoft Research China, P. R. China

Some stochastic models like Markov decision process (MDP) are used to model the dialogue manager. MDP-based system degrades fast when uncertainty about user¡¯s intention increases. We propose a novel dialogue model based on the partially observable Markov decision process (POMDP). We use hidden system states and user intentions as the state set, parser results and low-level information as the observation set, domain actions and dialogue repair actions as the action set. Here the low-level information is extracted from different input modals using Bayesian networks. Because of the limitation of exact algorithms, we focus on heuristic methods and their applicability in dialogue management.

Volume 3, page 2169 Session D45

Modeling of Conversational Strategy for the Robot Participating in the Group Conversation

Matsusaka Y, Fujie S, Kobayashi T
Waseda Univ., Japan

This paper describes a strategy for the conversation system to take part in human-to-human group conversation. One big characteristic of the group conversation system is that it can choose whether to observe or to take turn in the conversation. We implement the computational model combined with speech and gaze recognizers to keep the rules in turn taking, and define an interruption decision strategy based on an analysis of human needs. And finally, we realized a human-friendly group conversation system by combining multi-modal information processing/expression abilities of humanoid robot ROBITA.

Volume 3, page 2173 Session D45

Supporting the Construction of a User Model in Speech-only Interfaces by Adding Multi-modality

Terken J, Riele, te S
Technische Universiteit Eindhoven, the Netherlands

Comparing to graphical user interfaces, speech-only interfaces face several problems: robustness, making clear what functionality is available, and making clear how the functionality may be accessed. We explore a potential solution for these problems by presenting a visual representation of the domain of discourse and of the state of the dialogue. We describe an experiment in which uni-modal and multi-modal interfaces are compared in terms of effectiveness, efficiency and satisfaction. The results of the experiment show a strong learning effect. Subjects who start using the multi-modal interface subsequently have a strong advantage when switching to the uni-modal (speech-only) interface, compared to subjects who start by using the uni-modal interface, switching to the multi-modal interface later on. The results are discussed in terms of the need to establish an appropriate user model as early as possible. We discuss implications of this interpretation for interaction design.

Volume 3, page 2177 Session D45

A Word- and Turn-Oriented Approach to Exploring the Structure of Mandarin Dialogues

Tseng S-C
Institute of Linguistics, Academia Sinica, Taiwan

This paper investigates the structure of Mandarin spoken dialogues by analysing the distribution of words and turns used in dialogues. The results of an empirical quantitative study show that independent of speakers, there exists a kind of basic vocabulary for daily Mandarin conversations. It is proposed that this is the minimal set of a lexicon for the use of spoken Mandarin. Moreover, a number of words in the basic vocabulary were specifically used for marking various constituent boundaries in discourse, such as turn-initial and utterance-final. Discourse markers and disfluency are also taken into consideration for their highly frequent occurrences and their function of marking significant positions in spoken discourse. By means of lexical distribution and its interaction with turn taking, this paper demonstrates a new attempt to analyse the structure of Mandarin spoken dialogues.

Volume 3, page 2181 Session D45

A Rule Based Approach to Extraction of Topics and Dialog Acts in a Spoken Dialog System

Niimi Y, Oku T, Nishimoto T, Araki M
Kyoto Institute of Technology, Japan

This paper presents a rule based approach to extraction of dialog acts and topics from utterances in a spoken dialog system with a task-independent dialog controller based on an extension of the frame-driven method. We demonstrated it could control dialogs in several different task domains, only given a set of topic frames and a set of rules manually designed for the discourse analysis which were both task-dependent. In this paper we report an approach to semiautomatic derivation of a set of rules for the discourse analysis from information needed to specify a task, for

example, a set of topic frames, a conceptual tree of those words and case frames of those verbs which are likely to be used in the task domain. This method was examined with a corpus of twenty dialogs. Correct extraction rates were 82% for the topic and 80% for the dialog act.

Volume 3, page 2185 Session D45

Agent-based Error Handling in Spoken Dialogue Systems

Turunen M, Hakulinen J
Univ. of Tampere, Finland

In this paper, we introduce an agent-based error handling ar-chitecture for spoken dialogue systems. In this architecture, all the parts of the error-handling process on the different in-teraction levels (input, dialogue and output) are explicitly modeled. Error handling is divided into individual, preferably application independent components. The proposed architec-ture makes it possible to construct adaptive and reusable error handling components and entire error-handling toolkits. The architecture is especially suitable for multilingual applica-tions. The architecture is implemented as part of the Jaspis speech application development environment and it uses Jaspis' agent-based interaction model.

Volume 3, page 2189 Session D45

Iterative Implementation of Dialogue System Modules

Degerstedt L, Jönsson A
Linköping University, Sweden

This paper presents an approach to the implementation of modules for dialogue systems. The implementation method is divided into two distinct, but correlated, steps; Conceptual design and Framework customisation. Conceptual design and framework customisation are two mutually dependent sides of the same phenomena, where the former is an on-paper activity that results in a design document and the latter results in the actual implementation. The method is iterative and applicable in various phases of dialogue system development and also for different dialogue system modules. We also present the development of the dialogue management module in more detail. The development space for such modules involves issues on modularisation, knowledge representation and interface functionality internally, and between modules. Orthogonal to this are the various types of re-use for framework customisation; tools, framework template and code patterns. Taken together they form a scheme which is explored during the implementation process.

Volume 3, page 2193 Session D45

Off-Talk - a Problem for Human-Machine-Interaction?

Oppermann D, Schiel F, Steininger S, Beringer N
University of Munich, Germany

This paper is concerned with the definition and description of the phenomenon Off-Talk in human-machine-interaction. This phenomenon is considered to cause problems due to non-relevant information that is conveyed within these utterances. Besides the definition of Off-Talk our work aims to provide an analysis of transcribed audio data that is part of the SmartKom data collection. In the search for features that could indicate the occurrence of Off-Talk we looked at several speech levels e.g. acoustics, lexicon and prosody. Due to the small amount of available data only three features were examined, as there are: loudness, word frequency and filled pauses. The analysis revealed that a correlation might exist between Off-Talk and all features, so that they may serve as indicators for this phenomenon.

Volume 3, page 2197 Session D45

Automatic Analysis of Real Dialogues and Generating of Training Corpora

Schwarz J[1], Matousek V[2]
[1]*Technical University of Dresden, Germany*, [2]*University of West Bohemia in Pilsen, Czech Republic*

The development of computerized information retrieval dialogue systems communicating with the user in natural language requires the implementation of an effective training procedure with the aid of which the main modules of the dialogue system have to be partly automatically developed. The presented paper describes an attempt to create the generating sentence templates automatically, using a special program package implementing an especially developed method of a quantitative linguistic analysis of transcribed real dialogues. Firstly, the program package generates a set of formulas (templates) consisting of a special grammar and describing the syntactic structure of required sentences. Secondly, it generates a large corpus of unique training sentences using the sentence templates and a stochastic context-free grammar. The experimentally created corpus was used for the training of modules of a city information dialogue system.

Volume 3, page 2201 Session D45

Natural Language Understanding Using Statistical Machine Translation

Macherey K, Och F J, Ney H
RWTH Aachen, University of Technology, Germany

Over the past years, automatic dialogue systems have received increasing attention. In addition to a speech recognizer, such systems include a natural language understanding (NLU) component. One of the most investigated approaches to NLU are rule-based methods as stochastic grammars which are often written manually. However, the sole usage of rule-based methods can turn out to be inflexible when extending or changing the application's domain. Therefore, techniques are desirable which help to reduce the manual effort when creating an NLU component for a new domain. In this paper we investigate an approach to NLU which is derived from the field of statistical machine translation. Starting from a bilingual annotated corpus, we describe the problem of NLU as the translation from a source to a target sentence. Experiments were performed on the TABA corpus which is a text corpus in the domain of a German train timetable information system.

Volume 3, page 2205 Session D45

Improvements in Audio Processing and Language modeling in the CU Communicator

Zhang J, Ward W, Pellom B, Yu X, Hacioglu K
University of Colorado at Boulder, USA

This paper presents some up-to-date audio processing techniques which have been developed and integrated into the University of Colorado (CU) communicator system. The CU Communicator is an interactive human-machine dialogue system for airline, hotel and rental car information. The baseline system was fully functional in June 1999. Since then, many improvements have been made. The paper will concentrate on acoustic echo cancellation, voice activity detection (VAD) and language modeling techniques and provide a paradigm for speech and audio processing in a dialog system with barge-in capabilities. Specifically, a real-time block least-mean-square (LMS) algorithm is discussed. A robust voice activity detector using energy threshold is applied to detect user voice. Experimental results are presented and some real-time implementation issues are addressed.

Volume 3, page 2209 Session D45

Dialogue Session Management Using VoiceXML

Tsai A, Pargellis A N, Lee C-H, Olive J P
Bell Labs, Lucent Technologies, USA

A spoken dialogue system capable of maintaining a human-machine conversation over a telephone must simultaneously maintain many dialogue sessions with different callers and third-party servers running applications over the Internet. There are three main issues to be addressed. First, the caller's identity has to be passed between each dialogue turn. Secondly, the dialogue state and connections to third party servers have to be continuously maintained. Thirdly, the system interfaces have to follow some standards, especially Internet protocols. We use session tickets and session object hashes to address these issues. Each dialogue turn is encoded in a session data object and VXML page. The session tickets contain the session ID and security-related information passed between the VoiceXML interpreter and document server. The session object hash stores the third party connection handle, which is retrieved for subsequent dialogue turns.

Volume 3, page 2213 Session D45

Ambiguity Representation and Resolution in Spoken Dialogue Systems

Ammicht E, Potamianos A, Fosler-Lussier E
Bell Labs, Lucent Technologies, USA

Spoken natural language often contains ambiguities that must be addressed by a spoken dialogue system. In this work, we present the internal semantic representation and resolution strategy of a dialogue system designed to understand ambiguous input. These mechanisms are domain independent; task-specific knowledge is represented in parameterizable data structures. Speech input is processed through the speech recognizer, parser, interpreter, context tracker, pragmatic analyzer and pragmatic scorer. The context tracker combines dialogue context and parser output to yield raw attribute-value (AV) pairs from which candidate values are derived. The pragmatic analyzer adjusts the confidence associated with each AV candidate based on system intent, e.g., implicit confirmation, and on user input. Pragmatic confidence scores are introduced to measure the dialogue managers confidence for each AV: MYCIN-like scoring is used to merge multiple information sources. Pragmatic analysis and scoring is combined with explicit error correction capabilities to achieve efficient ambiguity resolution. The proposed strategies greatly improve dialogue interaction, eliminating about half of the errors in dialogues from a travel reservation task.

Volume 3, page 2217 Session D45

Session D46 - Poster
Thursday - 15.50 - 17.30

Speech Synthesis: Miscellaneous

Chair: Jan van Santen, OGI, USA

Feature Extraction by Auditory Modeling for Unit Selection in Concatenative Speech Synthesis

Tsuzaki M
ATR Spoken Language Translation Research Laboratories, Japan

A comprehensive computational model of the human auditory peripherals was applied to extract basic features of speech sounds. The auditory model extracts features by the auditory temporal coding mechanism in addition to features by the auditory place coding mechanism which has traditionally been used as spectral features. It also considers the nonlinearity of human auditory responses. Several speech databases of different talkers for a concatenative synthesis system were analyzed by the auditory model, and segmental characteristics were estimated by calculating the averages, standard deviations, and trends of individual feature parameters. The results were compared with results obtained by a physical model. A preliminary perceptual test suggested an advantage of auditory-based distances over physical distances.

Volume 3, page 2223 Session D46

Perceptual Cost Functions for Unit Searching in Large Corpus-based Text-to-Speech

Lee M
Bell Labs, Lucent Technologies, USA

In large corpus-based concatenative Text-to-Speech, unit selection is critical for the quality of synthetic speech. Dynamic programming algorithms have been used for unit-searching by minimizing a total cost (1) between target specification and candidate units and (2) between candidate units for concatenation. The cost function is often a weighted sum of sub-costs, which are the costs for each of the acoustic and phonetic features of units. The weights control the individual contribution of the sub-costs to the total cost. They also determine the relative sensitivity of a feature to the quality degradation when signal processing is applied to modify the feature. However, determining the weights for the cost function has not been a simple task. In this paper, we propose a new method for unit-searching based on a perceptual preference test. The proposed algorithm is designed to find the weights in more systematic and meaningful way. The algorithm searches for a set of weights that can produce a ranking of renditions, that is close to the perceptual test results. The downhill simplex method is used for the multi-dimensional search of the weights. A dissimilarity measure is proposed to evaluate the closeness of two rankings. About 83 percent of the cases, the unit selection algorithm using the optimal set of weights choose the same rendition that human listeners prefer. The results show that the proposed weight optimization algorithm can successfully predict the human preference pattern. The synthetic speech using the optimal weights consistantly showed smoother transition and higher voice quality than the one using manually determined weights.

Volume 3, page 2227 Session D46

Pruning of Redundant Synthesis Instances Based on Weighted Vector Quantization

Kim S[1], Lee Y[1], Hirose K[2]
[1]*Electronics and Telecommunications Research Institute, Korea,*
[2]*University of Tokyo, Japan*

A new method of pruning redundant synthesis unit instances in a large-scale synthesis database was proposed based on weighted vector quantization (WVQ). WVQ takes relative importance of each instance

into account when clustering the similar instances using vector quantization (VQ) technique. The proposed method was compared with two conventional pruning methods through objective and subjective evaluations of synthetic speech quality: one to simply limit maximum number of instance, and the other based on normal VQ-based clustering. For the same reduction rate of instance number, the proposed method showed the best performance. The synthetic speech with reduction rate 50% sounded with no perceptible degradation as compared to that without instance reduction.

Volume 3, page 2231 Session D46

Using Real Words for Recording Diphones

Fitt S
University of Edinburgh, UK

This paper focuses on the creation of word-lists for making diphone recordings for speech synthesis. Such lists often consist of nonsense words, which has the advantage that the phonetic environment can be constrained, and it is easy to produce lists containing all possible combinations. However, this approach has the disadvantage that non-experts may find it difficult to read the nonsense-word transcriptions. For this reason, we investigate here the issues associated with the use of real words in creating diphone recordings.

Volume 3, page 2235 Session D46

Application of the Trended Hidden Markov Model to Speech Synthesis

Dines J D S, Sridharan S, Moody M
RCSAVT, QUT, Australia

This paper presents our work on a speech synthesis system that utilises the trended Hidden Markov Model to represent the basic synthesis unit. We draw upon both speech recognition and speech synthesis research to develop a system that is able to synthesise intelligible and natural sounding speech. Acoustic units are clustered using the decision tree technique and speech data corresponding to these clusters is used for the training of trended Hidden Markov Model synthesis units. The overall system has been implemented in a PSOLA synthesiser and the resultant speech has been compared with that produced by a conventional diphone synthesiser to yield very encouraging results.

Volume 3, page 2239 Session D46

Two Features to Check Phonetic Transcriptions in Text to Speech Systems

Sandri S, Zovato E
Loquendo, Vocal Technology and Services, Italy

The paper describes a framework to overcome some problems in the analysis of speech corpora used in text-to-speech systems. In particular two kinds of errors that can produce disagreeable effect at synthesis level have been examined. The first of them is the incorrect transcription of pauses (and more generally low energy intervals) and the second one is the mismatch between voiced intervals and the phonetic symbol that should represent them. For the first problem a statistical approach has been used, by comparing some features of the detected low energy intervals (LE) with those of trained data. The second problem has been faced extracting the voiced/unvoiced intervals (VU) and checking the coherence with the phonetic transcription and segmentation.

Volume 3, page 2243 Session D46

Text-to-Speech Scripting Interface for Appropriate Vocalisation of e-Texts

Xydas G, Kouroupetroglou G
University of Athens, Greece

Electronic texts carry important meta-information (such as tags in HTML) that most of the current Text-to-Speech (TtS) systems ignore during the production of the speech. We propose an approach to exploit this meta-information in order to achieve a detailed auditory representation of an e-text. The e-Text to Speech and Audio (e-TSA) Composer has been designed and developed as an XML based scripting framework that can be adopted by existing TtS, with minor or major modifications. It provides a mechanism to create scripts using combined elements from e-texts and TtS systems. The e-TSA Composer can manipulate the behaviour of a TtS (e.g. the applied prosody) in order to define a finest vocalisation in response to specific e-texts.

Volume 3, page 2247 Session D46

Representation of Large Lexica Using Finite-State Transducers for the Multilingual Text-to-Speech Synthesis Systems

Rojc M, Kacic Z
University of Maribor, Slovenia

Large external language resources used for multilingual text processing in TTS systems represent a big problem because of needed space and slow look-up time. Representation of large lexica using finite-state transducers is mainly motivated by considerations of space and time efficiency. In the paper we present a method and results of compiling large German phonetic and morphology lexica (CISLEX) [4] into corresponding finite-state transducers (FSTs), both with about 300.000 words. For both lexica a great reduction in size and optimal access time was achieved. The starting size for German phonetic lexicon was 12.526 MB and 18.49 MB for morphology lexicon. The final size of the corresponding FST was only 2.78 MB for the phonetic lexicon and 6.33 MB for the morphology lexicon. At the same time the look-up time is optimal, since it depends only on the length of the input word and not on the size of the lexicon.

Volume 3, page 2251 Session D46

Corpus-Based Synthesis of Fundamental Frequency Contours Based on A Generation Process Model

Hirose K[1], Eto M[1], Minematsu N[1], Sakurai A[2]
[1]University of Tokyo, Japan, [2]Tsukuba R&D Center, Texas Instruments Japan, Japan

A mode-constrained corpus-based synthesis strategy was developed for F0 contours of Japanese sentences. In the training phase, the relationship between linguistic factors and the command values of F0 contour generation process model was learned using neural networks. Input parameters consist of linguistic information related to accentual phrases that can be automatically driven from text, such as the number of morae, and so on. In the synthesis phase, the network is used to generate the command values. The synthesis method was also realized based on multiple linear regression analysis to examine how each input parameter contributes to the F0 contour generation. The use of the parametric model restricts the degrees of freedom of the mapping between linguistic and prosodic features, and thus enables to generate appropriate values even with limited training data. Experimental results showed that the method could generate F0 contours quite close to those by the rule-based method.

Volume 3, page 2255 Session D46

Corpus-Based Database of Residual Excitations Used for Speech Reconstruction from MFCCs

Tychtl Z, Psutka J
University of West Bohemia, Czech Republic

This paper proposes a new approach to extraction of a corpus-based database of residual signal segments that are used as excitations of a production model to replay MFCC encoded speech signal with natural

sound. Neither extra information besides the MFCCs (like F0, voiced/unvoiced flag etc.) nor modification and/or extension of a MFCC computation algorithm is needed. The MFCC algorithm is considered to be in a commonly accepted form that was implemented for example in the HTK software. Because of mentioned restrictions we don't aim to achieve exact reconstruction of original signal but we seek to replay the speech signal in an intelligible and as natural as possible way. Moreover, the 'low-demanding' solution based on pulse/noise excitation is offered that employs a new method for making voiced/unvoiced decision using the MFCC vector only.

Volume 3, page 2259 Session D46

Mixed Excitation for HMM-based Speech Synthesis

Yoshimura T[1], Tokuda K[1], Masuko T[2], Kobayashi T[2], Kitamura T[1]
[1]Nagoya Institute of Technology, Japan, [2]Tokyo Institute of Technology, Japan

This paper describes improvements on the excitation model of an HMM-based text-to-speech system. In our previous work, natural spectral and pitch parameters have been generated from HMM by using a speech parameter generation algorithm. However, synthesized speech has a typical quality of ``vocoded speech'' since the system used a traditional excitation model with either a periodic impulse train or white noise. In this paper, in order to reduce the synthetic quality, a mixed excitation model used in MELP is incorporated into the system. Excitation parameters used in mixed excitation are modeled by HMMs, and generated from HMMs by a parameter generation algorithm in the synthesis phase. The result of a listening test shows that the mixed excitation model significantly improves quality of synthesized speech as compared with the traditional excitation model.

Volume 3, page 2263 Session D46

Aperiodicity Control in ARX-Based Speech Analysis-Synthesis Method

Ohtsuka T, Kasuya H
Utsunomiya University, Japan

We present an improved algorithm for a robust speech analysis-synthesis method based on an auto-regressive with exogenous input (ARX) speech production model proposed previously. The speech analysis-synthesis method is capable of making an automatic estimation of vocal tract (formant) and voice source parameters from a speech utterance, generating accurate formant values even for very high-pitched voices. The improved algorithm presented in this paper incorporates aperiodic components included in the voice source signal, taking the dynamic nature of the speech production process into account. Perceptual experiments show that implementation of the aperiodic components in the analysis-synthesis is very effective in improving the perceived quality of synthetic speech, particularly for soft voices, typical of female voice quality.

Volume 3, page 2267 Session D46

Generalized Source-Filter Structures for Speech Synthesis

Karjalainen M, Paatero T
Helsinki University of Technology, Finland

In this paper we discuss various digital filter principles as models for synthetic speech generation. Warped linear prediction (WLP) and frequency-warped filters have been introduced earlier as a method to reduce the filter order in high-quality wideband speech synthesis. In addition to analyzing WLP and frequency-warped filters we introduce new related structures and techniques for arbitrary frequency resolution allocation. Kautz filters can be considered as generalized structures for pole-zero modeling. This study focuses on residual-excited synthesis and diphone-oriented reconstruction of speech signals. Control strategies for text-to-speech synthesis are discussed briefly.

Volume 3, page 2271 Session D46

The Speech Synthesis Environment and Parametric Modeling of Coarticulation

Mikolaj W
Adam Mickiewicz University, Poland

A general description of the environment for speech processing called SLOPE is presented. The main area of application of the SLOPE environment is mid and low-level speech synthesis - the area between prosody modeling and a speech waveform generation. The final part of the article describes the idea of obtaining a strict parametric form of the utterance from the string of phones and prosodic information implemented on the basis of SLOPE

Volume 3, page 2275 Session D46

Session E11 - Oral
Friday - 09.00 - 10.40

ESE7 - Integration of Phonetic Knowledge in Speech Technology: Experiments and Experiences

Chair: Wim van Dommelen, Department of Linguistics, NTNU, Trondheim, Norway

Defining Constraints for Multilinear Speech Processing

Carson-Berndsen J, Walsh M
University College Dublin, Ireland

This paper presents a constraint model for the interpretation of multilinear representations of speech utterances which can provide important fine-grained information for speech recognition applications. The model uses explicit structural constraints specifying overlap and precedence relations between features in both the phonological and the phonetic domains in order to recognise well-formed syllable structures. In the phonological domain, these constraints together form a complete phonotactic description of the language, while in the phonetic domain, the constraints define the internal structure of phonologial features based on phonetic realisations. The constraints are enhanced by a constraint relaxation procedure to cater for underspecified input and allows output representations to be extrapolated based on the phonetic and phonological information contained in the constraints and the rankings which have been assigned to them. This approach thus addresses issues of robustness in speech recognition.

Volume 4, page 2281 Session E11

Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground.

Batliner A[1], Möbius B[2], Möhler G[2], Schweitzer A[2], Nöth E[1]
[1]University of Erlangen-Nuremberg, Germany, [2]University of Stuttgart, Germany

Automatic speech understanding and speech synthesis, two of the major speech processing applications, impose strikingly different constraints and requirements on prosodic models. The prevalent models of prosody and intonation fail to offer a unified solution to these conflicting constraints. As a consequence, prosodic models have been applied only occasionally in end-to-end automatic speech understanding systems; in contrast, they have been applied extensively in speech synthesis systems. In this paper we want to discuss the reasons for this state of affairs as well as possible strategies to overcome the shortcomings of the use of prosodic modelling in automatic speech processing.

Volume 4, page 2285 Session E11

Introducing Phonetically Motivated Information into ASR

Christensen H, Lindberg B, Andersen O
Aalborg University, Denmark

In this paper we present an approach to introducing more phonetically motivated information into automatic speech recognition in the form of a phonetic `expert'. To avoid the curse of dimensionality problem, the expert information is introduced at the level of the acoustic model. Two types of experts are used each providing discriminative information regarding groups of phonetically related phonemes. The phonetic expert is implemented using an MLP. A numbers recognition task shows that, when using the expert in conjunction with both a fullband and a multi-band system speech recognition performances are increased.

Volume 4, page 2289 Session E11

Integrating Contextual Phonological Rules in a Large Vocabulary Decoder

Gravier G[1], Yvon F[2], Jacob B[3], Bimbot F[1]
[1]IRISA, France, [2]ENST Paris, France, [3]LIUM, France

This paper presents an approach to the integratation of contextual phonological rules in the beam-search algorithm of a large vocabulary speech recognition system. The main interest of contextual transcription rules is that they implement constraints on pronunciations sequences which complement the bigram constraints on word sequences. As such, they should help avoiding acoustic confusions and reduce the search space. In our approach, contextual transcription do not incur any augmentation of the lexicon size. This approach is evaluated on a dictation task in French for two different sets of contextual phonological rules. Our results show that, given the current resources, the introduction of contextual rule deteriorates the recognition rate. We discuss the possible factors explaining this surprising result and outline the problems of defining a set of contextual phonological rules and integrating them in the search algorithm.

Volume 4, page 2293 Session E11

Automatic Learning of Finite State Automata for Pronunciation Modeling

Pastor-i-Gadea M, Casacuberta F
Universitat Politècnica de València, Spain

The great variability of word pronunciations in spontaneous speech is one of the reasons for the low performance of present speech recognition systems. The generation of dictionaries that take into account this variability can increase the robustness of such systems. A word pronunciation is a possible phone sequence that can appear in a real utterance, and represents a possible acoustic realization of the word. Here, word pronunciations are modeled using finite state automata. The use of such models allow for the application of grammatical inference methods and an easy integration with the others sources of acknowledge. The training samples are obtained from the alignment between the phone decodification of each training utterance and the corresponding canonical transcription. Models proposed in this work were applied in a translation-oriented speech task. The improvements achieved by these models were in the range betwen 2.7 to 0.6 points depending on the language model used.

Volume 4, page 2297 Session E11

Session E12 - Oral
Friday - 09.00 - 10.40

Speech Coding: Wideband Speech Coding

Chair: Bastiaan Kleijn, KTH, Stockholm, Sweden

AMR Wideband codec - leap in mobile communication voice quality

Rotola-Pukkila J[1], Vainio J[1], Mikkola H[1], Järvinen K[1], Bessette B[2], Lefebvre R[2], Salami R[3], Jelinek M[2]
[1]Nokia Research Center, Finland, [2]University of Sherbrooke, Canada, [3]VoiceAge Corporation, Canada

The Third Generation Partnership Project (3GPP) and European Telecommunications Standards Institute (ETSI) have carried out development and standardisation of a wideband speech codec for GSM and the third generation mobile communication WCDMA system since 1999. The Adaptive Multi-Rate Wideband (AMR-WB) codec algorithm was selected in December 2000, and the corresponding specifications were approved in March 2001. The AMR-WB codec was jointly developed by Nokia and VoiceAge. AMR-WB extends the audio bandwidth from 3.4 kHz to 7 kHz and gives superior speech quality and voice naturalness compared to existing 2nd and 3rd generation mobile communication systems. The wideband speech service provided by the AMR-WB codec will give mobile communication speech quality that even exceeds (narrowband) wireline quality.

Volume 4, page 2303 Session E12

Combined Speech and Audio Coding with Bit Rate and Bandwidth Scalability

Farrugia M, Kondoz A M
University of Surrey, UK

The growing demand for streaming multimedia services over the Internet and recently also over mobile networks has initiated a great interest in coding algorithms which are able to adapt to different transmission environments and to operate under multiple constraints of bit rate, complexity, delay, robustness to bit errors and diversity of input signals. In the light of these recent developments, we present a novel scalable representation for speech and audio signals with low delay. The algorithm operates in four modes, each based on backward-adaptive linear predictive coding (BA LPC). The first mode is referred to as the base-line narrowband (0--4kHz) coder. Wideband speech and audio signals (0--8kHz) are efficiently represented by the second mode which employs a QMF to split the spectrum into two equal bands. The remaining two modes use a two-stage QMF structure to decompose the bandwidth of 32kHz sampled signals into four bands. Scalability is achieved by means of discrete quantisation layers representing various levels of enhancements for each band and also flexibility in terms of complexity and bit allocation requirements depending on the particular application and on the network resources. The resulting bit rates range from 12 to 64kb/s. The performance of the coder is evaluated by comparing it to MPEG and ITU standards.

Volume 4, page 2307 Session E12

Joint Speech and Audio Coding Combining Sinusoidal Modeling and Wavelet Packets

Fék M[1], Várkonyi-Kóczy A R[1], Boucher J-M[2]
[1]Budapest University of Technology and Economics, Hungary, [2]ENST de Bretagne, France

This paper presents a joint speech and audio coding algorithm combining sinusoidal modeling and a perceptually adapted Wavelet Packet Transform (WPT). The input signal is limited to the band of 50-7000 Hz, and sampled at 16 kHz. The sinusoidal modeling uses a Sinusoidal Similarity Measure (SSM) to find stable sinusoidal components. A novel pitch harmonics based encoding is applied to encode the sinusoidal frequencies. The residual is obtained by extracting the re-synthesized sinusoids from the input, and is processed by a WPT simulating the critical bands of the Human Auditory System. Perceptual Noise Substitution (PNS) is applied in noisy WPT sub-bands to reduce the bit rate. The method provides nearly transparent quality for both speech and audio inputs. The mean bit rate of the compressed signal varies between 32-62 kbps depending on the input.

Volume 4, page 2311 Session E12

Temporal Decomposition: A Promising Approach to Low Rate Wideband Speech Compression

Ritz C H, Burnett I S
University of Wollongong, Australia

In this paper, we present new results on Temporal Decomposition (TD) applied to the Line Spectral Frequencies (LSFs) derived for wideband speech. The paper shows that by incorporating a dynamic programming search algorithm into TD, near transparent quantisation of wideband LSFs can be obtained at approximately 1 kbps. We also show that TD performs significantly better than Split Vector Quantisation at low bit rates. We propose that TD is a promising approach to low rate wideband speech coding for applications such as unicast streaming.

Volume 4, page 2315 Session E12

Wideband LSF Quantization by Generalized Voronoi Codes

Ragot S, Lahdili H, Lefebvre R
Univ. Sherbrooke, Canada

Presented a method for quantizing the wideband line spectrum frequencies (LSF) with a specific class of near-ellipsoidal lattice codes referred to as ``generalized Voronoi codes". Optimization procedures are described with respect to a weighted mean-square error (WMSE). The lattices D16, RE16 or RLambda16 are applied to quantize the LSF with no frequency splitting. Results indicate that near-ellipsoidal lattice quantization allows to develop efficient one-stage algebraic wideband LSF quantization at competitive bit rates.

Volume 4, page 2319 Session E12

Session E13 - Oral
Friday - 09.00 - 10.40

Dialogue Systems: Techniques and Strategies

Chair: Arne Jönsson, IDA, Linkøping, Sweden

Learning of User Formulations for Business Listings in Automatic Directory Assistance

Popovici C[1], Andorno M[1], Laface P[1], Fissore L[2], Nigra M[2], Vair C[2]
[1]*Politecnico di Torino, Italy*, [2]*Loquendo, Italy*

Automatic Directory Assistance (DA) for business listings poses many application specific problems. One of the main problem is that customers formulate their requests for the same listing with great variability. We present the results of a study aiming at automatic learning, from field data, of expressions typically used by customers to formulate their requests for the most frequent business listings. We use a clustering procedure that exploits the association of the phonetic string produced by a lexical unconstrained search for a given denomination pronounced by the user and the phone number provided by the system or by the human operator, in case of failure of the automatic DA service. We show that an unsupervised approach allows to detect user formulations that were not foreseen by the designers, and that can be added, as variants, to the denominations already included in the system to reduce its failures.

Volume 4, page 2325 Session E13

Mathematical modeling of Spoken Human - Machine dialogues including erroneous confirmations

Louloudis D[1], Tsopanoglou A[2], Fakotakis N[1], Kokkinakis G[1]
[1]*University of Patras, Greece*, [2]*Knowledge S.A., Greece*

In this paper, we present a method that enables the designer to investigate a spoken dialogue system's performance by employing diagnostic evaluation during the initial phases of the system's development. Results from glass box assessment (e.g.. recognition success rate), combined with the dialogue strategy in the proposed mathematical model, can be used to predict the system's performance. The model incorporates erroneous confirmations by the user, which affect the overall performance.

Volume 4, page 2329 Session E13

Limited Enquiry Negotiation Dialogues

Lewin I
Netdecisions Ltd, UK

We define a new dialogue management strategy Limited Enquiry Negotiation Dialogues designed for enabling simple man-machine dialogues in which the parameters (for which the user will supply values) of a query to a database are negotiated. The choice of which query to make next is also not pre-ordained. The strategy is simple and intuitive but permits interestingly complex dialogue behaviour. We propose it as an addition to a dialogue designer's standard components toolbox along with other well-known ideas such as menu-traversal and slot-filling. We illustrate the strategy by examining how it accounts for interesting but by no means rare data in a Wizard of Oz corpus of business trip planning dialogues. Finally, we discuss some more theoretical issues arising from the model.

Volume 4, page 2333 Session E13

A Comparison of some Different Techniques for Vector Based Call-Routing

Cox S[1], Shahshahani B[2]
[1]*University of East Anglia, UK*, [2]*Nuance Communications, USA*

Two approaches to vector-based call-routing are described, one based on matching queries to routes and the other on matching queries directly to stored queries. We argue that there are some problems with the former approach, both when used directly and when latent semantic analysis (LSA) is used to reduce the dimensionality of the vectors. However, the second approach imposes a higher computational load than the first and we have experimented with reducing the number of reference vectors (using the multi-edit and condense algorithm) and the dimensionality of the vectors (using linear discriminant analysis (LDA)). Results are presented for the task of routing queries on banking and financial services to one of thirty-two destinations. Best results (5.1% routing error) were obtained by first using LSA to smooth the query vectors followed by LDA to increase discrimination and reduce vector dimensionality.

Volume 4, page 2337 Session E13

Architecture for adaptive multimodal dialog systems based on VoiceXML

Niklfeld G[1], Finan R[2], Pucher M[1]
[1]*Telecommunications Research Center Vienna, Austria*, [2]*Mobilkom Austria AG, Austria*

This paper describes application oriented research on architectural building blocks and constraints for adaptive multimodal dialog systems that use VoiceXML as a component technology. The VoiceXML standard is well supported and promises to make the development of speech-enabled applications so easy that everyone with general web programming skills can accomplish it. The paper proposes a software architecture for multimodal interfaces that emphasizes modularity, in order to strengthen this potential by clearly separating different types of development tasks in a multimodal dialog system. The paper argues that adaptivity is a central design concern for multimodal dialog systems, but that adaptivity is not facilitated by the current VoiceXML standard. This and other examined limitations of VoiceXML for building multimodal dialog systems should be repaired in upcoming work on a successor standard that will explicitly target multimodal applications.

Volume 4, page 2341 Session E13

Session E14 - Oral
Friday - 09.00 - 10.40

Speech Recognition and Understanding: Robust ASR

Chair: John H.L. Hansen, RSPL-CSLR / Univ. of Colorado at Boulder, USA

Separating speaker and environment variabilities for improved recognition in non-stationary conditions

Rigazio L, Nguyen P, Kryze D, Junqua J-C
PSTL, USA

In this paper we address the problem of speaker adaptation in noisy environments. We estimate speaker adapted models from noisy data by combining unsupervised speaker adaptation with noise compensation. We aim at using the resulting speaker adapted models in environments that differ from the adaptation environment, without a significant loss in performance. The key idea is to separate speaker and environment variabilities and associate them to independent models. We show that linear models for both speaker and environment are critical for achieving this goal. Experiments for 2000 and 4000 isolated word tasks on real car noise show that unsupervised speaker adaptation combined with noise compensation can provide more than 20% error rate reduction compared with noise compensation only, and more than 50% error rate reduction compared with speaker adaptation only.

Volume 4, page 2347 Session E14

Robust speech recognition techniques applied to a speech in noise task

Rose R C[1], Kim H K[1], Hindle D[2]
[1]AT&T Labs - Research, USA, [2]AnswerLogic Inc., USA

This paper describes the design and evaluation of an automatic speech recognition (ASR) system on the Naval Research Laboratory Speech In Noise (SPINE) speech corpus. This corpus represents a task which involves human-human interaction on a constrained problem solving scenario under six different simulated noisy environments. Acoustic and language modeling were performed using a dataset taken entirely from a subset of the acoustic environments. Speech recognition was performed on continuous conversations by detecting speech utterances, performing acoustic feature analysis and normalization, and adapting HMM models in multiple passes over each conversation-side. The ASR word accuracy (WAC) ranged from 77 percent in an office environment to 61 percent in conditions that include significant levels of background speech and noise. An overall average WAC of 69.0 percent was obtained across all noise conditions.

Volume 4, page 2351 Session E14

Minimax Classification With Parametric neighborhoods For Noisy Speech Recognition

Afify M, Siohan O, Lee C-H
Bell Laboratories, Lucent Technologies, USA

In this paper we derive upper and lower bounds on the mean of speech signals corrupted by additive noise. The bounds are derived in the log spectral domain. Approximate bounds on the first and second order time derivatives are also developed. It is then shown how to transform these bounds to the MFCC domain to be used by conventional cepstrum-based speech recognizers. The proposed bounds define the mismatch neighborhood for minimax classification. Speech recognition experiments, using artificially added noise, and a real-life mismatch scenario, illustrate that this parametric neighborhood works quite well in practice. We also beleive that the proposed bounds will find various applications in noisy speech recognition.

Volume 4, page 2355 Session E14

Maximum Likelihood Non-linear Transformation for Environment Adaptation in Speech Recognition

Padmanabhan M, Dharanipragada S
IBM T. J. Watson Research Center, USA

In this paper, we describe an adaptation method for speech recognition systems that is based on a piecewise-linear approximation to a non-linear transformation of the feature space. The method extends a previously proposed non-linear transformation (NLT) technique by making the transformation function more sophisticated and by computing the transformation to maximize the likelihood of the adaptation data given its transcription. This method also differs from other linear techniques (such as MLLR, linear feature space transforms, etc.) in two ways - first, the computed transformation is non-linear, second, the tying structure of the transformation depends not on the phonetic class but rather on the location in the feature space. Experimental results show that the method performs well for the case of limited adaptation data, and the performance gains appear to be additive to those provided by MLLR - yielding upto 3.4% relative improvement over MLLR.

Volume 4, page 2359 Session E14

A Study of Speech Coding Parameters in Speech Recognition

Turunen J J[1], Vlaj D[2]
[1]Tampere University of Technology, Finland, [2]University of Maribor, Slovenia

Speech recognition over different transmission channels will set demands to the parametric encoded/decoded speech. The effects of different types of noise have been studied a lot and the effects of the parameterization process in speech has been known to cause degradation in decoded speech when compared to the original speech. But does the encoding/decoding process modify the speech so much that it will cause degradation in the speech recognition result? If it does what may cause the speech recognition degradation? We have studied the effect of the parameterization and the causes of the nine different codec configurations to isolated word recognition.

Volume 4, page 2363 Session E14

Session E15 - Poster
Friday - 09.00 - 10.40

Applications: Miscellaneous Applications

Chair: George Kokkinakis, Univ. of Patras, Greece

Some Practical Considerations in the Deployment of a Wireless-Communication Interactive Voice Response System

Garcia-Mateo C, Docio-Fernandez L, Cardenal-Lopez A
University of Vigo, Spain

In this paper, we describe the design procedure for a wireless communication interactive voice response (IVR) system. The application must work in a very noisy environment which has imposed many design constraints. We will address the sensible aspects of three components of the application: the voice activity detector (VAD), the automatic speech recognition (ASR) system, and the confidence measure (CM) determination. In order to get a satisfactory product, it has been necessary to reduce the important mismatch between available linguistic and acoustic resources and the operational environment. Adaptation techniques for the acoustic models of the speech recognition system have proven to be effective to speed up the application deployment time.

Volume 4, page 2369 Session E15

Caller Identification for the SCANMail Voicemail Browser

Rosenberg A[1], Hirschberg J[1], Bacchiani M[1], Parthasarathy S[1], Isenhour P[2], Stead L[1]
[1]AT&T Labs-Research, USA, [2]Virginia Tech, USA

SCANMail is a prototype system developed to provide useful tools for managing and searching through voicemail messages. Content is extracted from voicemail messages using various speech and text processing tools. One such content category is the identity of the message caller. This paper describes CallerID, the sever tool attached to SCANMail to provide caller labels for voicemail messages. CallerID makes use of text independent speaker recognition techniques. Two kinds of requests are handled by the CallerID server. A request triggered by the arrival of a new voicemail message results in the processing of the message to score it agains the models of callers assigned to the user (recipient) in order to propose the identity of the caller. A second request is initiated by a user who provides a caller label for a message he/she has reviewed. CallerID processes the message and uses it to train or adapt a speaker model for the caller.

Volume 4, page 2373 Session E15

Extractive Summarization of Voicemail using Lexical and Prosodic Feature Subset Selection

Koumpis K, Renals S, Niranjan M
University of Sheffield, UK

This paper presents a novel data-driven approach to summarizing spoken audio transcripts utilizing lexical and prosodic features. The former are obtained from a speech recognizer and the latter are extracted automatically from speech waveforms. We employ a feature subset selection algorithm, based on ROC curves, which examines different combinations of features at different target operating conditions. The approach is evaluated on the IBM Voicemail corpus, demonstrating that it is possible and desirable to avoid complete commitment to a single best classifier or feature set.

Volume 4, page 2377 Session E15

Business Listings in Automatic Directory Assistance

Scharenborg O, Sturm J, Boves L
University of Nijmegen, the Netherlands

So far most attempts to automate Directory Assistance services focused on private listings, because it is not known precisely how callers will refer to a business listings. The research described in this paper, carried out in the SMADA project, tries to fill this gap. The aim of the research is to model the expressions people use when referring to a business listing by means of rules, in order to automatically create a vocabulary, which can be part of an automated DA service. In this paper a rule-base procedure is proposed, which derives rules from the expressions people use. These rules are then used to automatically create expressions from directory listings. Two categories of businesses, viz. hospitals and the hotel and catering industry, are used to explain this procedure. Results for these two categories are used to discuss the problem of the over- and undergeneration of expressions.

Volume 4, page 2381 Session E15

EuTrans: a Speech-to-Speech Translator Prototype

Pastor-i-Gadea M, Sanchis A, Casacuberta F, Vidal E
Institut Tecnològic d'Informàtica Universitat Politècnica de València, Spain

EuTrans system is a telephone speech input translation prototype capable of translating telephone calls from one language to another. It assumes a human to human communication, each one speaking a different language, assisted by a system with translation capabilities. The prototype has been developed as a demonstrator for the European project with the same name. EuTrans achieves a response time close to real time for speaker-independent, medium complexity tasks (a few thousand words) and offers competitive accuracy. The acoustic, language and translation models are finite-state networks that are automatically learnt form training samples, this makes the system easily adaptable to news tasks. It runs on a standard PC with audio capability and a cheap modem. The system is currently available for two translation tasks: FUB task (Italian-English) and Traveler task (Spanish-English).

Volume 4, page 2385 Session E15

Speech Recognition over NetMeeting Connections

Metze F, McDonough J, Soltau H
University of Karlsruhe, Germany

In this paper we evaluate the performance of the ISL's German Verbmobil spontaneous speech recognizer on the Nespole! database. In this task, people talk to an agent in a tourist office to plan their holidays via a NetMeeting connection, also sharing screen contents (web-pages). Stereo recordings were made both before and after speech transmission over an IP connection using the G.711 codec, so that we are able to directly measure the loss in LVCSR performance due to NetMeeting's segmentation and compression. The aim of this work is to quantify this loss, which is a consequence of using protocols which were not designed for speech recognition purposes. We report on techniques employed to port our existing clean-speech recognizer to this new data quality, using about 1.5h of labeled adaptation data, but avoiding a complete retraining of the system.

Volume 4, page 2389 Session E15

DIARCA: A Component Approach to Voice Recognition

Díaz Martín J C[1], García Zapata J L[1], Rodríguez García J M[1], Álvarez Salgado J F[1], Espada Bueno P[1], Gómez Vilda P[2]
[1]Universidad de Extremadura, Spain, [2]Universidad de Politécnica de Madrid, Spain

Current voice recognition systems tend to be implemented as a PC desktop facility. This model is not suitable for the growing complexities of present and future developments: It is single-user, it is non portable,

and it assumes the workstation model, where all the CPU resources are supposed to be locally available. This work researches how a high performance speech recognition system can be redesigned and implemented as a time-critical network service shared through ordinary data transmission media with three main design goals: Scalability, predictability and POSIX portability. The whole idea has been tested by rebuilding IVORY, a well known robust desktop voice recognition methodology, as a distributed component.

Volume 4, page 2393 Session E15

The mVprotek : m-commerce Voice verification system

Kyung Y[1], Jung J[2], Sohn S M[1], Chun H J[1], Moon S Y[2], Kim M H[1], Sull W H[1]
[1]SK Telecom Platform R&D Center, Korea, [2]Infinity Telecom ICN Lab., Korea

In this paper, we developed speaker verification system for m-commerce (mobile commerce) via wireless internet and WAP. We implemented the system as client-server architecture. The clients are mobile phone simulator and PDA. As the needs for wireless Internet service is increasing, the needs for secure m-commerce is also increasing. Conventional security technique are reinforced by biometric security technique. This paper utilized the voice as biometric security techniques. The verification results are obtained by integrating the mVprotek system with SK Telecom's CDMA system. Utilizing F-Ratio and Virtual cohort model normalization showed much better performance than conventional background model normalization technique.

Volume 4, page 2397 Session E15

Real-time Multilingual Communication by means of Prestored Conversational Units

Alm N[1], Iwabuchi M[2], Andreasen P N[1], Nakamura K[3], Murray I R[1]
[1]University of Dundee, UK, [2]Hiroshima University, Japan, [3]Kagawa University, Japan

A computer mediated communication system has been developed which can offer real time multilingual communication, as long as users stay within the boundaries of prestored conversational units. The system was designed originally to give non-speaking people a multi-lingual capability. However, the system could also be used by people whose only communication disadvantage is not being able to speak a foreign language. It is based on research into conversational modelling and utterance prediction, making use of prestored material. In comparison with a multi-lingual phrase book, the system helped users to have more natural conversation, and to take more control of the interaction. This project is an interesting example of the way in which systems developed for people with severe disabilities can often have useful general applications.

Volume 4, page 2401 Session E15

Writing script-based dialogues for AAC

Murray I R, Arnott J L, Alm N, Dye R, Harper G
University of Dundee, UK

AAC (Augmentative and Alternative Communication) devices are often used by disabled people who are non-speaking, in order to assist them to communicate. However, many such systems require the user to build up an utterance word-by-word each time, and are thus often laborious for the user and slow (and thus less effective) in communication. For this reason, an AAC system was developed which relies on pre-stored scripts and an engaging user interface to predict and guide the user through many standard dialogue situations with a minimum of effort. In order to allow individuality in using the system, an authoring package has been developed. This allows users (or their carers) to modify existing scripts or to add new scripts into the system, and has also been designed to facilitate exchange of scripts between users. The systems developed utilise speech synthesisers for output, and are commercially available in English, Dutch and German versions.

Volume 4, page 2405 Session E15

Communication Aid for non-vocal people using corpus-based concatenative speech synthesis

Iida A[1], Sakurada Y[1], Campbell N[2], Yasumura M[1]
[1]Keio University, Japan, [2]ATR International, Japan

This paper reports on the development of Chatako-AID, a communication aid for non-vocal people using corpus-based cocatenative speech synthesis by creating a speech corpus especially designed for such use. The concept of Chatako-AID; synthesis with the user's voice, which makes use of precomposed texts, is highly appreciated by the target user. This confirms that the recording of a minimum set of phonetically balanced sentences is insufficient for speech synthesis in the proposed method and that a combination of the above recording and a recording of well-read continuous-text material produces more natural sounded synthesised speech.

Volume 4, page 2409 Session E15

Social Effects on Vocal Rate with Echoic mimicry Using Prosody-only Voice

Suzuki N[1], Kakehi K[2], Takeuchi Y[3], Okada M[3]
[1]ATR MI&C / Nagoya University, Japan, [2]Nagoya University, Japan, [3]ATR MI&C, Japan

We have been studying some of the essential factors that constitute interpersonal relations between humans and computers by focusing on social bonding in proto-communications (the interaction between adults and human infants or pets). From this viewpoint, this paper presents psychological experiments on the interaction between humans and animated characters that mimic the human voice at the prosodic level using prosody-only voice under different vocal rate of character's voice: (a) faster than normal, (b) normal speed, and (c) slower than normal. We examine the subjects' impression towards animated characters with the above conditions of their voice using post-questionnaire and analyse the change of the speech rate of subjects. The results indicate that most humans may prefer an animated character with a faster voice to that with a slower voice. Moreover, the speech rate of humans changes to opposite of vocal rate of animated characters' voice. I.e. the speech rate of subjects becomes slower when the voice of character becomes faster, and vice versa.

Volume 4, page 2413 Session E15

Everyday Life Sounds and Speech Analysis for a Medical Telemonitoring System

Castelli E, Istrate D
CLIPS/IMAG, France

In order to improve patients' life conditions and to reduce the costs of the long hospitalization, the medicine is more and more interested in the telemonitoring techniques. These will allow the old people or the high risk patients to stay at home, and to benefit from a remotely and automated medical supervision. We develop in collaboration with TIMC-IMAG laboratory, a system of telemonitoring in a habitat equipped with physiological sensors, position encoders of the person, and microphones. The originality of our approach consists in replacing the video camera monitoring, hardly accepted by the patients by microphones recording the sounds (speech or noises) in the apartment. The microphones carry out a multichannel sound acquisition system which, thanks to the sound information coupled with physical information, will enable us to identify a situation of distress. We describe the practical solutions chosen for the acquisition system and the recorded corpus of situations.

Volume 4, page 2417 Session E15

Speaking While Driving - Preliminary Results on Spellings in the German SpeechDat-Car Database

Draxler C[1], Bengler K[2]
[1]*University of Munich, Germany*, [2]*BMW Group, Munich, Germany*

Abstract Voice-operated devices are of particular interest in mobile environments, e.g. vehicles. They promise a natural and intuitive interface to devices and services, and they offer hands-free operation, a legal prerequisite for in-car usage in many European countries. Spelling is a common task for the operation of voice operated devices, especially under unfavorable communication conditions. This paper presents a first analysis of the error and fluency rate for 4502 utterances from the German SpeechDat-Car database. The error rate was found to be between 1.7% and 4.4% for the spelling of natural items, and between 3.6% and 7.9% for artificial letter sequences. Only 3.6% of the utterances contained hesitations. These results suggest that spelling while driving might be a suitable means of fallback interaction if specific error recovery mechanisms are implemented.

Volume 4, page 2421 Session E15

Session E16 - Poster
Friday - 09.00 - 10.40

Signal Analysis: Pitch and Speech Analysis

Chair: Paavo Alku, HUT, Finland

Efficient Periodicity Extraction Based on Sine-Wave Representation and its Application to Pitch Determination of Speech Signals

Chazan D[1], Tzur M[2], Hoory R[1], Cohen G[1]
[1]*IBM Research, Israel*, [2]*BIT Innovation Technologies, Israel*

This paper presents a novel low-complexity method for extracting periodicity of signals based on their sine-wave representation. In this representation, the signal is modeled as a finite sum of sine-waves, with time-varying amplitudes, phases and frequencies. We describe how one can modify the familiar spectral-comb analysis method to obtain a guaranteed and effective procedure to find the fundamental-frequency which gives the best harmonic approximation of the signal spectrum. The search is efficiently carried out in the frequency domain. The procedure obtains a successive refinement of possible pitch values which are consistent with an increasing number of sine wave components. Other pitch intervals are pruned at an early stage of the search. The advantage of this algorithm is its high accuracy achieved at a relatively low complexity. We also briefly describe one possible application in the area of pitch determination of speech signals.

Volume 4, page 2427 Session E16

Viseme Recognition Using Multiple Feature Matching

Shdaifat I[1], Grigat R -R[1], Luetgert S[2]
[1]*TU Hamburg Harburg, Vision Systems, Germany*, [2]*Philips Semiconductors, Systems Laboratory Hamburg, Germany*

In this paper, we present a technique for the extraction of the five main visemes produced in natural speech for German. The method belongs to the LDA (Linear Discriminant Analysis) family. The intensity, the edges, and the line segments are used to locate the lips automatically and for viseme classification. Using many features in the recognition maximizes the probability of recognition rate. The corners of the mouth are used in case of small rotation and scale. An experiment has been carried out on different people, to understand the part of the speech that the human being use. The people grouped the phonemes into five different visemes. The number of distinguished visemes is not the same for each speaker. Everyone express the speech in a different visemes. Good recognition rate has been achieved on different speaker.

Volume 4, page 2431 Session E16

The fundamental frequency of cough by autocorrelation analysis

Van Hirtum A, Berckmans D
KULeuven, Belgium

The presented research evaluates the quantitative characterization of human cough sounds by estimating the fundamental frequency or pitch. The fundamental frequency was determined by autocorrelation analysis on both the rough time-signal and the linear predicted time-signal. Differences between `spontaneous' and `voluntary' cough sounds are put forward. The experimental cough database was registered in the free acoustical field on respectively 3 pathological and 9 healthy non-smoking subjects.

Volume 4, page 2435 Session E16

A Fundamental Frequency Estimation Method for Noisy Speech Based on Instantaneous Amplitude and Frequency

Ishimoto Y[1], Unoki M[2], Akagi M[1]
[1]*Japan Advanced Institute of Science and Technology, Japan*,
[2]*University of Cambridge, UK*

This paper proposes a robust and accurate F0 estimation method for noisy speech. This method uses two different principles: (1) an F0 estimation based on periodicity and harmonicity of instantaneous amplitude for a robust estimation in noisy environments, and (2) an F0 estimation based on stability of instantaneous frequency as an accurate estimation method. The proposed method also uses a comb filter with controllable pass-bands to combine the two estimation methods. Simulations were carried out to estimate F0s from real speech in noisy environments and to compare the proposed method with other methods. The results showed that this method can not only estimate F0s for clean speech with similar accuracy as the method using only instantaneous frequency but also robustly estimate F0s from noisy speech in comparison with the other methods such as the cepstrum method.

Volume 4, page 2439 Session E16

Robust LP Analysis Using Glottal Source HMM with Application to High-Pitched and Noise Corrupted Speech

Sasou A, Tanaka K
National Institute of Advanced Industrial Science and Technology, Japan

This paper presents a robust feature extraction method effective to speech signal with high fundamental frequency and/or corrupted by additive white noise. The method represents the glottal source wave using HMM in order to model the non-stationary properties. The nodes of HMM are concatenated in a ring state to represent the periodicity of voiced sounds. The method can accurately extract glottal source wave and vocal tract characteristics from speech signals even in high fundamental frequency as ranging up to 750Hz. From identification theory, estimation of vocal tract characteristics from speech corrupted by additive noise requires glottal source wave that can not be observed directly, so that it needs to be estimated. Therefore, estimation accuracy of vocal tract characteristics highly depends on the estimation accuracy of glottal source wave. We apply the glottal source HMM to extracting the glottal source wave from corrupted speech, and confirmed the feasibility of the method.

Volume 4, page 2443 Session E16

Fast Harmonic Estimation Using a Low Resolution Pitch for Low Bit Rate Harmonic Coding

Choi Y-S[1], Youn D-H[2]
[1]*LG Electronics Inc., Korea*, [2]*Yonsei University, Korea*

This paper describes a fast harmonic estimation, referred to Delta Adjustment (DA), using a low resolution pitch. The presented DA method is based on modification of the Generalized Dual Excitation (GDE) technique [1] which was proposed to improve speech enhancement performance. We introduce the GDE technique and modify it to be suitable for low bit rate harmonic coding that uses only an integer pitch estimate. Unlike the GDE, the DA matches a frequency-warped version of the original spectrum that conforms to a fixed pitch at all harmonic bands. In addition, complexity and performance of the presented method are described in comparison with those of the conventional Fractional Pitch (FP) based harmonic estimation. Experimental results showed that the DA algorithm significantly reduces the complexity of the FP method while maintaining the performance.

Volume 4, page 2447 Session E16

Comparative evaluation of F0 estimation algorithms

de Cheveigné A[1], Kawahara H[2]
[1]*Ircam - CNRS, France*, [2]*CREST - Wakayama University, Japan*

This paper reports the comparative evaluation of several speech F0 evaluation algorithms over a wide database of laryngograph-labeled speech. Included are several classic algorithms that are available in software on the net, as well as two new algorithms that offer greatly reduced error rates. Particular attention is given to the methodology of evaluation.

Volume 4, page 2451 Session E16

Identification of Accent and Intonation in sentences for CALL systems

Ishi C T, Minematsu N, Nishide R, Hirose K
University of Tokyo, Japan

In order to construct a CALL (Computer Aided Language Learning) system that can teach learners accent and intonation of Japanese, it's necessary to automatically identify accent types and intonation types in sentence utterances. For this purpose, several acoustic (prosodic) features of speech were investigated taking their effects on human perception into account. For the accent type identification method, the use of average values of F0 in mora and target values of F0 in mora final was evaluated in CV and VC units. Average values of VC units and target values of CV units showed better performance in the identification task. As for the intonation identification, several acoustic features were investigated to represent 6 types of sentence final tones, each conveying different information of intention and perceptual impression. The proposed acoustic features for relative duration and sentence final pitch change showed good correspondence to perceptual features.

Volume 4, page 2455 Session E16

Systematic F0 Glitches around Nasal-Vowel Transitions

Kawahara H[1], Zolfaghari P[2]
[1]*Wakayama University / ATR-ISD / CREST, Japan*, [2]*CIAIR, Nagoya University, Japan*

High-resolution F0 analysis using a speech database with simultaneously recorded EGG (Electroglottogram) signals indicated that there are systematic F0 glitches around nasal-vowel transitions. The durations of the glitches are 10 to 20 ms and they introduce 5 to 10 Hz F0 shifts. A detailed series of analyses of these glitches indicated that the major contributing factor of these glitches is sudden changes of group delay values of the vocal tract transfer function in the vicinity of the fundamental frequency at nasal-vowel transitions. It is also suggested that the Doppler effects due to apparent changes of vocal tract length are marginal, even if they exist. Finally, issues in evaluating high resolution F0 extraction algorithms and applications to high quality speech manipulation methods are discussed.

Volume 4, page 2459 Session E16

Using Aerial and Geometric Features in Automatic Lip-reading

Wojdel J C, Rothkrantz L J M
Delft University of Technology, the Netherlands

In this paper we present the lip-reading experiments with different sets of the features extracted from the video sequence. In our experiments we use a simple color based filtering techniques to extract the feature vectors from the incoming video signal. Some of those features are directly related to the geometrical properties of the lips (their position and visible thickness). Other features represent the information that relates to the visibility of the other components of the speech production system. The visibility of the teeth and vocal tract for example is described by means

of the area they occupy in the image, we call them therefore the aerial features.

Volume 4, page 2463 Session E16

Inverse Filtering of Tube Models with Frequency Dependent Tube Terminations

Schnell K, Lacroix A
Johann Wolfgang-Goethe Universitaet, Germany

The tube model, realized by lattice filters in discrete time, can be used to describe the propagation of plane sound waves through the vocal tract. The tube model which is treated in this contribution contains two prescribed terminations. One for the lip opening and one for the constriction at the glottis. These two terminations are frequency dependent. To estimate the parameters of this tube model, standard algorithms like the Burg-method and related methods are not applicable. Therefore a procedure is proposed to estimate the parameters of these tube models in an adequate way. The procedure is based on inverse filtering, which is carried out iteratively. The analysis of consonants shows, that the corresponding constrictions in the vocal tract area functions can be observed. Using these estimated constrictions it is possible to synthesize VCV transitions too, which implies the typical formant movements.

Volume 4, page 2467 Session E16

Formant Estimation using Gammachirp Filterbank

Ouni K, Lachiri Z, Ellouze N
LSTS, ENIT, Tunisie

This paper proposes a new method for formants estimation using a decomposition of speech signal in Gammachirp functions base. It is a spectral analysis method performed by a gammachirp filterbank. A similar approach to the modified spectrum estimation, which allows a smooth and an average spectrum is adopted. In fact, instead of using an uniform window commonly used in short-time Fourier analysis, a bank of gammachirp filters is applied on the signal. A temporal average of the estimated spectra is then applied to obtain one spectrum highlighting the formants structure. This approach is validated by its application on synthesized vowels. The formants are detected with good estimation in comparison with the values given in synthesis. In the same way, this analysis is applied on natural vowels. All the results are compared to three traditional methods, LPC, cepstral and spectral one's and also to a same analysis given by a gammatone filterbank. The tracking of formants shows that this method, which based on gammachirp filters, gives a correct estimation of the formants compared to traditional methods.

Volume 4, page 2471 Session E16

Autoregressive Time-Frequency Interpolation in the Context of Missing Data Theory for Impulsive Noise Compensation

Potamitis I, Fakotakis N
University of Patras, Greece

The present paper reports on a novel technique for the identification and replacement of spectral coefficients degraded by impulsive noise. The problem is viewed from the perspective of Missing Feature Theory (MFT). The analysis is carried out in the linear spectrum prior to, or after applying the mel-scale filter-bank depending on whether one aims at improving the quality of perception of speech recordings or at Automatic Speech Recognition (ASR). Each filter-bank output is considered to be a time series drawn from an Auto-Regressive process (AR). A validation corpus of undistorted recordings is used to derive a-priori bounds on the expected prediction error of each AR model. In operational conditions, the prediction procedure is monitored and the violation of the statistical bounds indicates band corruption and entails the substitution of the degraded spectral coefficients by the prediction of the corresponding AR model. ASR experiments and informal listening tests demonstrate large improvement in terms of word recognition performance and Itakura-Saito divergence at very low Signal to Impulsive Noise Ratios (SINRs). Data, and implementation code can be found at: ftp://wcl.ee.upatras.gr/

Volume 4, page 2475 Session E16

Analysis of the Voiced Speech using the Generalized Fourier Transform with Quadratic Phase

Petrinovic D, Cuperman V
University of California Santa Barbara, USA

One significant problem with sinusoidal modeling of the speech signal is due to the use of standard Fourier Transform for a quasi-periodic signal. The analysis accuracy is severely limited by the lack of stationarity of the analyzed segment, since the analysis is based on the conventional Fourier Transform. An improved analysis technique based on the Generalized Fourier Transform (GFT) with quadratic phase will be discussed in this paper. Speech signal is modeled as a sum of harmonic cosines but with nonlinear phases. A technique for estimation of the time-varying model parameters from the GFT spectrum is proposed. It will be shown that the modeling gain can be improved significantly by inclusion of a single additional parameter in the analysis procedure.

Volume 4, page 2479 Session E16

Session E21 - Oral
Friday - 11.10 - 12.30

ESE7 - Integration of Phonetic Knowledge in Speech Technology: Is Phonetic Knowledge any use? Panel discussion

Chair: Bill Barry, Universität des Saarlandes, Germany

From Here to Utility - Melding Phonetic Insight with Speech Technology

Greenberg S
International Computer Science Institute, USA

An historic tension exists between science and technology with respect to spoken language. Over the coming decades this tension is likely to dissolve into a collaborative relationship melding linguistic knowledge with machine-learning and statistical methods as a means of developing mature science and technology pertaining to human-machine communication. In the process many mysteries surrounding the form and substance of spoken language are likely to be solved through the concerted efforts of scientists and engineers focused on the creation of “flawless” speech technology.

Volume 4, page 2485 Session E21

Session E22 - Oral
Friday - 11.10 - 12.30

Speech Coding: Speech Transmission Systems

Chair: Isabel Trancoso, INESC, Lisboa, Portugal

Speech Quality Measure for VoIP using Wavelet based Bark Coherence Function

Park S-W, Park Y-C, Youn D-H
Yonsei University, Korea

The Bark Coherence Function (BCF) defines a coherence function with loudness speech as a new cognition module, robust to linear distortions due to the analog interface of digital mobile system. Preliminary experiments have shown the superiority of BCF over current measures. In this paper, a new BCF suitable for VoIP is developed. The new BCF is based on the wavelet series expansion that provides good frequency resolution while keeping good time locality. The proposed Wavelet based Bark Coherence Function (WBCF) is robust to variable delay often observed in internet telephony such as VoIP. We also show that the refinement of time synchronization after signal decomposition can improve the performance of the WBCF. The regression analysis was performed with VoIP speech data. The correlation coefficients and the standard error of estimates computed using the WBCF showed noticeable improvement over the PSQM that is recommended by ITU-T.

Volume 4, page 2491 Session E22

A Proposed Method for Measuring Language Dependency of Narrow Band Voice Coders

Van Wijngaarden S, Steeneken H
TNO Human Factors, the Netherlands

Narrow band voice coders that use vector quantization techniques may suffer from language dependency: the performance of the coder (in terms of speech intelligibility) may depend on the language spoken. For multinational applications, this is undesirable. A test method is proposed that may be used to determine to which extent a vocoder is language dependent. The proposed method, based on a subjective speech intelligibility test in multiple languages, is shown to be feasible by application on known language dependent 'systems': non-native (human) speakers and listeners. The method is shown to be able to significantly prove differences in language dependency, even when using only three languages and nine speaker/listener combinations.

Volume 4, page 2495 Session E22

An Efficient Transcoding Algorithm For G.723.1 And G.729A Speech Coders

Yoon S W, Jung S K, Park Y C, Youn D H
Yonsei University, Korea

To set a valid communication channel between two endpoints employing different speech coders, decoder and encoder of each endpoint need to be placed in tandem. However, tandem coding is often associated with problems such as poor speech quality, high computational load, and additional transmission delay. In this paper, we propose an efficient transcoding algorithm for a legitimate communication between 5.3 kbps G.723.1 and 8 kbps G.729A coders. The proposed transcoding algorithm is composed of four parts: LSP conversion, open-loop pitch conversion, fast adaptive codebook search, and fast fixed codebook search. In each part of the transcoding algorithm, parameters of the target coder are obtained directly from the parameters of the source coder. The efficient transcoding algorithm is supported via the computational reduction of about 25-35% in the encoding part. Subjective preference tests as well as objective quality evaluation confirmed that the proposed transcoding

algorithm can produce equivalent speech quality to the tandem coding with the shorter processing delay and less computational complexity.

Volume 4, page 2499 Session E22

Joint Source-Channel Coding for Low Bit-Rate Coding of LSP Parameters

Perez-Cordoba J L, Rubio A J, Peinado A M, de la Torre A
Universidad de Granada, Spain

This work presents a quantization technique for LSP parameters which results in a low bit-rate transmission while providing protection against channel errors. As a generalization of the so called Channel Optimized Vector Quantization (COVQ), Channel Optimized Matrix Quantization (COMQ) can remove intraframe and interframe LSP redundancy with the target of protecting the information sent through a channel in the presence of noise. Split COMQ is used in order to reduce storage requirements and complexity. Results show that Split COMQ gives better performance under certain error conditions and a lower bit rate transmission in all channel conditions compared to the reference quantization techniques.

Volume 4, page 2503 Session E22

Session E23 - Oral
Friday - 11.10 - 12.30

Speaker Recognition: Alternative Trends in Verification - II

Chair: Michael Wagner, University of Canberra, Australia

A Segmental Mixture Model for Speaker Recognition

Stapert R P, Mason J S
University of Wales Swansea, UK

Standard Gaussian mixture modelling does not possess time sequence information (TSI) other than that which might be embedded in the acoustic features. Dynamic time warping relates directly to TSI, time-warping two sequences of features into alignment. Here, a hybrid system embedding DTW into a GMM is presented. Improved automatic speaker verification performance is demonstrated. Testing 1000 speakers in a fully text independent, world-model-adapted mode shows an equal error improvement over a standard GMM from 4.1% to 3.8%.

Volume 4, page 2509 Session E23

Tree Based Score Computation for Speaker Verification

Blouet R, Bimbot F
IRISA (CNRS & INRIA), France

This paper proposes an original approach to the task of speaker verification, in which the training process consists in a direct modeling of the score function. It divides the parameter space in disjoint regions where a score can be obtained as a simple function of the vector position in the region. The aim of this approach is, on the one hand to overcome some undesirable properties of the gaussian mixture models (GMMs), and on the other hand, to speed up the decision process. First, we present the formalism of probabilistic speaker verification and we discuss some motivations for exploring alternative approaches. We then describe a method currently under investigation, which is based on a binary recursive partition of the acoustic parameter space into regions to which an elementary scoring function is associated. Finally, we provide illustrations and preliminary results of the method, together with conclusions and perspectives.

Volume 4, page 2513 Session E23

Phonetic Speaker Recognition

Andrews W D, Kohler M A, Campbell J P
Department of Defense, USA

This paper introduces a novel language-independent speaker-recognition system based on differences among speakers in dynamic realization of phonetic features (i.e., pronunciation) rather than spectral differences in voice quality. The system exploits phonetic information from six languages to perform text independent speaker recognition. All experiments were performed on the NIST 2001 Speaker Recognition Evaluation Extended Data Task. Recognition results are provided for each of the six language front ends and for various fusions. The fusion results demonstrate that speaker recognition capability for speech in languages outside the system is successful.

Volume 4, page 2517 Session E23

Speaker Recognition based on Idiolectal Differences between Speakers

Doddington G
National Institute of Standards and Technology, USA

Familiar speaker information is explored using non-acoustic features in NIST's new extended data speaker detection task. Word unigrams and bigrams, used in a traditional target/background likelihood ratio framework, are shown to give surprisingly good performance. Performance continues to improve with additional training and/or test data. Bigram performance is also found to be a function of target/model sex and age difference. These initial experiments strongly suggest that further exploration of familiar speaker characteristics will likely be an extremely interesting and valuable research direction for recognition of speakers in conversational speech.

Volume 4, page 2521 Session E23

Session E24 - Oral
Friday - 11.10 - 12.30

Speech Recognition and Understanding: Rhythm and Timing in ASR

Chair: Guenther Ruske, TU Munich, Germany

An investigation of modelling aspects for rate-dependent speech recognition

Wrede B, Fink G A, Sagerer G
Bielefeld University, Germany

For the modelling of speech rate variation in speech recognition many approaches have been suggested. However, the training of speech-rate dependent models has by far received most of the attention. In order to investigate problematic aspects related with the classification of the speech data which represents one of the major problems of these approaches extensive experiments were carried out on a German corpus of read speech. The results indicate that while the kind of the model-driven speech-rate measure is only of minor importance a data-driven classification of the speech data significantly improves the performance of rate-dependent models. Further results suggest a detailed modelling of speech rate based on more general models. This means that it might be possible to model speech rate adaptation by means of a transformation based on a continuous measure.

Volume 4, page 2527 Session E24

Speaking Rate Dependent Acoustic Modeling for Spontaneous Lecture Speech Recognition

Nanjo H, Kato K, Kawahara T
Kyoto University, Japan

The paper addresses large vocabulary spontaneous speech recognition focusing on acoustic modeling that considers the speaking rate. Using the real lecture speech corpus collected under the priority research project in Japan, we have made baseline acoustic model, and evaluated on the automatic transcription of oral presentations by experienced speakers and obtained word accuracy of 58.2%. Compared with read speech, we have observed significant difference in the speaking rate. To handle fast and poorly articulated phone segments, several extensions of the modeling are explored. Specifically, we introduce state-skipping modeling, speech rate-dependent model, and syllable sub-word modeling. As a result, we reduced the word error rate by absolute 0.8%-2.0%. We also address a language modeling especially on effective use of various large text corpora.

Volume 4, page 2531 Session E24

Analysis of N-Best Output Hypotheses for Fast Speech in Large Vocabulary Continuous Speech Recognition

Fábián T[1], Pfau T[2], Ruske G[1]
[1]Technical University of Munich, Germany, [2]International Computer Science Institute, USA

The performance of speech recognition systems often deteriorate considerably with fast speech. Particularly when the recognizer is run in mismatched conditions, e.g. fast speech, the performance can be improved by properly selecting one of the N-best recognition output hypotheses. For the selection of the best hypothesis, different speech rate measures were taken into account. First, to show the potential of the speech rate as a selection criterion, an "ideal" speech rate value is assumed, which is calculated from the known transcription. Phoneme and vowel rate are compared. Second, a phoneme recognizer is used to estimate the speaking rates of unknown sentences. Tests on the spontaneously spoken German Verbmobil material showed a relative decrease of 6.6% in the word error rate for fast speech, when taking the

estimated vowel rate which is almost as good as using the "ideal" vowel rate (relative improvement of 7.64%). The most accurate match of N-best output hypotheses shows that the word error rate could ideally be decreased by 26.75%.

Volume 4, page 2535 Session E24

Automatic Rhythm Modeling for Language Identification

Farinas J[1], Pellegrino F[2]
[1]IRIT, France, [2]DDL, France

This paper deals with an approach to Automatic Language Identification based on rhythmic modeling. Beside phonetics and phonotactics, rhythm is actually one of the most promising features to be considered for language identification, but significant problems are unresolved for its modeling. In this paper, an algorithm of rhythm extraction is described. Experiments are performed on read speech for 5 European languages. They show that salient features may be automatically extracted and efficiently modeled from the raw signal: a Gaussian mixture modeling of the extracted features results in a 81% percent of correct language identification for the 5 languages, using 20 s duration utterances.

Volume 4, page 2539 Session E24

Session E25 - Poster
Friday - 11.10 - 12.30

Speech Recognition and Understanding: Confidence Measures and OOV

Chair: Jan Nouza, TU Liberec, Czech Republic

Confidence Measure (CM) Estimation for Large Vocabulary Speaker-Independent Continuous Speech Recognition System

Zhang Y[1], Lee R[2], Madievski A[2]
[1]Motorola China Research Center, P. R. China, [2]Motorola Australian Research Center, Australia

In this paper we report a study for confidence measure estimation in a large vocabulary speaker-independent continuous speech recognition system. A hybrid confidence measure estimation algorithm was developed. The final confidence measure consists of a number of confidence parameters which are generated from the different processing levels of the recognition system. A Parameter Reliability Analysis (PRA) algorithm was proposed to combine the confidence parameters to form the final confidence measure. The approach was applied to a large vocabulary speaker-independent continuous speech recognition system and obtained superior performance.

Volume 4, page 2545 Session E25

Experimental Evaluation on Confidence of Agreement among Multiple Japanese LVCSR Models

Kodama Y, Utsuro T, Nishizaki H, Nakagawa S
Toyohashi University of Technology, Japan

For many practical applications of speech recognition systems, it is quite desirable to have an estimate of confidence for each hypothesized word. Unlike previous works on confidence measures, this paper studies features for confidence measures that are extracted from outputs of {it more than one} LVCSR models. More specifically, this paper experimentally evaluates the agreement among the outputs of multiple Japanese LVCSR models, with respect to whether it is effective as an estimate of confidence for each hypothesized word. The results of experimental evaluation show that the agreement between the outputs with two acoustic models which have different units in HMMs, such as phonemes and syllables, can achieve quite reliable confidence.

Volume 4, page 2549 Session E25

Detection of Recognition Errors and Out of the Spelling Dictionary Names in a Spelled Name Recognizer for Spanish

San-Segundo R, Macías-Guarasa J, Ferreiros J, Martín P, Pardo J M
Departamento de Ingeniería Electrónica. UPM., Spain

This paper deals with improved confidence assessment for detecting recognition errors and out of dictionary names in a Spanish Recognizer of continuously spelled names over the telephone. We present a hypothesis-verification approach for spelled name recognition. We evaluate the system for several dictionaries, obtaining more than 90.0% recognition rate for a 10,000 name dictionary. For confidence scoring, we consider several features obtained from the different recognition stages. The paper investigates the ability of each feature set to detect recognition errors and names out of the spelling dictionary. We use a neural network to combine all the features in order to obtain the best confidence annotation. Using the data collected from 1,000 phone calls, it is shown that 57.9% incorrectly recognized names and 68.3% out of the spelling dictionary names are detected at a 5% false rejection rate.

Volume 4, page 2553 Session E25

Use of acoustic prior information for confidence measure in ASR applications

Mengusoglu E, Ris C
Faculté Polytechnique de Mons - TCTS lab, Belgium

In this paper, we propose a new acoustic confidence measure of ASR hypothesis and compare it to approaches proposed in the literature. This approach takes into account prior information on the acoustic model performance specific to each phoneme. The new method is tested on two types of recognition errors: the out-of-vocabulary words and the errors due to additive noise. We then propose an efficient way to interpret the raw confidence measure as a correctness prior probability.

Volume 4, page 2557 Session E25

Improving Performance of a Keyword Spotting System by Using a New Confidence Measure

Ferrer L, Estienne C
School of Engineering, University of Buenos Aires, Argentina

This work describes a HMM-based keyword spotting system. In this system, keywords are modeled as concatenations of phoneme models, consequently, no specific databases are needed to train the system. In addition no filler models are required, therefore small computational requirements are necessary. Two main stages define the whole system. The first stage extracts segments from the utterance corresponding to possible keywords based on the maximization of a confidence measure. Those segments are used as input hypotheses for the second stage in order to get a new confidence measure. This second measure is determined by comparing the vector of emission probabilities for an hypothesis over the keyword model and the vector of emission probabilities for the best sequence of phonemes, in the segment where the hypothesis was detected. The first measure is linearly combined with the second one resulting in a new confidence measure which performs significatively better than that one.

Volume 4, page 2561 Session E25

Word Level Confidence Measures Using N-Best Sub-Hypotheses Likelihood Ratio

Tan B T, Gu Y, Thomas T
Vocalis Ltd., UK

This paper proposes an efficient confidence measure applied at the word level by combining various likelihood ratio tests. The estimates are derived from the local N-best sub-hypotheses. This approach allows the confidence measures to take into account the effect of neighboring words and still provides the estimate localized around the word to be verified. It produces an effective confidence measure that is usable for various tasks. We compared the results with other likelihood ratio based confidence measures including garbage model, N-best homogeneity and online garbage models. The proposed method gave more than 30% relative false accept rate reduction over other methods and the rejection performance was less task-dependent.

Volume 4, page 2565 Session E25

Confidence Based Lattice Segmentation and Minimum Bayes-Risk Decoding

Goel V[1], Kumar S[2], Byrne W[2]
[1]I.B.M. T.J.Watson Research Center, USA, [2]Johns Hopkins University, USA

Minimum Bayes Risk (MBR) speech recognizers have been shown to yield improvements over the conventional maximum a-posteriori probability (MAP) decoders in the context of N-best list rescoring and A* search over recognition lattices. Segmental MBR (SMBR) procedures have been developed to simplify implementation of MBR recognizers, by segmenting the N-best list or lattice, to reduce the size of the search space over which MBR recognition is carried out. In this paper we describe lattice cutting as a method to segment recognition word lattices into regions of low confidence and high confidence. We present two SMBR decoding procedures that can be applied on low confidence segment sets. Results obtained on the Switchboard conversational telephone speech corpus show modest but significant improvements relative to MAP decoders.

Volume 4, page 2569 Session E25

A Data Selection Strategy for Utterance Verification in Continuous Speech Recognition

Jiang H, Soong F, Chin-Hui L
Bell Labs, Lucent Technologies Inc., USA

In this paper, we propose the concept of rival for verifying hypothesis in speech recognition. A likelihood ratio test, based on the rivals model, are investigated for utterance verification in continuous speech recognition. We present an efficient method to train rival model automatically from training data as well as a single pass strategy of utterance verification, namely verification-in-search, for continuous speech recognition. Some preliminary experiments on DARPA Communicator travel task have shown the rival models give better verification performance in terms of identifying mis-recognized words from the output of our baseline recognizer.

Volume 4, page 2573 Session E25

Improved Speech Recognition Using Iterative Decoding Based on Confidence Measures

Ogata J, Ariki Y
Ryukoku University, Japan

In this paper, a decoding method incorporating word-level confidence measures for improved speech recognition is presented. At first, we focus on the estimation of confidence measures from the word graph and evaluate them in word graph rescoring (2nd-pass in 2-pass search system). Next, we propose the lexical tree search (1st-pass in 2-pass search system) incorporating the word-level confidence measures and an iterative decoding based on the confidence measures, resulting in the re-construcion of the word graph. The experimental results showed that this method achieved a slight improvement at word accuracy.

Volume 4, page 2577 Session E25

Detection of OOV Words Using Generalized Word Models and a Semantic Class Language Model

Schaaf T
University of Karlsruhe, Germany

This paper describes an approach to detect out-of-vocabulary words in spontaneous speech using a language model built on semantic categories and a new type of generalized word models consisting of a mixture of specific and general acoustic units. We demonstrate the construction of the generalized word models as replacements for surnames in a German spontaneous travel planning task GSST. We show that the use of our generalized word models improves recognition accuracy in cases where out-of-vocabulary words appear and does not lead to a degradation of the overall recognition accuracy. In our experiments we measured recall and precision rates of OOV-detection which are close to their theoretic optimum. Furthermore, we compared the effect of using cross-word-triphones vs. using context-independent cross-word models. We show that when using generalized word models with cross-word-triphones, the expected number of consequential errors following an OOV word can be reduced significantly by 37%.

Volume 4, page 2581 Session E25

Effects of OOV rates on Keyphrase Rejection Schemes

Bouwman G, Sturm J, Boves L
University of Nijmegen, the Netherlands

Recognising directory listings for national telephone number inquiry is slowly getting within reach for modern ASR technology. Two key factors for a successful system design are (1) optimal extent of lexical modelling and (2) an effective utterance rejection method. In this paper we show how a choice for the first has consequences for the second. We have taken the approach of building a lexicon with multiword expressions for the most frequently requested telephone listings, stepwise extended with filler words and less frequently addressed listings. In doing so, we keep track of the consequences that different Out of Vocabulary (OOV) rates have on two diverging keyphrase rejection schemes. Experimental results on field data clearly show that tasks with high OOV rates benefit most from acoustic confidence measures, while tasks with low OOV rates are better off with N-best list-based rejection schemes.

Volume 4, page 2585 Session E25

Session E26 - Poster
Friday - 11.10 - 12.30

Signal Analysis: Source Localisation and Beam Forming

Chair: Yunxin Zhao, University of Missouri, USA

A New Auditory Based Microphone Array and Objective Evaluation Using E-RASTI

Sanchez-Bote J-L, Gonzalez-Rodriguez J, Simon-Zorita D
ATVS-DIAC-Universidad Politecnica de Madrid, Spain

Two are the goals of the work presented in this paper. The first one is the implementation of a new method of speech enhancement using microphone arrays. This method gets noise reduction of speech signal using the masking properties of the human auditory system. The second goal of the paper is to use RASTI index (RApid Speech Transmission Index) for objective evaluation of speech signal quality through E-RASTI evaluation. What is new is that E-RASTI is applied to speech signals and no to RASTI-like signals. The E-RASTI index is specially suited to test reverberant speech and has been used here to evaluate the reverberation reduction produced by a microphone array based on all-pass and minimum-phase decomposition or multichannel liftering. Noise reduction evaluation has been performed with the E-RASTI index and also with more traditional methods, based on Signal to Noise Ratios (SNR). Results have demonstrated the good performance of the noise suppressor and the E-RASTI objective quality evaluator.

Volume 4, page 2591 Session E26

Equivalence between Frequency Domain Blind Source Separation and Frequency Domain Adaptive Null Beamformers

Araki S[1], Makino S[1], Mukai R[1], Saruwatari H[2]
[1]NTT Communication Science Laboratories, Japan, [2]Nara Institute of Science and Technology, Japan

Frequency domain Blind Source Separation (BSS) is shown to be equivalent to two sets of frequency domain adaptive microphone arrays, that is, Adaptive Null Beamformers (ANB). The unmixing matrix of the BSS and the filter coefficients of the ANB converge to the same solution in the mean square error sense if the two source signals are ideally independent. This understanding clearly explains the poor performance of the BSS in a real room with long reverberation. The fundamental difference exists in the adaptation period when they should adapt. That is, the ANB can adapt in the presence of a jammer but the absence of a target, whereas the BSS can adapt in the presence of a target and jammer, and also in the presence of only a target.

Volume 4, page 2595 Session E26

Separation and Dereverberation Performance of Frequency Domain Blind Source Separation for Speech in a Reverberant Environment

Mukai R, Araki S, Makino S
NTT Communication Science Laboratories, Japan

In this paper, we investigate the performance of an unmixing system obtained by frequency domain Blind Source Separation (BSS) based on Independent Component Analysis (ICA). Since ICA is based on statistics, i.e., it only attempts to make outputs independent, it is not easy to predict what is going on in a BSS system. We therefore investigate the detailed components in the processed signals of a whole BSS system by measuring four impulse responses of the system. In particular, we focus on the direct sound and reverberation in the target and jammer signals. As a result, we reveal that the direct sound and reverberation of the

jammer can be reduced compared to a null beamformer (NBF), while the reverberation of the target can not be reduced.

Volume 4, page 2599 Session E26

Blind Source Separation for Speech Based on Fast-Convergence Algorithm with ICA and Beamforming

Saruwatari H, Kawamura T, Shikano K
Nara Institute of Science and Technology, Japan

We propose a new algorithm for blind source separation (BSS), in which independent component analysis (ICA) and beamforming are combined to resolve the low-convergence problem through optimization in ICA. The proposed method consists of the following three parts: (1) frequency-domain ICA with direction-of-arrival (DOA) estimation, (2) null beamforming based on the estimated DOA, and (3) integration of (1) and (2) based on the algorithm diversity in both iteration and frequency domain. The inverse of the mixing matrix obtained by ICA is temporally substituted by the matrix based on null beamforming through iterative optimization, and the temporal alternation between ICA and beamforming can realize fast- and high-convergence optimization. The results of the signal separation experiments reveal that the signal separation performance of the proposed algorithm is superior to that of the conventional ICA-based BSS method, even under reverberant conditions.

Volume 4, page 2603 Session E26

Noise Reduction Using Paired-microphones for both Far-field and Near-field Sound Sources

Mizumachi M, Nakamura S
ATR Spoken Language Translation Research Laboratories, Japan

The near-field problem is a source of anxiety with beamforming techniques. We propose a strategy to solve the near-field problem by using a subtractive beamforming technique. The authors earlier proposed a method for noise reduction using paired-microphones. In this paper, the method is improved for near-field sound sources. The proposed method can maintain a high performance regardless of the distance between the sound source and an array, but the performance of a Delay-and-Sum beamformer declines even if the amplitude of the target signal is normalized. The concept of ``paired-microphones" in the proposed method is the key for solving the near-field problem.

Volume 4, page 2607 Session E26

Statistical Sound Source Identification in a Real Acoustic Environment for Robust Speech Recognition Using a Microphone Array

Nishiura T[1], Nakamura S[1], Shikano K[2]
[1]ATR Spoken Language Translation Research Laboratories, Japan,
[2]Nara Institute of Science and Technology, Japan

It is very important for a hands-free speech interface to capture distant talking speech with high quality. A microphone array is an ideal candidate for this purpose. However, this approach requires localizing the target talker. To cope with this problem, we propose a new talker localization method consisting of two algorithms. One algorithm is for multiple sound source localization based on CSP (Cross-power Spectrum Phase) analysis. The other algorithm is for sound source identification among localized multiple sound sources towards talker localization. In this paper, we particularly focus on the latter statistical sound source identification among localized multiple sound sources with statistical speech and environmental sound models based on GMMs (Gaussian Mixture Models) and a microphone array towards talker localization.

Volume 4, page 2611 Session E26

Speech Enhancement and Source Separation based on Binaural Negative Beamforming

Álvarez-Marquina A, Gómez-Vilda P, Martínez-Olalla R, Nieto-Lluís V, Rodellar-Biarge V
Universidad Politécnica de Madrid, Spain

Negative Beamformers are well known for their high angular selectivity, which makes them potentially suitable for speech enhancement applications in noisy backgrounds and for directional source separation. On the other hand, Spectral Subtraction is a well-known method for removing noise from a noise-corrupted speech signal. The scheme that is proposed in this paper combines both techniques in order to obtain large gains in the SNR at a reasonable low computational cost. This method may be used to eliminate or enhance a specific signal using a binaural array. The fundamentals of the technique are reviewed, and a structure to control and improve its angular selectivity is presented. Results obtained in a real situation are also commented. Applications of this technique may be found in Security Systems, Domotic Control and also, to improve Speech Recognition.

Volume 4, page 2615 Session E26

Multiple source separation in the frequency domain using Negative Beamforming

Gómez-Vilda P, Álvarez-Marquina A, Nieto-Lluís V, Rodellar-Biarge V, Martínez-Olalla R
Universidad Politécnica de Madrid, Spain

The localization of acoustic sources in a room is essential in many applications, as security monitoring, video conferencing, automatic scene analysis, reverberation canceling, or robust Speech Recognition under multiple-party effect. Through the present paper the design and operation of a negative beamformer for multiple source speech separation will be presented. The problems found for its proper operation when multiple sources are present on the same band will be pointed out and the solutions found will be commented and discussed showing the results of real experiments carried out on a recording scenario.

Volume 4, page 2619 Session E26

Planar Superdirective Microphone Arrays for Speech Acquisition in the Car

Martin R[1], Petrovsky A[2], Lotter T[1]
[1]Aachen University of Technology, Germany, [2]Belarussian State University of Informatics and Radioelectronics, Republic of Belarus

In this paper we investigate a small broadside planar (2D) superdirective microphone array for speech acquisition in the car and compare its performance to linear arrays. The objective of this investigation is to replace an expensive directional microphone by a small array of inexpensive omnidirectional sensors. Since the array was designed to be used in the car environment it has to satisfy restrictions with respect to size and to the number of microphones. For all array configurations we present theoretical gains, actual measured gains using low-cost microphones, and beam patterns. For a fixed number of microphones and fixed array dimensions we show that the planar design leads to slightly superior array gains.

Volume 4, page 2623 Session E26

Is Speech Data Clustered? - Statistical Analysis of Cepstral Features

Kinnunen T, Fränti P
University of Joensuu, Finland

Abstract: Speech analysis applications are typically based on short-term spectral analysis of the speech signal. Feature extraction process outputs one feature vector per frame. The features are further processed by application-dependent techniques, such as hidden Markov models or

vector quantization. Independent from the application, it is often desirable that the feature vectors form separable clusters in the feature space. In this work, we study whether data is really clustered in the feature space and, if so, what is the number of the clusters in typical speech data. We consider different forms of the widely used cepstral features. Keywords: Speech analysis, pattern recognition, short-term features, cluster analysis.

Volume 4, page 2627 Session E26

Maximum Likelihood Adaptation for Distant Speech Recognition of Stationary and Moving Speakers in Reverberant Environments

Nokas G, Dermatas E, Kokkinakis G
University of Patras, Greece

In this paper, a feature transformation method is presented for distant speech recognition in reverberant and noisy environments. In the Maximum Likelihood framework the optimum bias parameters are obtained on-line, using a small number of successive speech frames. The stochastic matching is achieved by assuming a mixture of Gaussians pdf for the clean speech features. The proposed method was evaluated on the Mel-scaled Frequency Cepstral Coefficient (MFCC) features as well as on MFCC after cepstral mean subtraction and after RASTA filtering. The experiments, carried out in several adverse conditions including room acoustics and additive factory noise for stationary and moving speakers, have shown significant improvement of the recognition accuracy for isolated word speech recognition. In the experiments, the proposed method improves the recognition score of a standing speaker by more than 50%, when SNR is higher than 10db. In the case of the moving speaker the improvement is 8.6% using MFCC while the score reach 91.05% using RASTA fetures.

Volume 4, page 2631 Session E26

Model-based Blind Estimation of Reverberation Time: Application to Robust ASR in Reverberant Environments

Couvreur L[1], Ris C[1], Couvreur C[2]
[1]Faculte Polytechnique de Mons, Belgium, [2]Lernout and Hauspie Speech Products, Belgium

This paper presents a method for blind estimation of reverberation times in reverberant enclosures. The proposed algorithm is based on a statistical model of short-term log-energy sequences for echo-free speech. Given a speech utterance recorded in a reverberant room, it computes a Maximum Likelihood estimate of the room full-band reverberation time. The estimation method is shown to require little data and to perform satisfactorily. The method has been successfully applied to robust automatic speech recognition in reverberant environments by model selection. For this application, the reverberation time is first estimated from the reverberated speech utterance to be recognized. The estimation is then used to select the best acoustic model out of a library of models trained in various artificial reverberant conditions. Speech recognition experiments in simulated and real reverberant environments show the efficiency of our approach which outperforms standard channel normalization techniques.

Volume 4, page 2635 Session E26

Using the Modulation Complex Wavelet Transform for Feature Extraction in Automatic Speech Recognition

Momomura Y[1], Okada K[1], Arai T[1], Kanedera N[2], Murahara Y[1]
[1]Sophia Univ., Japan, [2]Ishikawa National College of Technology, Japan

In this paper we examine robust feature extraction methods for automatic speech recognition (ASR) in noise-distorted environments. Previous research showed that combining the coefficients of multi-resolutional modulation frequency band. We show that this multi-resolutional approach can be achieved using a wavelet transform instead of the Fourier transform. Taking the FFT phase into consideration, we applied the Gabor function, which is a complex function, as mother wavelet. This approach yielded a 1.7% increase in recognition accuracy compared to the FFT-based multi-resolutional approach.

Volume 4, page 2639 Session E26

Separating Three Simultaneous Speeches with Two Microphones by Integrating Auditory and Visual Processing

Okuno H G, Nakadai K, Lourens T, Kitano H
Kitano Symbiotic Systems Project, JST, Japan

This paper addresses the problem of automatic recognition of three simultaneous speeches with two microphones, that is, that of sound source separation where the number of sound sources is greater than that of microphones. The approach used is the {it direction-pass filter}, which is implemented by hypothetical reasoning on the interaural phase difference (IPD) and interaural intensity difference (IID). Auditory processing calculates IPD and IID for each subband, and generates hypotheses for precalculated IPD and IID for every direction including one obtained by visual processing. Then the system calculates the belief factor of hypothesis by Dempster-Shafer theory and determines the direction of each subband. Subbands of the specific direction are collected and then converted to a wave form by inverse FFT. With 200 benchmarks of three simultaneous utterances of Japanese words, the average 1-best and 10-best recognition rates of extracted speeches are 60% and 81%, respectively.

Volume 4, page 2643 Session E26

Session E31 - Oral
Friday - 13.30 - 14.50

Signal Analysis: Speech Features and Modelling

Chair: Matti Karjalainen, HUT, Finland

A Time-Varying Complex AR Speech Analysis Based on GLS and ELS Method

Funaki K
University of The Ryukyus, Japan

We have already developed three kinds of time-varying complex AR (TV-CAR) parameter estimation algorithms for analytic speech signal, which are based on minimizing mean square error (MMSE), Huber's robust M-estimation and Instrumental Variable (IV) method. This paper presents novel robust TV-CAR model parameter estimation algorithms on the basis of a Generalized Least Square (GLS) and Extended Least Square (ELS) method, in which the equation error is modeled by complex AR model with white Gaussian input to whiten the equation error. The experiments with natural speech corrupted by white Gaussian demonstrate that the proposed methods achieve robust spectral estimation against additive white Gaussian.

Volume 4, page 2649 Session E31

Vocal Tract Normalization Equals Linear Transformation in Cepstral Space

Pitz M, Molau S, Schlüter R, Ney H
RWTH Aachen - University of Technology, Germany

We show that vocal tract normalization (VTN) frequency warping results in a linear transformation in the cepstral domain. For the special case of a piece-wise linear warping function, the transformation matrix is analytically calculated. This approach enables us to compute the Jacobian determinant of the transformation matrix, which allows the normalization of the probability distributions used in speaker-normalization for automatic speech recognition.

Volume 4, page 2653 Session E31

An Algorithm for finding Line Spectrum Frequencies of Added Speech Signals and its Application to Robust Speech Recognition

Yu A-T, Wang H-C
National Tsing Hua University, Taiwan

Line Spectrum Frequencies (LSFs) are efficient and popular for representing the spectral envelope in low bit-rate (LBR) speech coding. It is also attractive to use LSFs in the task of speech or speaker recognition. Although, the LBR speech coding does not deteriorate the recognition performance substantially, additive noise does degrade the performance. This paper presents an algorithm for finding LSFs of the addition of two speech signals. This algorithm can be used to adapt the model of noisy speech in LSFs domain, thereby improve the robustness of speech recognition in noisy environments. The experiments on Mandarin digits recognition have proved the effectiveness of the proposed algorithm.

Volume 4, page 2657 Session E31

Improved entropic gain for speech signals analysis/synthesis based on an adaptive time-frequency segmentation scheme

Gonon G[1], Montrésor S[2], Baudry M[1]
[1]*Laboratoire d'Informatique de l'Université du Maine, France,*
[2]*Laboratoire d'Acoustique de l'Université du Maine, France*

In the search for adaptive representation of speech signals, the Wavelet Packet Decomposition (WPD) has been proved to be a efficient tool because of its frequency adaptation skills through the best basis search algorithm. The entropic minimization of this algorithm is bounded by two artifacts : the dyadic structure of the decomposition and the lack of temporal segmentation. We propose here a low cost extended tree in the WPD which improves the best basis search by reducing the entropy of the base and which is still compatible with the classical WPD. The decomposition also allows perfect reconstruction. The entropic test is updated to take into account the new basis. The preliminary use of a temporal segmentation, based on the Local Entropic Criterion highly improves the entropic gain of the global analysis. The results are shown on experimental speech signals comparing the gain of our scheme versus a usual WPD.

Volume 4, page 2661 Session E31

Session E32 - Oral
Friday - 13.30 - 14.50

Speech Recognition and Understanding: Kids, Toys and Emotions

Chair: Roger Moore, 20/20 Speech, United Kingdom

Automatic Word Acquisition from Continuous Speech

Lucke H, Omote M
Sony Corporation, Japan

A method for learning lexical representations of unknown words in an unsupervised manner is described. The unknown words are automatically extracted from continuous speech and a clustering algorithm is used to derive word clusters and lexical representations based on the set of phonetic units used in the system. In experiments, we verify the robustness of the approach. An interesting feature is that extraction errors usually do no harm, as wrongly extracted words tend to inhabit clusters by themselves and thus do not adversely effect the modeling of correctly extracted words.

Volume 4, page 2667 Session E32

Why is automatic recognition of children's speech difficult?

Li Q, Russell M
University of Birmingham, UK

This paper is concerned with automatic recognition of children's speech. The paper begins with a comparison of vowel formant frequencies for adult and children's speech, and notes that in many cases, the average value of F3 for children is greater than 4kHz. Next it is shown that recognition accuracy for children's speech degrades rapidly as bandwidth is reduced to less that 6kHz. Finally, it is demonstrated that the choice of front-end signal processing parameters such as analysis window length, and mel-scale filter widths, have little effect on recognition accuracy for children's speech. It is concluded that bandwidth reduction is a major contributor to the difficulty of recognition of children's speech.

Volume 4, page 2671 Session E32

Politeness and frustration language in child-machine interactions

Arunachalam S, Gould D, Andersen E, Byrd D, Narayanan S
USC, USA

Children represent a potentially crucial user segment for conversational interfaces. Computer systems interacting with children need to be tailored for these users so that they will understand child intent and so that the child will have a positive and successful experience with the system. This study focuses on discourse analysis of spoken-language child- machine interactions. In particular, politeness and frustration markers were analyzed using a database of child-machine conversations obtained from 160 children using a computer game in a wizard-of-Oz set up. Results indicate that younger children less likely to use overt politeness markers and more polite information requests compared to the older ones, with no apparent gender differences. Younger children, on the other hand, expressed frustration verbally more than the older ones; furthermore, frustration language was more predominant in male children.

Volume 4, page 2675 Session E32

Speech Emotion Recognition Using Hidden Markov Models

Albino N, Asunción M, Antonio B, José B. M
Research Center TALP UPC, Spain

This paper introduces a first approach to emotion recognition using RAMSES, the UPC's speech recognition system. The approach is based on standard speech recognition technology using hidden semi-continuous Markov models. Both the selection of low level features and the design of the recognition system are addressed. Results are given on speaker dependent emotion recognition using the Spanish corpus of INTERFACE emotional speech synthesis database. The accuracy recognising seven different emotions---the six ones defined in MPEG-4 plus neutral style---exceeds 80% using the best combination of low level features and HMM structure. This result is very similar to that obtained with the same database in subjective evaluation by human judges.

Volume 4, page 2679 Session E32

Session E33 - Oral
Friday - 13.30 - 14.50

Applications: Media Applications

Chair: Helmut Mangold, DaimlerChrysler, Germany

Speech Enhanced Remote Control for Media Terminal

Ibrahim A, Lundberg J, Johansson J
Linköpings Univ., Sweden

A media terminal box combines digital television and services on the World Wide Web. This device will be available in many homes and the interaction with it occurs via a remote control and a visual presentation. The problem is the navigation difficulties among the huge number of television channels. The aim of this study is to investigate whether spoken commands could solve the navigation problem. In this study two input techniques were tested: remote control and speech input. The results showed that speech input was more effective as steps to complete tasks were less and shortcuts were used more often in the speech condition. However, the subjective data showed that the subjects were more satisfied with the remote control input. In conclusion, we recommend multimodal interaction where of speech input to complement the remote control unit.

Volume 4, page 2685 Session E33

The development of a Portuguese version of a media watch system

Amaral R[1], Langlois T[2], Meinedo H[2], Neto J[2], Souto N[3], Trancoso I[2]
[1]INESC ID/IPS, Portugal, [2]INESC ID/IST, Portugal, [3]INESC ID, Portugal

This paper summarizes the work that has been done concerning the Portuguese language in the scope of the ALERT project during its first year. The media watch system that is the goal of this project comprises many different modules, some of them common among the three languages of the project. This paper concentrates on the definition and collection of the necessary linguistic resources for Portuguese, and the development of the speech recognition, topic and jingle detection modules. The first version of the ALERT demo for European Portuguese is also described.

Volume 4, page 2689 Session E33

Classification of Video Genre using Audio

Roach M, Mason J
University of Wales Swansea, UK

In this paper we propose an approach to high-level classification of video into genre: sport, cartoon, news, commercial and music. An important issue for automatic high-level classification systems is the amount of time needed to classify a video. Here we investigate classification performance as a function of the test sequence length. In addition we present performance against different orders and combinations of static and dynamic mel-frequency cepstral coefficients (MFCC). We find that static and delta MFCCs perform well for this classification task. A test sequence length of approximately 25 seconds for the 5 class problem gives approximately 80% correct identification.

Volume 4, page 2693 Session E33

Prosody in Finger Braille and Teletext Receiver for Finger Braille

Horiuchi Y, Ichikawa A
Chiba University, Japan

In this paper, we introduce durational rules in text-to-Finger-Braille. Finger Braille is one of the communication methods for the deaf blind and it seems to be the medium most suited to real-time communication and for expressing the feelings of the speaker because of its prosody existing similarly to spoken languages. First, we analyzed duration between two Braille codes in Finger Braille and found that it can be changed according to the structure and meaning of the sentences. Second, we construct durational rules in Finger Braille based on these results. Third, the effectiveness of the rule was examined in listening experiments with a deaf blind person. As a result, it is suggested that durational prosody help listeners to have clear understanding. Finally, we made a prototype of Finger Braille receiver for teletext broadcasting system as a practical application applying this rule.

Volume 4, page 2697 Session E33

Session E34 - Oral
Friday - 13.30 - 14.50

Speech Recognition and Understanding: Distributed Speech Recognition

Chair: Gerhard Rigoll, Univ. Duisburg, Germany

Joint Channel Decoding - Viterbi Recognition For Wireless Applications

Bernard A, Alwan A
UCLA, USA

We introduce the concept of joint channel decoding and Viterbi recognition, by which the Viterbi recognizer is modified to take into account the confidence in the decoded feature after channel transmission. We present a metric for evaluating such confidence based on soft decision decoding. As a case study, we quantize MFCCs using predictive VQ. The overall source-channel coding scheme operating at a combined rate of 1 kbps is shown to provide good recognition accuracy over a wide range of Rayleigh fading channels.

Volume 4, page 2703 Session E34

MMSE-Based Channel Error Mitigation for Distributed Speech Recognition

Peinado A M, Sanchez V, Segura J C, Perez-Cordoba J L
Universidad de Granada, Spain

Recently, the first version of an ETSI standard for Distributed Speech Recognition has been proposed. The main benefit of this approach is the possibility of maintaining a high recognition performance when accessing remote information systems. The use of a digital channel for transmission of the encoded speech parameters implies the introduction of several channel distortions. Our paper deals with the mitigation of such distortions. We study the application of MMSE estimation to this problem and propose a new MMSE procedure that obtains the probabilities needed for MMSE from a forward-backward algorithm. We show that MMSE estimation obtains better performance than the mitigation algorithm described in the ETSI standard under different channel conditions.

Volume 4, page 2707 Session E34

Distributed Speech Recognition using Traditional and Hybrid Modeling Techniques

Stadermann J[1], Meermeier R[2], Rigoll G[1]
[1]University of Duisburg, Germany, [2]SpeechWorks International, USA

We compare the performance of different acoustic modeling techniques on the task of distributed speech recognition (DSR). The DSR technology is interesting for speech recognition tasks in mobile environments, where features are sent from a thin client to a server where the actual recognition is performed. The evaluation is done on the TI digits database which consists of single digits and digit-chains spoken by American-English talkers. We investigate clean speech and speech added with white noise. Our results show that new hybrid or discrete modeling techniques can outperform standard continuous systems on this task.

Volume 4, page 2711 Session E34

Graceful Degradation of Speech Recognition Performance Over Lossy Packet Networks

Riskin E, Boulis C, Otterson S, Ostendorf M
University of Washington, USA

This paper explores packet loss recovery in client-server Automatic Speech Recognition (ASR) systems. A forward error correction (FEC) system is designed and tested over several channel loss models, at variable amounts of data acquisition delay. In experiments with simulated packet loss, the FEC system provides robust ASR performance which degrades gracefully as packet loss rates increase. Comparing this scheme to several alternatives under low and medium loss channel conditions, we found one approach (multiple transmission plus interpolation) that yielded similar performance, but the FEC system should scale better to lower bit rate conditions.

Volume 4, page 2715 Session E34

Session E35 - Poster
Friday - 13.30 - 14.50

Speech Recognition and Understanding: Prosody and Cross-Language in ASR

Chair: Phil Green, University of Sheffield, United Kingdom

Experiments on Cross-language Acoustic Modeling

Schultz T, Waibel A
Carnegie Mellon University, USA

With the distribution of speech products all over the world, the portability to new target languages becomes a practical concern. As a consequence our research focuses on rapid transfer of LVCSR systems to other languages. In former studies we evaluated the performance if limited adaptation data is available. Particularly for very time constrained tasks and minority languages, it is even reasonable that no training data is available at all. In this paper we examine what performance can be expected in this scenario. All experiments are run in the framework of the GlobalPhone project which investigates LVCSR systems in 15 languages.

Volume 4, page 2721 Session E35

Crosslingual Speech Recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering

Zgank A[1], Imperl B[1], Johansen F T[2], Kacic Z[1], Horvat B[1]
[1]University of Maribor, Slovenia, [2]Telenor Research and Development, Norway

The paper describes our ongoing work on crosslingual speech recognition based on multilingual triphone hidden Markov models. Multilingual acoustic models were built using two different clustering procedures: agglomerative triphone clustering and tree-based triphone clustering. The agglomerative clustering procedure is based on measuring the similarity of triphones on a phoneme level where the monophone similarity is estimated by the Houtgast algorithm. The tree-based clustering procedure is based on common broad classes. The Slovenian, German and Spanish 1000 FDB SpeechDat(II) databases were used for training. The crosslingual speech recognition was performed on the Norwegian 1000 FDB SpeechDat(II) database. No adaptation or training with the Norwegian database was used. The mapping of Norwegian phonemes was done with the IPA scheme. Five different Norwegian recognition vocabularies were generated. The best crosslingual system achieved a recognition rate of 45.03%, while the reference Norwegian system achieved 78.32%.

Volume 4, page 2725 Session E35

Comparing parameter tying methods for multilingual acoustic modelling

Harju M[1], Salmela P[1], Leppänen J[1], Viikki O[2], Saarinen J[1]
[1]Tampere University of Technology, Finland, [2]Nokia Research Center, Finland

In this paper, we compare the state-level and model-level tying of continuous density hidden Markov models for the multilingual acoustic modelling. Using the model-level tying technique, the number of the language dependent (LD) phoneme models of five European languages were reduced to the desired number. This tying was based on dissimilarity measure between the LD phoneme models in a bottom-up agglomerative clustering technique. This system provided 87.3% word recognition accuracy on the test set, while a comparable multilingual recognition based on the SAMPA phone inventory obtained 84.6% accuracy on the same set. The above model-level tying technique was also used for obtaining an alternative phone inventory to SAMPA such that both inventories have an equal number of phones for these five languages. The multilingual recognition systems trained for the SAMPA and alternative phone invetonries obtained 80.9% and 83.7% word accuracies on the same test set, when state-level tying was used for reducing the number of the parameters from 199k to 76k in both systems. The original LD recognition systems obtained 89.0% recognition rate with the same test set, which contained approximately 200 isolated words from SpeechDat(II) databases for each of the five languages. In this paper, the test set results are also given for the recognition systems after performing MAP language adaptation for the multilingual phone models.

Volume 4, page 2729 Session E35

Accent-Independent Universal HMM-Based Speech Recognizer for American, Australian and British English

Chengalvarayan R
Lucent Technologies Inc., USA

This paper addresses the problem of speech recognition under accent variations in English language. It has been demonstrated in previous research efforts that the multi-transitional model architecture is one of the solutions for robust speech recognition. In this study, we describe an universal hybrid system that is trained with data from American, Australian, and British accented speech. Experimental results on connected-digit recognition task show an average string error rate reduction of about 62% and 8% when compared to our best monolingual and multi-transitional systems respectively. The result indicates that the universal model is about three times faster and half time smaller than the multi-transitional or multilingual models and this makes it an ideal choice for practical accent-independent speech recognition applications.

Volume 4, page 2733 Session E35

The Effect of Time Stress on Automatic Speech recognition Accuracy When Using Second Language

Chen F[1], Sääv J[2]
[1]Swedish Center for Human Factors, Sweden, [2]Linköping Univ., Sweden

The purpose of the present study is to compare the ASR performance when Swedish people speaking Swedish and English under time-stress and due-task performance. Fifteen university students (20 to 40 years of age, native Swedish language speaking) participated in the experiment. Three factors were studied: time-stress, which was manipulated by PWSP program. Two models of presenting the commands, one is by displaying the text on the screen and another is by headphone voice. Swedish and English languages were tested on Philips FreeSpeech 2000 speech recognition system. There is no individual voice file training and pre-designed grammar file for the speech recognition system. The results show that there are no interactions between any of the factors. The individual differences are large. There is a significant decrease of recognition accuracy ($p<0.05$) for both languages during stress. The recognition accuracy on Swedish language is significant higher ($p<0.01$) than English Language due to the Swedish accents.

Volume 4, page 2737 Session E35

The Effect of Pitch and Lexical Tone on Different Mandarin Speech Recognition Tasks

Wong Y W[1], Chang E[2]
[1]The Chinese University of Hong Kong, Hong Kong, P. R. China, [2]Microsoft Research China, P. R. China

Tone is an important component in Mandarin speech recognition. It is necessary to recognize the five lexical tones to disambiguate between confusing words. Tone is acoustically characterized by the pitch contour. The use of pitch has been shown to be helpful in Mandarin syllable

recognition. In this paper, a comprehensive set of investigations on the effect of pitch on diverse Mandarin speech recognition tasks, namely large vocabulary continuous speech recognition (LVCSR) and isolated word recognition, is reported. In this paper, various techniques to utilize pitch in acoustic modeling are examined. In particular, modeling of tone context dependence and normalization of pitch value are investigated. The experimental result shows that with the incorporation of pitch, an error reduction of 26% can be achieved in tonal syllable recognition. The same level of error reduction is attained in isolated word recognition. On the other hand, the gain from using pitch in an LVCSR task is less. The result suggests that without a language model, the use of pitch is more beneficial in Mandarin speech recognition, thus speech recognizers may be designed to dynamically make use of the pitch feature to obtain the best tradeoff between accuracy and computation.

Volume 4, page 2741 Session E35

Acoustic Modeling of Foreign Words in a German Speech Recognition System

Stemmer G, Nöth E, Niemann H
University of Erlangen-Nürnberg, Germany

The paper deals with the development of acoustic models of foreign words for a German speech recognizer. The recognition quality of foreign words is crucial for the overall performance of a system in application fields like spoken dialogue systems, when foreign words occur as proper names. One of the main problems in the modeling of foreign words is the limitation of training data, which must contain samples of the non-native pronunciation of the foreign sounds. In order to obtain robust acoustic models, which are still precise enough, we compare several methods to map or to merge the models of phonemes, which are pronounced in a similar way by German speakers. We utilize an entropy-based distance measure between sets of phoneme models. The best approach yields a reduction of 16.5 % word error rate, when compared to a baseline system.

Volume 4, page 2745 Session E35

Semi-Automatic Grammar Induction for Bi-directional English-Chinese Machine Translation

Siu K-C, Meng H
The Chinese University of Hong Kong, Hong Kong, P. R. China

We have previously designed a methodology for semi-automatic grammar induction from un-annotated corpora belonging to a restricted domain. The induced grammar contains both semantic and syntactic structures, and experiments with the Air Travel Information Service (ATIS-3) corpus demonstrated the viability of our approach [1] for natural language understanding. This work explores the portability of our grammar induction approach to Chinese, based on a corpus of translated ATIS-3 queries. To assess grammar quality, we developed a framework bi-directional English-Chinese example-based machine translation using the induced grammars. Our translation framework can handle word order differences between the language pair during translation. Translations based on the ATIS-3 test sets showed a high percentage (76% to 91%) of user-accepted translations.

Volume 4, page 2749 Session E35

F0 Feature Extraction by Polynomial Regression Function for Monosyllabic Thai Tone Recognition

Charnvivit P, Jitapunkul S, Ahkuputra V, Maneenoi E, Thathong U, Thampanitchawong B
Chulalongkorn University, Thailand

This paper presents a monosyllabic Thai tone recognition system. The system is composed of three processes, fundamental frequency (F0) extraction from input speech signal, analysis of F0 contour for feature extraction, and classification of each tone using the extracted features. In the F0 feature extraction, the polynomial regression functions are employed to fit the segmented F0 curve where its coefficients are used as a feature vector. In tone recognition, we used the maximum a posteriori probability classifier (MAP) to classify a tone. The vocabulary set is composed of the short vowel words, the long vowel words and have the effect of initial and final consonant on the shape of F0 contour. The experimental results show that by using the system as a speaker-dependent system, the maximum recognition rate is 96.20% using three-dimension feature vector. The speaker-independent recognition rates are 79.99% for male and 82.80% for female using four-dimension feature vector.

Volume 4, page 2753 Session E35

The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition

Kim J-H, Woodland P C
Cambridge University, UK

In this paper, we discuss a combined system for punctuation generation and speech recognition. This system incorporates prosodic information with acoustic and language model information. Experiments are conducted for both the reference transcriptions and speech recogniser outputs. For the reference transcription case, prosodic information is shown to be more useful than language model information. When these information sources are combined, we can obtain an F-measure of up to 0.7830 for punctuation recognition. A few straightforward modifications of a conventional speech recogniser allow the system to produce punctuation and speech recognition hypotheses simultaneously. The multiple hypotheses are produced by the automatic speech recogniser and are re-scored by prosodic information. When prosodic information is incorporated, the F-measure can be improved by 19% relative. At the same time, small reductions in word error rate are obtained.

Volume 4, page 2757 Session E35

Lexical Stress Modeling for Improved Speech Recognition of Spontaneous Telephone Speech in the Jupiter Domain

Wang C, Seneff S
MIT Laboratory for Computer Science, USA

This paper examines an approach of using lexical stress models to improve the speech recognition performance on spontaneous telephone speech. We analyzed the correlation of various pitch, energy, and duration measurements with lexical stress on a large corpus of spontaneous English utterances, and identified the most informative features of stress using classification experiments. We incorporated the stress models into the recognizer first-pass Viterbi search and obtained modest but statistically significant improvements over a state-of-the-art real-time performance on the Jupiter weather information domain.

Volume 4, page 2761 Session E35

Modeling Auxiliary Information in Bayesian Network Based ASR

Stephenson T A, Magimai Doss M, Bourlard H
Dalle Molle Institute for Perceptual Artificial Intelligence (IDIAP), Switzerland

Automatic speech recognition bases its models on the acoustic features derived from the speech signal. Some have investigated replacing or supplementing these features with information that can not be precisely measured (articulator positions, pitch, gender, etc.) automatically. Consequently, automatic estimations of the desired information would be generated. This data can degrade performance due to its imprecisions. In this paper, we describe a system that treats pitch as an auxiliary information within the framework of Bayesian networks, resulting in improved performance.

Volume 4, page 2765 Session E35

A New Dynamic HMM Model for Speech Recognition

Chen F[1], Chang E[2]
[1]*Shanghai JiaoTong University / Bell-lab Joint Lab, P. R. China,*
[2]*Microsoft Research China, P. R. China*

In this paper, we describe a new method to do speech recognition based on Dynamic HMM architecture. Pitch values are treated as hidden layer and used to modify the parameter of observation probability functions. The results show that the new model achieves approximately 10 percent relative error reductions both in base-syllable recognition task and tonal syllable recognition task. The new method can be used compatibly with conventional HMM based EM training algorithm and Viterbi decoding algorithm.

Volume 4, page 2769 Session E35

Multi-Keyword Spotting of Telephone Speech Using Orthogonal Transform-Based SBR and RNN Prosodic Model

Wang W-J, Chen S-H
National Chiao Tung University, Taiwan, ROC

In this paper, orthogonal transform-based signal bias removal (OTSBR) approach and RNN prosodic model are proposed for multi-keyword spotting of telephone speech. OTSBR is employed in the pre-processing stage of acoustic decoding and aimed at channel bias estimation to eliminate the acoustic mismatch between training and testing environments. The RNN prosodic model is adopted in the post-processing stage of the acoustic decoding to detect word boundaries for reordering the keyword candidates from the keyword spotter. Simulations on the real speech database collected from the Phone Directory Assistant Service developed in Chunghwa Telecommunication Laboratories (CTL-PDAS) were performed to evaluate the proposed methods. Experimental results showed that 71.0% of keyword detection rate and 81.8% of top 5 keywords inclusion rate can be attained by incorporating OTSBR and RNN prosodic model into the system.

Volume 4, page 2773 Session E35

Recognition of Slovenian Speech: Within and Cross - Language Experiments on Monophones using the SpeechDat(II)

Iskra A[1], Petek B[1], Brøndsted T[2]
[1]*University of Ljubljana, Slovenia,* [2]*Aalborg University, Denmark*

Though the Slovenian SpeechDat(II) database is the largest spoken language resources for Slovenian ever recorded, it belongs to the smaller speech data collections made available by the European LE2-4001 project (http://www.speechdat.org/). The aim of this paper is to analyze this new Slovenian resource and explore the possibilities of supplementing it with data recorded for other languages. The donor languages being considered are English, German, and Danish. For each of these languages four time as much speech data has been recorded (4000 speakers compared to the Slovenian 1000 speaker database). Our purely data-driven cross language tests show that serious problems are involved when porting data across languages. The problems are partly due to differences in the recording conditions (telephone line noise). Other problems can be explained by the different phonological structures of the analyzed languages.

Volume 4, page 2777 Session E35

Boiling down Prosody for the Classification of Boundaries and Accents in German and English

Batliner A, Buckow J, Huber R, Warnke V, Nöth E, Niemann H
University of Erlangen-Nuremberg, Germany

In the focus of this paper is a comparison of the most relevant prosodic features/feature classes for the classification of boundaries and accents in German and in English. Principal components were computed based on a large prosodic feature vector; these principal components were used as predictor variables in a Linear Discriminant analysis as well as in a Classification and Regression Tree. The number of the most relevant principal components was between three and five; for both languages and for boundary and accent classification alike, most important were principal components modelling duration, in combination with energy, followed by pauses and F0.

Volume 4, page 2781 Session E35

Session E36a - Poster
Friday - 13.30 - 14.50

Education: Education and Training

Chair: Valerie Hazan, University College London, United Kingdom

JavaSpeakerRecognition - Interactive Workbench for Visualizing Speaker Recognition Concepts on the WWW

Drygajlo A, Garcia Molina G
EPFL, Switzerland

The purpose of this paper is to introduce a user-friendly computer assisted learning (CAL) workbench in order to support traditional teaching in the Speaker Recognition area. The workbench (an interactive on-line laboratory) is based on Java and Java-enabled Web browser. The first prototype demonstrator developed at the Swiss Federal Institute of Technology Lausanne (EPFL) for speaker identification training consists of four modules: Dynamic Time Warping (DTW), hidden Markov Modelling (HMM), Vector Quantization (VQ) and Gaussian Mixture Modelling (GMM). These four modules aim at presenting, visualizing and investigating in a uniform way basic concepts of speaker recognition in a single user-friendly environment and allow for easy and highly illustrative learning through experiments with real speech data. They can be used for conventional classroom experiments, in the students' laboratory or can provide self-study means for distance learning applications or for further free exploration.

Volume 4, page 2787 Session E36a

Prototype of a Vocal-Tract Model for Vowel Production Designed for Education in Speech Science

Arai T, Usuki N, Murahara Y
Sophia University, Japan

We present a manipulative tool for use in education in speech science. The tool consists of several, square, plastic disks each of which has holes of various diameters. Combined, the holes in 10-17 disks simulate the vocal tract by creating an acoustic tube. Students may study the effect the shape of an acoustic tube has on acoustic output by reordering the disks. After demonstrating the disks in an actual classroom, results show that the tool helped students grasp the relationship between vocal tract configuration and acoustic output. This suggests that students are better able to understand abstractions regarding the acoustics of speech if they have access to a 3-dimensional model such as ours.

Volume 4, page 2791 Session E36a

A tool for automatic feedback on phonemic transcription

Cooke M[1], Garcia Lecumberri M L[2], Maidment J[3]
[1]University of Sheffield, UK, [2]University of the Basque Country, Spain, [3]University College London, UK

A tool which provides relevant feedback on learners' attempts at phonemic transcription is described. The tool aims to complement courses in transcription which are currently taught in both linguistics and language learning settings. A variety of types of feedback are provided. These can be staged by a tutor in order to support customization for different groups of learners and course levels. The tool consists of two similar standalone applications (for tutors and learners). The system performs an optimal alignment of student versus model transcriptions using a dynamic programming algorithm, modified to handle optional and alternative pronunciations. As a result, it computes a summary of errors and their locations within the attempt. Portability and internationalization are key design goals, supported in practice through the use of Java and XML. The tool is currently being tested in a controlled experiment which will provide considerable information on its actual usefulness and necessary refinements.

Volume 4, page 2795 Session E36a

Speech Lab in a Box: A Mandarin Speech Toolbox to Jumpstart Speech Related Research

Chang E, Shi Y, Zhou J, Huang C
Microsoft Research China, P. R. China

The necessity of gathering data has been an impediment for researchers and students who are interested in getting started in the fields related to speech recognition. We are proposing a new approach of distributing data that is designed to quickly help researchers and students achieve a set of baseline results to build upon. Furthermore, by leveraging publicly available programs, all researchers will be able to exactly reproduce results that are described in this paper. We also aim to facilitate comparison of recognition results in the field of Mandarin speech recognition by including a testing set in the toolbox. We describe a toolbox that includes Mandarin speech data from 125 speakers, suitable language model, scripts and data files required for recreating a set of baseline experiments, and a copy of Microsoft SAPI 5.0 SDK that can help professors and students who wish to jumpstart research programs in speech technologies. By lowering the barrier of entry to the field, we hope to encourage more participation in the study of Mandarin speech recognition.

Volume 4, page 2799 Session E36a

Relating PhonePass Scores Overall Scores to the Council of Europe Framework Level Descriptors

de Jong J H A L[1], Bernstein J[2]
[1]Language Testing Services, the Netherlands, [2]Ordinate Corporation, USA

This study is a preliminary report on an experiment relating PhonePass SET-10 scores to the scale of level descriptors in the Council of Europe Framework. This scale describes the content and level of second language proficiency from a functional communicative perspective. Speech samples from each of 121 non-native English speakers were (1) scored in SET-10, an automatic speaking test, and (2) rated under the European Framework through the conciliation of three independent raters. Rater reliability in using the Council of Europe scale and the comparability of the human and automatic measures are reported.

Volume 4, page 2803 Session E36a

A Multilingual, Multimodal, Speech Training System, SPECO

Vicsi K[1], Roach P[2], Öster A-M[3], Kacic Z[4], Csatári F[1], Sfakianaki A[2], Veronik R[4], Gordos G[1]
[1]Budapest University of Technology and Economics, Hungary, [2]University of Reading, UK, [3]Kungl. Tekniska Högskolan, Sweden, [4]University of Maribor, Slovenia

The SPECO Project was funded by the EU through the INCO-COPERNICUS program (Contract no. 977126) in 1999. In the frame of the project a system has been developed which is an audio-visual pronunciation teaching and training system for 5-10 year old children. Correction of disordered aspects of speech is done by real time visual presentation of the speech parameters, in a way that is understandable and interesting for young children, while remaining correct from the acoustic-phonetic point of view. The development of the speech by our teaching method is made mainly on the basis of visual information using the intact visual channel of the hearing impaired child. However during practice we use their limited auditory channel too, by giving auditory information synchronised with the visual one. This multi modal training and teaching system has been developed for languages of all the SPECO partners; these are English, Swedish, Slovenian and Hungarian.

Volume 4, page 2807 Session E36a

Instantaneous Estimation of Accentuation Habits for Japanese Students to Learn English Pronunciation

Nakamura N[1], Minematsu N[2], Nakagawa S[1]
[1]Toyohashi University of Technology, Japan, [2]University of Tokyo, Japan

More and more efforts have been recently made to apply speech technologies to language learning. The authors have been especially focusing on Japanese manners of generating English word stress. This is because accentuation habits inevitable to Japanese learners can be easily found in their stress generation. In our previous studies, a stressed syllable detector and an accentuation habit estimator were developed, where the estimated habits of individual learners accorded well with their English pronunciation proficiency rated by English teachers. However, the estimation methods in our previous studies required several dozens of word utterances or a relatively large amount of computation even when a single word utterance enabled the estimation. In this paper, we investigated a method which required only a single word utterance with a small computation cost. Results showed that similar tendencies can be found between the habits estimated in our previous study and in the current one.

Volume 4, page 2811 Session E36a

Automatic Construction of CALL System from TV News Program with Captions

Tanaka T, Mori K, Kobayashi S, Nakagawa S
Toyohashi University of Technology, Japan

Many language learning materials have been published in Japan. However, they are limited in their scope and content. In addition, we doubt whether the speech sounds found there are natural in various situations. These days, some TV news programs (by NHK, CNN, ABC, etc.) have closed/open captions corresponding to the speech of the announcer. We have developed a system that automatically makes CALL (Computer Assisted Language Learning) materials from such captioned newscasts. Materials compiled by this system have the following functions: repetition listening, consulting an electronic dictionary, and automatic construction of a dictation test. The materials have the following advantages: polite and natural speech sound, various and timely topics, and abundant materials. In this paper, we describe the organization of our new system.

Volume 4, page 2815 Session E36a

Session E36b - Poster
Friday - 13.30 - 14.50

Speaker Recognition: Features and Robustness

Chair: George Kokkinakis, Univ. of Patras, Greece

Pitch-Dependent GMMs for Text-Independent Speaker Recognition Systems

Arcienega M, Drygajlo A
EPFL, Switzerland

Gaussian mixture models (GMMs) and ergodic hidden Markov models (HMMs) have been successfully applied to model short-term acoustic vectors for speaker recognition systems. Prosodic features are known to carry information concerning the speaker's identity and they can be combined with the short-term acoustic vectors in order to increase the performance of the speaker recognition system. In this paper, a statistical approach using pitch-dependent GMMs for modeling speakers is presented. This new approach is capable of simultaneously modeling the statistical distributions of the short-term acoustic vectors and long-term prosodic features

Volume 4, page 2821 Session E36b

Towards combining pitch and MFCC for speaker recognition systems

Ezzaidi H[1], Rouat J[1], O'Shaughnessy [2]
[1]DSA, ERMETIS, Université du Québec à Chicoutimi, Canada,
[2]INRS-Télécommunications, Université du Québec, Canada

Usually, speaker recognition systems do not take into account the dependence between the vocal source and the vocal tract. A feasibility study that retains this dependence is presented here. A model of joint probability functions of the pitch and the feature vectors is proposed. Three strategies are designed and compared for all female speakers taken from the SPIDRE corpus. The first operates on all voiced and unvoiced speech segments (baseline strategy). The second strategy considers only the voiced speech segments and the last includes the pitch information along with thestandard MFCC. We use two pattern recognizers: LVQ--SLP and GMM. In all cases, we observe an increase in the identification rates and more specifically when using a time duration of 500ms (6% higher).

Volume 4, page 2825 Session E36b

Formant-Broadened CMS Using Peak-Picking in LOG Spectrum

Kim Y-J, Jung H-K, Chung J-H
Inha Univ., Korea

In this paper, we propose a method to remove the residual speech effects of the channel cepstrum for speaker recognition in the Cepstral Mean Subtraction framework. The proposed Formant-Broadened CMS(FBCMS) is based on the facts that the formants can be found easily in log spectrum which is transformed from the cepstrum and the formants correspond to the dominant poles of all-pole model which is usually modeled vocal tract. The FBCMS evaluates only poles to be broadening from the log spectrum without polynomial factorization and makes a formant-broadened cepstrum by broadening the bandwidths of formant poles. Using 8 simulated telephone channels, we compared the relative errors of estimating channel cepstrum, speaker identification and computational efficiency for CMS, PFCMS, and the proposed method respectively on two databases. The proposed method has shown to yield improved speaker recognition rates with lower computational burden.

Volume 4, page 2829 Session E36b

Improvements in the speaker identification rate using feature-sets

Mashao D J, Baloyi N T
University of Cape Town, South Africa

In this paper we look at the parameterized feature-set that has been used in connected alpha-digit speech recognition and evaluate it on a speaker identification system. Compared to the popular mel-scaled feature-set (MFCC) the parameterized feature-set gives over 21% improvement in identification rate on the NTIMIT database in some cases. On average it shows a 14.0% improvement in identification rate. This demonstrates the improvement in performance that can be obtained using feature-sets.

Volume 4, page 2833 Session E36b

Minimum Classification Error Training for Speaker Identification Using Gaussian Mixture Models Based on Multi-Space Probability Distribution

Miyajima C, Tokuda K, Kitamura T
Nagoya Institute of Technology, Japan

In our previous work, we have proposed a speaker modeling technique using spectral and pitch features for text-independent speaker identification based on Multi-Space Probability Distribution Gaussian Mixture Models (MSD-GMMs). We have presented a maximum likelihood (ML) estimation procedure for the MSD-GMM parameters and demonstrated its high recognition performance. In this paper, we describe an minimum classification error (MCE) training procedure for the MSD-GMM speaker models. MCE training is also applied to automatically estimate mixture-dependent stream weights for spectral and pitch streams. The MCE-based MSD-GMM speaker models are evaluated for a text-independent speaker identification task. Experimental results show that MCE training of the MSD-GMM parameters significantly reduces identification errors and system performance is further improved by appropriately weighting spectral and pitch streams using MCE training.

Volume 4, page 2837 Session E36b

Speaker Recognition based on Feature Space Trace

Wu Y, Li Z
Shanghai Jiaotong University, P. R. China

This paper presents a multiple templates matching algorithm based on feature space trace, which is used in speaker recognition. It extracts the cepstrum coefficient as feature parameter. We normalize the sequence of feature parameter based on feature space trace. The fuzzy c-means method is adopted in generating the multiple templates and the multiple matching method is applied to match the templates. Experiments show this method is a robust, high recognizable and valuable way to implement text-dependent speaker recognition.

Volume 4, page 2841 Session E36b

Additive and Convolutional Noise Canceling in Speaker Verification Using a Stochastic Weighted Viterbi Algorithm

Yoma N B, Villar M
University of Chile, Chile

This paper replaces the ordinary output probability with its expected value if the addition of noise is modeled as a stochastic process, which in turn is merged with the HMM in the Viterbi algorithm. The method, which can be seen as a weighted matching algorithm, is applied in combination with spectral subtraction and RASTA to improve the robustness to additive and convolutional noise of a text-dependent speaker verification system. Reductions around 10% or 20% in the error rates and improvements as high as 30% or 50% in the stability of the decision thresholds are reported when the ordinary Viterbi algorithm is replaced with the weighted one. When compared with the baseline system, reductions of 70% or 80% are shown.

Volume 4, page 2845 Session E36b

A Multi-SNR Subband Model for Speaker Identification under Noisy Environments

Yoshida K, Takagi K, Ozeki K
The University of Electro-Communications, Japan

The model presented in this paper consists of a set of subband GMMs trained on speech data corrupted with white Gaussian noise at several SNRs. In the recognition stage, an optimal GMM that yields the maximum accumulated likelihood on the whole input frames is selected for each subband. Then the likelihood is recombined over the subbands to give a speaker identification score. To evaluate the performance of this model, text independent speaker identification experiments were conducted under 5 different noisy environments. For comparison, performance evaluation was also conducted on 3 other models: a subband model trained on clean speech, a multi-SNR fullband model, and a fullband model trained on clean speech. Results show that the multi-SNR subband model is very effective under a wide variety of noisy environments. Additional improvement was observed when an optimal GMM was selected on a short term basis instead of a whole input basis.

Volume 4, page 2849 Session E36b