CONFERENCE PROGRAM AND ABSTRACT BOOK

September 17 – 21, 2006  Ninth International Conference on Spoken Language Processing

http://www.interspeech2006.org
Interspeech 2006 — ICSLP
Table of Contents

Welcome Messages ........................................................................................................... 5
Organizing and Review Committees ............................................................................... 9
Corporate sponsors ......................................................................................................... 14
Satellite Events ............................................................................................................... 17
Future Interspeech Meetings ......................................................................................... 19
General Information ....................................................................................................... 21
Maps of Pittsburgh and the Conference Venues ........................................................... 23
Arrival and parking ......................................................................................................... 24
Information about Pittsburgh ......................................................................................... 25
Maps of the Conference Rooms at the Westin .............................................................. 28
General Interspeech 2006 Information .......................................................................... 29
Interspeech 2006 Schedule ........................................................................................... 32
Tutorials ......................................................................................................................... 37
Plenary Sessions ........................................................................................................... 43
Special Sessions ............................................................................................................ 47
Session Table of Contents ........................................................................................... 50
Abstracts ....................................................................................................................... 52
Author Index .................................................................................................................. 163
A Message from the President of ISCA

On behalf of the International Speech Communication Association (ISCA), I welcome you to Pittsburgh and to INTERSPEECH 2006. INTERSPEECH conferences (formerly known as Eurospeech and ICSLP conferences) have established themselves as one of the major conference series in the field of speech communications science and technology, starting with the first Eurospeech conference in Paris in 1989 and the first ICSLP conference in 1990 in Kobe. ISCA as an organization builds upon the earlier organization ESCA and the Permanent Council of the ICSLP to form a truly international organization, supporting conferences, workshops, and other scientific activities throughout the world. The primary goal of ISCA is “to promote Speech Communication Science and Technology, both in the industrial and academic areas” in all speech technologies, including acoustics, phonetics, phonology, natural language processing, cognitive science, signal processing, and pattern recognition. Previous INTERSPEECH conferences have been held in Genoa (1991), Banff (1992), Berlin (1993), Yokohama (1994), Madrid (1995), Philadelphia (1996), Rhodes (1997), Sydney (1998), Budapest (1999), Beijing (2000), Aalborg (2001), Denver (2002), Geneva (2003), Jeju (2004), and Lisbon (2005). Future conferences will be held in Antwerp (2007), Brisbane (2008), and Brighton (2009). We welcome bids from any region for INTERSPEECH 2010.

In the past year, the ISCA board has pursued a number of projects aimed at making ISCA more international and more responsive to the needs of its members. These include:

- A renovated website which makes it possible to become an ISCA member, to update one’s personal information online, and to apply for ISCA grants on line, with future improvements in membership services to come;
- A student committee with its own activities and website, including pointers to shareable software and job postings;
- ISCA-sponsored workshops in Aalborg, Denmark (ASIDE 2005: Applied Spoken Language Interaction in Distributed Environments); Stellenbosch, South Africa (MULTILING 2006: ISCA Workshop on Multilingual Speech and Language Processing); Toulouse, France (ITRW on Speech Recognition and Intrinsic Variation); Athens, Greece (ITRW on Experimental Linguistics); as well as workshops held in conjunction with this conference: Berlin, Germany (ITRW on Perceptual Quality of Systems); Athens, Greece (SAPS 2006: ITRW on Statistical and Perceptual Audition). Additional workshops planned include: NOLISP 2007 in Paris, France (May 22-25); SSW-6; the 6th ISCA Speech Synthesis Research Workshop in Bonn, Germany (August 22-24, 2007); and a ISCA ITRW on Robustness, Santiago, Chile (Nov, 2007)
- Continued additions to the ISCA archive of current and older conference and workshop papers, available in full paper form to members;
- New Special Interest Groups (SIGs) in Iberian Languages and in Indian Languages;
- New international committees representing Eastern Europe, Western Asia and Northern Africa, Southern Asia, Central and Southern Africa.
• More student paper awards (3) to be presented at INTERSPEECH conferences, beginning at this year’s conference in Pittsburgh;
• Continued surveys of the membership to see what other activities and services ISCA can support and continued distributing to members of ISCA’s monthly newsletter, ISCApad.

For INTERSPEECH 2006, we offer many thanks to the organizers, a large team of researchers from academics and industry led by Rich Stern (CMU). They have done a wonderful job of planning a conference that will be scientifically interesting and have many opportunities for fun as well. We also thank the organizations that have helped to fund INTERSPEECH 2006, the many members of the international community who participated in reviewing papers submitted to the conference, and who organized special sessions and satellite workshops. Most of all, we thank all of you who have come to attend this conference and we extend our best wishes for a very interesting and enjoyable stay in Pittsburgh.

Julia Hirschberg

ISCA President
Welcome Message from the General Chair

On behalf of everyone on the Interspeech 2006 Organizing Committee, it is truly my pleasure to welcome you to Pittsburgh to participate in Interspeech 2006, the Ninth International Conference on Spoken Language Processing (ICSLP). The ICSLP conferences, along with the Eurospeech conferences that have taken place in alternate years, have become the primary focal point for the exchange of ideas in a broad array of fields centered around human-human and human-machine speech communication. Having attended every ICSLP since the initial conference in Kobe in 1990, I am quite familiar with both the very high level of scientific and technical discourse that these meetings foster, as well as the valuable opportunities for both formal and informal personal interaction with colleagues and friends from around the world. The ICSLP and Eurospeech meetings have been sponsored by the International Speech Communication Association (ISCA) since 2000, and are now collectively known as the Interspeech conferences. Our meeting in Pittsburgh follows well-organized Interspeech conferences in Beijing, Aalborg, Denver, Geneva, Jeju Island, and Lisbon. The Interspeech community is truly an interdisciplinary community, and our annual conferences bring together scientists in speech and natural language processing, acoustic phoneticians, linguists and psycholinguists, electrical engineers and computer scientists, and researchers in auditory physiology and perception, among many other disciplines. This is a unique opportunity for us both to deepen our understanding of our core research foci and at the same time broaden our perspectives by exposure to the best work in complementary disciplines.

As is the case with every conference, the primary attraction for most attendees is the high-quality papers that are presented and discussed. This year we were extremely fortunate to have had the advice of 448 expert technical reviewers who reviewed 1033 submitted papers, accepting 662 of them (64 percent) for presentation at Interspeech 2006. The accepted papers were molded into a coherent conference program by the 29 members of the Technical Program Committee whose names appear later in this volume. The technical program includes 35 oral sessions, 32 poster sessions, and 7 special sessions that have been organized by their Chairs around issues of vital topical importance. The four days of regular technical sessions are preceded by a day of six tutorial presentations that were selected by the Organizing Committee from a much larger number of excellent proposals. Finally, we encourage all participants to attend the plenary talks by John Ohala (this year’s ISCA Medalist), Mike Phillips, Elissa Newport, and Raj Reddy, which span topics from fundamental speech science to the industrial application of speech and language technology, and to their impact on society.

The primary technical program will be complemented by a set of social events, including an opening reception Monday evening at Heinz Hall, the home of the Pittsburgh Symphony, a reception organized by and for students at the nightclub Dowe’s on Ninth on Tuesday evening, and the conference banquet that will take place Wednesday evening in the Senator John Heinz Pittsburgh Regional History Center. All of these locations are an easy walk from the conference venue at the Westin Pittsburgh Hotel and Convention Center.

I am especially pleased to be bringing the Interspeech community to Pittsburgh, Pennsylvania, the Renaissance City. The convergence of the Allegheny and Monongahela Rivers not far from the
conference site was one of the major focal points of the struggle between the British and French for
domination of North America in the 18th century. The city of Pittsburgh, which developed around the
military fortifications, became the leading center in the world for steel production in the 19th century and
remained so until about 25 years ago. Pittsburgh today is a picturesque city of three rivers and hundreds
of bridges that is now centered on computing and information technology, biomedical research, and
education. Pittsburgh has also become a world-renown center for speech and language processing through
the work of the Language Technologies Institute along with other departments and centers at Carnegie
Mellon University and at the University of Pittsburgh, and the many local speech and language-based
companies. We will be highlighting much of this work at a reception Sunday evening in the Westin
Hotel.

I only now fully appreciate what an enormous effort the organization of an Interspeech conference
entails! The Interspeech 2006 meeting came about only with the help of many people and organizations,
a few of which I would like to acknowledge here. First, I must thank all of the members of the
Interspeech 2006 Organizing Committee who individually and collectively have worked for almost three
years now to bring this meeting together. Everyone has been generous with their time and talents beyond
my most optimistic expectations, and it is hard to believe that our work together on this project is about to
come to its conclusion. We could not have proceeded without ISCA’s institutional and organization
leadership, and I would like to give special thanks to present and past ISCA leaders and conference
organizers Hiroya Fukisaki, Sadaoki Furui, John Hansen, Julia Hirschberg, and Isabel Trancoso for their
advice, encouragement, and support throughout these three years. I also have received a crash course in
the financial stresses associated with organizing an Interspeech meeting, and I thank our corporate
sponsors, which range from some of the world’s largest companies to young local startups, for their
support of this conference and the Interspeech meetings in general. Finally, I thank the Language
Technologies Institute, the Electrical and Computer Engineering Department, and the Institute for
Software Research International at CMU for their willingness to underwrite this meeting.

So once again, welcome to Interspeech 2006 and to Pittsburgh! I hope that you find the technical
discussions and personal interchanges to be productive and rewarding. While in town I also hope that you
will take a run along the beautiful North Shore of the Allegheny River, visit the nearby Andy Warhol
Museum, watch the Pittsburgh Pirates play baseball, or attend one of the numerous local concerts and
theatrical productions before and after the meeting.

Richard Stern

General Chair, Interspeech 2006
Interspeech 2006 Organizing Committee

General Chair
Richard M. Stern
Carnegie Mellon University

General Co-Chairs
Alan W Black
Carnegie Mellon University
Tanja Schultz
Carnegie Mellon University

Technical Program Chair
Amro El-Jaroudi
University of Pittsburgh

Special Sessions
Maxine Eskenazi
Carnegie Mellon University
Diane J. Litman
University of Pittsburgh

Tutorials
Monika Woszczyna
Multimodal Technologies Inc

Satellite Workshops
Laura Mayfield Tomokiyo
Cepstral LLC

Finance
Alexander Rudnicky
Carnegie Mellon University

Publicity
Lori L. Holt
Carnegie Mellon University

Registration and Webmaster
Robert Frederking
Carnegie Mellon University

Website Implementation
Brian Langner
Carnegie Mellon University

Exhibits and Industrial Liaison
Alex Acero
Microsoft Research
Michael Picheny
IBM T. J. Watson Research Center

Volunteer Coordinator
Ian Lane
Carnegie Mellon University

Members at Large
Kevin Lenzo
Cepstral LLC
Lazaros Polymenakos
Athens Information Technology

Interspeech 2006 was organized by:
Carnegie Mellon University
http://www.cmu.edu
The University of Pittsburgh
http://www.pitt.edu
with the support of
The International Speech Communication Association
http://www.isca-speech.org
The Westin Convention Center Pittsburgh
http://www.westin.com/pittsburgh
Conferon, Inc.
http://www.conferon.com
Technical Program Committee

Amro El-Jaroudi (Chair)
University of Pittsburgh

Srinivas Bangalore
AT&T Research

Alan W Black
Carnegie Mellon University

Antonio Bonafonte
Universitat Politècnica de Catalunya

Lou Boves
University of Nijmegen

Dirk van Compernolle
University of Leuven

Maxine Eskenazi,
Carnegie Mellon University

Hiroya Fujisaki
University of Tokyo

Sadaoki Furui
Tokyo Institute of Technology

Mark Gales
University of Cambridge

Dafydd Gibbon
University of Bielefeld

John Hansen
University of Texas, Dallas

Julia Hirschberg
Columbia University

Michael Johnston
AT&T Research

Sanjeev Khudanpur
The Johns Hopkins University

Katrin Kirchhoff
University of Washington

Chin-Hui Lee
Georgia Institute of Technology

Pedro Moreno
Google, Inc.

Hermann Ney
RWTH Aachen

Mosur Ravishankar
Carnegie Mellon University

Doug Reynolds
Lawrence Livermore National Lab

Giuseppe Riccardi
University of Trento

Tanja Schultz
Carnegie Mellon University

Mike Seltzer
Microsoft Research

Liz Shriberg
SRI International

Richard M. Stern
Carnegie Mellon University

Isabel Trancoso
L2F, INESC ID

Dilek Hakkani-Tür
International Computer Science Institute

Bayya Yegnanarayana
International Institute of Information Technology
Scientific Review Committee

Alex Acero
Gilles Adda
Martine Adda-Decker
Masato Akagi
Paavo Alku
Abeer Alwan
Eliathamby Ambikairajah
Noam Amir
Tim Anderson
Amália Andrade
Ernesto Andrade
Elisabeth Andre
Takayuki Arai
Masahiro Araki
Victoria Arranz
Khaled Assaleh
Matthew Aylett
Pierre Badin
KeunSung Bae
Gerard Bailly
Janet Baker
Srinivas Bangalore
Plinio Barbosa
Etienne Barnard
Dante Barone
Anton Batliner
Steve Beet
Nuno Beires
Niels Ole Bernsen
Laurent Besacier
Alan W Black
Mats Blomberg
Gerrit Bloothooft
Antonio Bonafonte
Jean-Francois Bonastre
Zinny Bond
Christos Boukis
Hervé Bourlard
Lou Boves
Udo Bub
Tim Bunnell
Joseph P. Campbell
Nick Campbell
William Campbell
Jean Carletta
Rolf Carlson
Francisco Casacuberta
Diamantino Caseiro
Joyce Chai
Fang Chen
Yan-Ming Cheng
Rathi Chengalvarayan
Jen-Tzung Chien
Tseng Chiu-yu
Gerard Chollet
Khalid Choukri
Hyun-Yeol Chung
Robert Clark
Mark A. Clements
Jordan Cohen
Phil Cohen
Luisa Coheur
Andrew Cole
Thomas Colthurst
Alistair Conkie
Martin Cooke
Ricardo de Cordoba
Stephen Cox
Catia Cucchiarini
Suzanne Curtin
Walter Daelemans
Christophe D'Alessandro
Paul Dalsgaard
Marelie Davel
Chris Davis
Renato De Mori
Michael Deisher
Kris Demuynck
Matthias Denecke
Li Deng
Laurence Devillers
Giuseppe Di Fabbrizio
Vassilis Digalakis
Simon Doclo
Hans Dolfing
Christine Doran
Ellen Douglas-Cowie
Christoph Draxler
Jasha Droppo
Andrzej Drygajlo
Sorin Dusan
Thierry Dutoit
Laila Dybkjær
Robert Eklund
Amro El-Jaroudi
Dan Ellis
Thomas Eriksson
Maxine Eskenazi
Kevin Farrell
Marcos Faundez-Zanuy
Junlan Feng
Eric Fosler-Lussier
Horacio Franco
Nathan Franzen
Robert Frederking
Diamantino Freitas
Juergen Fritsch
Sónia Frota
Christian Fuegen
Hiroya Fujisaki
Toshiaki Fukada
Sadaoki Furui
Sharon Gannot
Jean-Luc Gauvain
Panayiotis Georgiou
Dafydd Gibbon
Herbert Gish
Diego Giuliani
James Glass
Christer Gobl
Roland Goecke
Yifan Gong
Joaquin Gonzalez-Rodriguez
Allen Gorin
Silke Goronzy
Mária Gösy
Philippe Gournay
Bjorn Granstrom
Phil Green
Steven Greenberg
Ellen Haas
Reinhold Haeb-Umbach
Thomas Hain
Dilek Hakkani-Tür
Wael Hamza
John H.L. Hansen
William Hardcastle
Ramalingam Harirhan
Mary Harper
Jean-Paul Haton
Valerie Hazan
Timothy Hazen
Matthieu Hebert
<table>
<thead>
<tr>
<th>Scientific Review Committee (continued)</th>
</tr>
</thead>
<tbody>
<tr>
<td>José Manuel Pardo</td>
</tr>
<tr>
<td>Hansang Park</td>
</tr>
<tr>
<td>Jong Cheol Park</td>
</tr>
<tr>
<td>Sarangarajan Parthasarathy</td>
</tr>
<tr>
<td>Steffen Pauws</td>
</tr>
<tr>
<td>David Pearce</td>
</tr>
<tr>
<td>Jason Pelecanos</td>
</tr>
<tr>
<td>François Pellegrino</td>
</tr>
<tr>
<td>David Pepper</td>
</tr>
<tr>
<td>Fernando Perdigão</td>
</tr>
<tr>
<td>Pascal Perrier</td>
</tr>
<tr>
<td>Joe Picone</td>
</tr>
<tr>
<td>Roberto Pieraccini</td>
</tr>
<tr>
<td>Aristodemos</td>
</tr>
<tr>
<td>Pneumatiakakis</td>
</tr>
<tr>
<td>Louis C.W. Pols</td>
</tr>
<tr>
<td>Thomas Portele</td>
</tr>
<tr>
<td>Daniel Povey</td>
</tr>
<tr>
<td>Rohit Prasad</td>
</tr>
<tr>
<td>Kristin Precoda</td>
</tr>
<tr>
<td>Patti Price</td>
</tr>
<tr>
<td>Qi féng Zhu</td>
</tr>
<tr>
<td>Thomas F. Quatieri</td>
</tr>
<tr>
<td>Hugo Quene</td>
</tr>
<tr>
<td>Sitaram R.N.V.</td>
</tr>
<tr>
<td>Lawrence R. Rabiner</td>
</tr>
<tr>
<td>Bhiksha Raj</td>
</tr>
<tr>
<td>Bhuvana Ramabhadran</td>
</tr>
<tr>
<td>Mosur Ravishankar</td>
</tr>
<tr>
<td>Daniel Recasens</td>
</tr>
<tr>
<td>Steve Renals</td>
</tr>
<tr>
<td>Doug Reynolds</td>
</tr>
<tr>
<td>Fernando Gil Vianna</td>
</tr>
<tr>
<td>Resende Jr.</td>
</tr>
<tr>
<td>Douglas Reynolds</td>
</tr>
<tr>
<td>Giuseppe Riccardi</td>
</tr>
<tr>
<td>Gerhard Rigoll</td>
</tr>
<tr>
<td>Michael Riley</td>
</tr>
<tr>
<td>Olivier Rosec</td>
</tr>
<tr>
<td>Antti-Veikko Rosti</td>
</tr>
<tr>
<td>Alexander Rudnicky</td>
</tr>
<tr>
<td>Yoshinori Sagisaka</td>
</tr>
<tr>
<td>Emilio Sanchis Arnal</td>
</tr>
<tr>
<td>Eric Sanders</td>
</tr>
<tr>
<td>Rubén San-Segundo</td>
</tr>
<tr>
<td>George Saon</td>
</tr>
<tr>
<td>Shimon Sapiro</td>
</tr>
<tr>
<td>Ruhi Sarikaya</td>
</tr>
<tr>
<td>Hiroshi Saruwatari</td>
</tr>
<tr>
<td>Thomas Schaaf</td>
</tr>
<tr>
<td>Ralf Schlüter</td>
</tr>
<tr>
<td>Astrid Schmidt-Nielsen</td>
</tr>
<tr>
<td>Marc Schröder</td>
</tr>
<tr>
<td>Juergen Schroeter</td>
</tr>
<tr>
<td>Tanja Schultz</td>
</tr>
<tr>
<td>Holger Schwenk</td>
</tr>
<tr>
<td>Mike Seltzer</td>
</tr>
<tr>
<td>Stephanie Seneff</td>
</tr>
<tr>
<td>Cheol Jae Seong</td>
</tr>
<tr>
<td>Antônio Serralheiro</td>
</tr>
<tr>
<td>Izhar Shafran</td>
</tr>
<tr>
<td>Hamid Sheikhzadeh</td>
</tr>
<tr>
<td>Chilin Shih</td>
</tr>
<tr>
<td>Koichi Shinoda</td>
</tr>
<tr>
<td>Elizabeth Shriberg</td>
</tr>
<tr>
<td>Carlos Silva</td>
</tr>
<tr>
<td>Olivier Siohan</td>
</tr>
<tr>
<td>R.N.V. Sitaram</td>
</tr>
<tr>
<td>Malcolm Slaney</td>
</tr>
<tr>
<td>Raymond Slyh</td>
</tr>
<tr>
<td>M. Mohan Sondhi</td>
</tr>
<tr>
<td>Victor Sorokin</td>
</tr>
<tr>
<td>Mark Steedman</td>
</tr>
<tr>
<td>Volker Steinbiss</td>
</tr>
<tr>
<td>Andreas Stergiou</td>
</tr>
<tr>
<td>Brad Story</td>
</tr>
<tr>
<td>Helmer Strik</td>
</tr>
<tr>
<td>Sebastian Stüker</td>
</tr>
<tr>
<td>Yannis Stylianou</td>
</tr>
<tr>
<td>Arun Suren-drnan</td>
</tr>
<tr>
<td>Marc Swerts</td>
</tr>
<tr>
<td>Ann Syrdal</td>
</tr>
<tr>
<td>David Talkin</td>
</tr>
<tr>
<td>Paul Taylor</td>
</tr>
<tr>
<td>Antônio Teixeira</td>
</tr>
<tr>
<td>Carlos Teixeira</td>
</tr>
<tr>
<td>Erik Thiessen</td>
</tr>
<tr>
<td>Matthias Thomae</td>
</tr>
<tr>
<td>Jilei Tian</td>
</tr>
<tr>
<td>Tomoki Toda</td>
</tr>
<tr>
<td>Keiichi Tokuda</td>
</tr>
<tr>
<td>Laura Tomokiyo</td>
</tr>
<tr>
<td>Isabel Trancoso</td>
</tr>
<tr>
<td>Roger Tucker</td>
</tr>
<tr>
<td>Gokhan Tur</td>
</tr>
<tr>
<td>Dirk van Compernolle</td>
</tr>
<tr>
<td>Antal van den Bosch</td>
</tr>
<tr>
<td>Henk van den Heuvel</td>
</tr>
<tr>
<td>Saeed Vaseghi</td>
</tr>
<tr>
<td>Eric Vatikiotis-Bateson</td>
</tr>
<tr>
<td>Jennifer Venditti</td>
</tr>
<tr>
<td>Céu Viana</td>
</tr>
<tr>
<td>Marina Vigário</td>
</tr>
<tr>
<td>Hsiao-Chuan Wang</td>
</tr>
<tr>
<td>Hsin-min Wang</td>
</tr>
<tr>
<td>Kuansan Wang</td>
</tr>
<tr>
<td>Ye-Yi Wang</td>
</tr>
<tr>
<td>Nigel Ward</td>
</tr>
<tr>
<td>Christian Wellekens</td>
</tr>
<tr>
<td>Stanley Wenndt</td>
</tr>
<tr>
<td>Martin Westphal</td>
</tr>
<tr>
<td>Robert Whitman</td>
</tr>
<tr>
<td>Briony Williams</td>
</tr>
<tr>
<td>Monika Woszczyka</td>
</tr>
<tr>
<td>Johan Wouters</td>
</tr>
<tr>
<td>John Zhiyong Wu</td>
</tr>
<tr>
<td>Bing Xiang</td>
</tr>
<tr>
<td>Bayya Yegnanarayana</td>
</tr>
<tr>
<td>Néstor Becerra Yoma</td>
</tr>
<tr>
<td>Dongsuk Yook</td>
</tr>
<tr>
<td>Dong Yu</td>
</tr>
<tr>
<td>Hua Yu</td>
</tr>
<tr>
<td>Young-Sun Yun</td>
</tr>
<tr>
<td>Stephen Zahorian</td>
</tr>
<tr>
<td>Milos Zelezny</td>
</tr>
<tr>
<td>Andrej Zgank</td>
</tr>
<tr>
<td>Tong Zhang</td>
</tr>
<tr>
<td>Yunxin Zhao</td>
</tr>
<tr>
<td>Jing Zheng</td>
</tr>
<tr>
<td>Bowen Zhou</td>
</tr>
<tr>
<td>Imed Zitouni</td>
</tr>
<tr>
<td>Geoffrey Zweig</td>
</tr>
</tbody>
</table>
Sponsors

The Interspeech 2006 Organizing Committee sincerely thanks all sponsors for their generous support and contributions.
Pittsburgh Area Sponsors

The Greater Pittsburgh region is home to many vibrant enterprises in speech and language technologies. We are grateful for our local sponsors for their support of Interspeech 2006 and for their having nurtured a dynamic and growing professional community.

Acusis® is a Higher Standards™ global medical transcription services company with offices in the United States and a transcription work force and software development team in India. Acusis offers world-class quality service with guaranteed 365/24/7 reports with tailored turnaround time of 24 hours or less. By using its own unique software – AcuSuite™ – Acusis offers customers unparalleled flexibility with access to transcription tracking during the entire process. Acusis is dedicated solely to the business of quality medical transcription for its’ customers – hospitals, clinics and physician practices – using all methods of data collection including digital recording, telephone dictation and PC transcription. With over 750 associates, Acusis offers the immediate capacity to address a wide range of transcription needs, from overflow to total outsourced solutions. http://www.acusis.com

Akustica, Inc. develops silicon microphones for sale directly to cell phone, mobile computer, and digital media Original Equipment Manufacturers (OEMs.) Silicon microphones are smaller and more reliable than current electret condenser microphones (ECMs), and can be customized with advanced sound capture features and noise reduction capabilities.

Through a revolutionary technology known as Sensory Silicon™, Akustica products enable electronic products to sense and respond to the world around them. By leveraging standard complementary metal oxide semiconductor (CMOS) processes and micro-electromechanical systems (MEMS) technology, Akustica acoustic system-on-chip solutions combine the functionality of microphones with microelectronics and software onto a single chip. Only the Akustica CMOS MEMS platform enables single-chip solutions with arrays of transducers and integrated signal processing that disrupt both conventional microphone and speaker technologies. http://www.akustica.com

Carnegie Speech is a Carnegie Mellon University spinoff producing software that uses speech technology to help students learn a new language. Its products detect specific errors in non-native speech and show the student where those errors are and how to correct them. The use of intelligent tutoring enables students to go at their own individual speed, spending time on what they specifically need to learn. NativeAccent™ is being used around the world to help adults to speak more understandable English. More information on the Carnegie Speech Company can be found at http://www.carnegiespeech.com
At Cepstral®, Text-to-Speech is our only focus. We make realistic synthetic voices that can say anything, anywhere, with personality and style. From the smallest device to large installations and high-end interactive media, Cepstral voices can bring fresh content to your ears, on demand.  
http://www.cepstral.com

Lucas Systems, the voice of Jennifer™, provides voice-directed operations solutions to warehouses and distribution centers, combining best-of-breed devices and voice technology with our applications, integration skills and industry knowledge to provide our customers a tailored solution. Jennifer is used by thousands of employees to greatly improve accuracy, efficiency, and reliability. Lucas has extensive experience across industries, including food and beverage, retail, manufacturing, and healthcare. We have been providing voice-directed solutions longer and more successfully than any other company, with customer results to back it up. Most importantly, our customer successes have been our greatest asset and engine for growth. http://www.lucasware.com

M*Modal is a fast-moving speech technology company based in Pittsburgh, PA. Our portfolio of conversational speech recognition, natural language understanding, and biometrics technologies is widely recognized as the most advanced in the industry. We are a leading innovator in the field of conversational documentation services (CDS) – where speech recognition and natural language understanding are combined in a unique setup targeted to truly understand conversational speech and turn it directly into actionable and meaningful data. Our proprietary speech understanding technology – operating on M*Modal's computing grid located at our national data center – is already redefining the way clinical information is captured in healthcare. Hundreds of hours of speech are processed daily and are converted into structured and encoded clinical documents. If you think you know everything about speech already, you're in for a surprise. http://www.mmodal.com

Since 1987, Voccoliect, Inc. has delivered dramatic improvements in productivity, accuracy, cost reduction and worker satisfaction for mobile employees. Voccoliect’s voice-directed work applications literally talk people through their daily tasks, replacing traditional work lists and cumbersome data capture methods with personal voice dialogs. The company’s global network of resellers and supply chain performance experts enables more than 100,000 people on six continents to use voice to improve work every day. Voccoliect is headquartered in Pittsburgh, PA USA, and supports its clients and resellers through regional offices in Europe, Latin America and Asia. http://www.voccoliect.com
Satellite Events

Interspeech 2006 is proud to be associated with the following satellite events:

**SAPA2006**

**Date:** Saturday, September 16  
**Location:** Allegheny III  
**Organizers:** Bhiksha Raj, Daniel Ellis, Paris Smaragdis, Malcom Slaney

The objective of the Statistical and Perceptual Audition (SAPA) workshop is to bring together researchers considering perceptually-motivated problems in sound and speech analysis and understanding, employing statistical and machine learning tools.


**Blizzard Challenge 2006**

**Date:** Saturday, September 16  
**Location:** Allegheny I  
**Organizer:** Alan W Black

In a continuing effort to evaluate corpus-based speech synthesis techniques on a common data set, the Blizzard Challenge 2006 workshop will report results and systems used in this year's challenge. 14 different systems will be presented with results.

[http://www.festvox.org/blizzard/](http://www.festvox.org/blizzard/)

**Young Researchers’ Roundtable on Spoken Dialog Systems**

**Date:** Saturday, September 16  
**Location:** University of Pittsburgh Campus  
**Organizers:** Matthias Denecke, Alex Gruenstein, Verena Rieser, Mihai Rotaru, Matt Stuttle

The Young Researchers’ Roundtable on Spoken Dialog Systems is an annual workshop designed for students, post docs, and junior researchers working in applied spoken dialog systems research. The roundtable provides an open forum where participants discuss their research interests, current work and future plans. The workshop is meant to foster creative thinking about current issues in spoken dialog systems research, and help create a stronger international network of young researchers in the field.

Satellite Events (continued)

Dialogue on Dialogues: Multidisciplinary Evaluation of Advanced Speech-based Interactive Systems

**Date:** Sunday, September 17  
**Location:** Cambria Room  
**Organizers:** Michael McTear, Kristiina Jokinen, Jim Larson

The goal of this one-day workshop is to encourage dialogue on dialogues: multidisciplinary discussions on the evaluation criteria for speech-based systems, descriptions of completed or on-going research, and debates on what is still needed. The workshop aims to bring together researchers and practitioners working on the development of communication models that support robust and efficient interaction in spoken natural language both for commercial dialogue systems and in basic research and, through discussions and working sessions, identify and enumerate criteria for evaluating robust and efficient interaction in spoken dialogue systems.

[http://www.ling.helsinki.fi/~kjokinen/ICSLP06-DoD/](http://www.ling.helsinki.fi/~kjokinen/ICSLP06-DoD/)

Planning meeting for a Computational Linguistics Olympiad

**Date:** Sunday, September 17  
**Location:** Butler Room  
**Organizer:** Lori Levin

This is an invitation-only workshop to plan a high school level Olympiad contest in Computational Linguistics (CLO) that will take place in 2007. Because computational linguistics is not part of high school curricula, the goal of the CLO is to identify talented high school students, introduce them the possibility of careers in human language technologies, and help them prepare for their undergraduate and graduate education.
INTERSPEECH 2007 is the eighth conference in the annual series of INTERSPEECH events and also the tenth biennial EUROspeech conference. It is jointly organized by partners from the Netherlands and Belgium and it will be held in Antwerp, Belgium, under the sponsorship of the International Speech Communication Association (ISCA).

INTERSPEECH 2007 will take place in the Flanders Congress and Concert Centre (FCCC) in the heart of Antwerp, near the Antwerp Zoo and Antwerp Central Station. Antwerp is the city of Rubens, a renowned world diamond center and one of the largest European harbors. It is a midsize town with a lively, well preserved, medieval center small enough to maintain a cozy atmosphere. Antwerp is easily reachable from the airport of Brussels and through a high speed train link also from the airports of Amsterdam and Paris.

INTERSPEECH 2007 will cover all aspects of speech science and technology. The conference will include plenary talks by internationally renowned experts, tutorials, exhibits, special sessions, and parallel oral and poster sessions about the following topics:

- Human speech production and perception
- Human speech communication
- Speech coding and speech enhancement
- Speech and audio signal processing
- Automatic speech and speaker characterization
- Speech synthesis
- Automatic speech recognition
- Speech technology applications
- Speech and multimodal resources
- Other relevant topics

A more detailed description of the topics can be found at the conference website.

**PAPER SUBMISSION**

The deadline for submission of a 4-page full paper is March 23, 2007. Paper submission will be through the conference website, using the submission guidelines provided. Previously-published papers should not be submitted. Notice of paper acceptance will be given on May 25, 2007.

**PROPOSALS FOR TUTORIALS AND SPECIAL SESSIONS**

We solicit proposals for half-day pre-conference tutorials to be held on Monday, August 27, 2007. One-page descriptions of proposed tutorials should be sent to tutorials@interspeech2007.org. We also welcome proposals for special sessions covering interdisciplinary topics and/or important new emerging areas of interest. Descriptions should be sent to special_sessions@interspeech2007.org.

**IMPORTANT DATES**

- Proposals for special sessions: November 1, 2006
- Proposals for tutorials: January 8, 2007
- Four-page paper deadline: March 23, 2007
- Notification of paper acceptance: May 25, 2007
- Early registration deadline: June 22, 2007
- Tutorial Day: August 27, 2007
- Main conference: August 28-31, 2007

For further information: http://www.interspeech2007.org

Email address: info@interspeech2007.org
The Tenth International Conference on Spoken Language Processing (INTERSPEECH-2008 ~ ICSLP) will be held in Brisbane, Australia.

INTERSPEECH-2008 ~ ICSLP follows on from INTERSPEECH-2006 ~ ICSLP in Pittsburg, USA, and INTERSPEECH-2007 ~ Eurospeech in Antwerp, Belgium.

INTERSPEECH-2008 will cover all the scientific and technological aspects of speech. We are expecting more than 1000 participants from all over the world and more than 700 papers presented in oral and poster sessions. Several satellite workshops and exhibitions will further enrich the conference content.

Brisbane, the capital of Queensland, is a modern, dynamic, sophisticated and cosmopolitan city with a friendly and relaxed lifestyle. Australia’s only sub-tropical capital city, Brisbane is located on the doorstep of the nation’s premier tourist region, and within an hour of some of the world’s most famous beaches, the legendary Australian Outback, and Barrier Reef Islands, making it the ideal destination for each access to pre and post conference touring.

www.interspeech2008.org

Local Organizers: Australasian Speech Science and Technology Association
Chairperson: Professor Denis Burnham, University of Western Sydney
General Information

Climate
Pittsburgh climate is moderately warm in September. The average high temperature is \(76 \degree F / 24 \degree C\) and the average low temperature is \(56 \degree F / 13 \degree C\). The record high temperature for September 19 is \(87 \degree F / 30 \degree C\) (in 1991) and the record low temperature is \(45 \degree F / 7 \degree C\) (in 2004). Pittsburgh does not have a dry season, so it may rain on any day and at any time. It is always prudent to carry a small umbrella.

Time Zone
Pittsburgh is the Eastern Time Zone and will be following Daylight Savings time in September (GMT – 4). It is in the same time zone as Boston, New York, Washington DC, Atlanta, Detroit, and Miami, one hour ahead of Chicago and Dallas, two hours ahead of Denver, and three hours ahead of San Francisco, Los Angeles, and Seattle.

Language
The official language is English. While a sizable minority also speaks Spanish with a variety of Central and South American dialects, Spanish is not frequently encountered in Pittsburgh.

Electrical Current
110 Volts AC, 60 Hz. Electrical outlets accept two-pronged plugs (without ground) or three-pronged plugs as shown in the picture below.

![Electrical plugs and outlet](image)

Currency Exchange
The only currency accepted is US Dollars. Foreign currency may be exchanged at some banks and exchange windows at the airport. Unfortunately, the banks charge a large premium on each transaction. It may be advisable to purchase dollars before your arrival. Credit cards are almost universally accepted (even at fast food establishments like McDonalds and Subway). There are Automatic Teller Machines (ATMs) available at most bank branches and at various establishments, including the conference hotel. These machines will dispense US Dollars.

Tipping
Tipping is expected in taxis, and at restaurants and bars. The usual rate is 15 percent, more for excellent service, or less for below-average service.

Telecommunications
Wireless internet connectivity is provided free of charge by the hotel in the guestrooms and in the public areas. Mobile phones in the USA use a variety of standards with GSM compatibility supported by some but not all providers.
Medical Care
Pittsburgh is a major medical center in the US with many of its hospitals ranked among the top national institutions. The emergency number is 911. It is advisable to check with your insurance carrier regarding any coverage limitations during your travel.

Smoking
Smoking is not permitted in the guestrooms and public spaces in the Westin Hotel. Many but not all restaurants in Pittsburgh offer a smoking area. The Allegheny County Council (which governs the county which includes Pittsburgh) is currently considering the imposition of a smoking ban at all restaurants. Bars continue to allow smoking.

Shopping
The Westin Hotel is located in downtown Pittsburgh in close proximity to many large department stores and specialty boutiques.

Sales Tax
The local sales tax is 7 percent. This tax applies to all purchases except for clothing and groceries. Purchases by mail or over the internet from vendors outside Pennsylvania that are shipped to Pittsburgh usually are not taxed.

Passport and visas
Please see the visa requirements page on the conference website
Downtown Pittsburgh
and the Interspeech 2006 Venues

Facilities
1. David L. Lawrence Convention Center
2. Heinz Field
3. PNC Park

Downtown Hotels
4. Courtyard by Marriott
5. Doubletree Hotel Pittsburgh City Center
6. Hilton Pittsburgh
7. Marriott Pittsburgh City Center
8. Omni William Penn Hotel
9. Renaissance Pittsburgh Hotel
10. Sheraton Hotel at Station Square
11. Springhill Suites by Marriott
12. Westin Convention Center

Interspeech 2006 Venues
- Heinz History Center
- Westin Hotel and Convention Center
- Dowe's on Ninth
- Heinz Hall

Map of Downtown Pittsburgh and Interspeech 2006 Venues
Arrival and Parking

The Pittsburgh International Airport

The Pittsburgh International Airport has been a favorite for travelers since its opening in 1992, and has consistently ranked among the top airports worldwide in national and international surveys. Extensive information about the airport and its facilities may be found at http://www.flypittsburgh.com.

Getting to the Westin Hotel and Convention Center

There are several easy ways of getting from the Pittsburgh International Airport to the Westin Hotel and Convention Center. Here are a few options, listed in order of increasing cost. All transportation services (except for rental cars) are accessed on the baggage level of the arrival terminal, on the “Commercial Curbs” side of the terminal building.

- The local public bus company, the Port Authority of Allegheny County, runs an express bus, the 28X Airport Flyer from the airport to within a few blocks of the Westin Hotel, the intersection of Sixth Avenue and Smithfield Street. Fare is $2.25, exact change only, collected on the bus. Ask the bus driver where to get off for the Westin Hotel. More information may be found at http://www.ridegold.com.

- Express Shuttle USA runs a shuttle from the airport to the Westin Hotel for the fare of $19 per person one way, $34 round trip. Further information about this service may be found at http://www.pghtrans.com/airportshuttle.html

- A taxi can be taken from the airport to the hotel for about $30-35 per carload, including the tip. During rush hours the taxis may be no faster than the buses, which run on reserved roadways for part of the trip, but taxis are less expensive than the Express Shuttle service for two or more people traveling together.

- All major car rental companies are represented at the airport. Renting a car is not recommended unless you wish to use it to tour the outlying attractions (such as Fallingwater), as a car is not really necessary for the meeting itself and parking can be expensive (see below).

Parking in downtown Pittsburgh

Pittsburgh is located within a short drive of a very large portion of the US population, so many attendees may choose to drive to the meeting. Parking at the Westin hotel is expensive, $18 per day during the weekdays. Some other parking lots in the area offer better prices. For example, the nearby lot on Penn Avenue between Seventh and Eight Streets is only $10 per day on most days. Further information on parking in downtown Pittsburgh (including locations and prices) may be found at http://www.downtownpittsburgh.com/Content/DowntownMap.aspx and http://www.alcoparking.com/downtown.cfm
Some Facts About Pittsburgh

Welcome to one of the USA's most delightful surprises. No longer the industrial steel town of old, Pittsburgh is now a true renaissance city: a city of modern cathedrals and old world, neighborly charms, filled with high-tech companies, friendly faces, fun and adventure.

Further information about Pittsburgh’s attractions may be found at the VisitPittsburgh Web site, 

Pittsburgh Basics

- Founded: 1758
- Incorporated: 1816
- City Population: 350,363
- Allegheny County Population: 1,336,449
- Rank: Thirteenth Largest City in the USA
- Elevation: 1,223 Feet
- Port: Pittsburgh is the USA's largest inland port

Sightseeing in Pittsburgh

Pittsburgh is home to great vistas and attractions. Below is partial list of museums and natural attractions in town, many of which are located in the Downtown Pittsburgh area.
(Source http://www.pittsburgh.worldweb.com/SightsAttractions/Museums/)

Three Rivers Heritage Trail (http://www.friendsoftheriverfront.org/new_pages/heritage.htm)
The Three Rivers Heritage Trail is a wonderful pedestrian trail and greenway system around Pittsburgh. Running for 37 miles along both sides of the Allegheny, Monongahela and Ohio Rivers, it is an almost complete public route for cyclists, walkers, runners, and in some places rollerbladers.

Fort Pitt Museum
With a series of displays and artifacts, Fort Pitt Museum teaches visitors about the French and Indian War, the American Revolution and early military tactics. Located at the Point State Park, the museum is housed in a re-created version of the Monongahela Bastion, a fortress built in the mid-18th century. 101 Commonwealth Place; 412 281-9284.

The Andy Warhol Museum
Located just across the Seventh Street Bridge (now the Andy Warhol Bridge), The museum opened in 1994 and features a large permanent collection of works by Andy Warhol and his contemporaries. Featured in the museum is over 4000 pieces of art, consisting of paintings, drawings, prints, photography, sculptures and film. 117 Sandusky Street; 412 237-8342.

Senator John Heinz Pittsburgh Regional History Center
The Center, site of the Interspeech 2006 banquet, first opened in 1996 and is located in the old Chautauqua Lake Ice Company building. In total, over 160,000 square feet of space is located in the museum. The museum features a children's Discovery Palace, Historical Society Library and Archives. 1212 Smallman Street; 412 454-6000.
The Western Pennsylvania Sports Museum
Located in the Heinz History Center, the Western Pennsylvania Sports Museum takes up two floors, taking visitors in depth through Pittsburgh's sports heritage. Inside are hundreds of artifacts, displays of stories and legends, hands-on exhibits and multi-media areas, as well as a Wall of Fame. 1212 Smallman Street; 412 454-6000.

Children's Museum of Pittsburgh
This museum offers interactive exhibits which are designed for children of all ages. Exhibits include the Studio, Attic, Backyard, Garage/Workshop, Nursery, Theater and Welcome to Mister Rogers’ Neighborhood. Interactive artwork by Camille Utterback, Tim Kaulen and more, classes, puppet shows, storytelling and traveling exhibits make this a destination for children of all ages. 10 Children's Way, (North Side); 412 322-5058.

Mattress Factory Museum
The museum features contemporary art. Established in 1977, the museum is located on the famous Mexican War Streets. Featured exhibitions include work from Patrice Carre, Robin Minard and Takehisa Kosugi. 500 Sampsonia Way (North Side); 412 231-3169.

Carnegie Museum of Art
Art collections include architecture, decorative arts, film and video, paintings, sculptures and more. The main attraction at the museum is contemporary art. A replicate of the Parthenon is located in the Hall of Sculpture. 4400 Forbes Avenue (Oakland area); 412 622-3131.

Soldiers & Sailors National Military Museum & Memorial
Through a myriad of artifacts, equipment, uniforms, arts and personal belongings, the Soldiers & Sailors National Military Museum & Memorial seeks to honor soldiers from the Civil War to modern day battles. 4141 Fifth Avenue (Oakland area); 412 621-4253.

Carnegie Museum of Natural History
Known for containing information on Dinosaurs. Located in the museum's Dinosaur Hall are ten full skeletons. The Bendedum Hall features geological information about the earth, such as rocks and fossils. Hall of North American Wildlife includes information on animal wildlife of North America. 4400 Forbes Avenue (Oakland area); 412 622-3131.

The Frick Art and Historical Center
Located on six-acres, the center features the home of Henry Clay Frick. The home was restored. Also reconstructed was the Alden and Harlow Greenhouse, which dates back to 1897. The visitor center was formerly the bowling alley in the home. Both the Car and Carriage Museum and Frick Art Museum are located in the center. 7227 Reynolds Street (in Pittsburgh’s East End); 412 371-0600.

Frick Car and Carriage Museum
Located next door to the Frick Art and Historical Center, the Car and Carriage Museum contains more than 20 vintage machines, taking visitors through the history of automobiles. The collection includes Henry Clay Frick's 1914 Rolls Royce Silver Ghost, with the oldest exhibit featuring an 1881 Brougham car. 7227 Reynolds Street (in Pittsburgh’s East End); 412 371-0600.
Pittsburgh Firsts

Pittsburgh was the first city in the world to do many interesting things! Here are a few of the best known:

- **First Internet Emoticon** – 1980. The Smiley :-) was the first Internet emoticon, created by Carnegie Mellon University computer scientist (and LTI faculty member) Scott Fahlman.
- **First Mr. Yuk Sticker** – 1971. Mr. Yuk was created at the Poison Center at Children's Hospital of Pittsburgh after research indicated that the skull and crossbones previously used to identify poisons had little meaning to children who equate the symbol with exciting things like pirates and adventure.
- **First Big Mac** – 1967. Created by Jim Delligatti at his Uniontown McDonald's, the Big Mac debuted and was test marketed in three other Pittsburgh-area McDonald's restaurants in 1967. By 1968 it was a mainstay of McDonald’s menus throughout the country.
- **First Pull-Tab on Cans** – 1962. The pull-tab was developed by Pittsburgh-based Alcoa Aluminum and was first used by Iron City Brewery in 1962. For many years, pull-tabs were only used in this area.
- **First Retractable Dome** – 1961. Pittsburgh’s Civic Arena, home of the Pittsburgh Penguins hockey team (at least for now) is the world’s first auditorium with a retractable roof.
- **First US Public Television Station** – April 1, 1954. WQED, was the first community-sponsored educational television station in America.
- **First Polio Vaccine** – March 26, 1953. The polio vaccine was developed by Dr. Jonas E. Salk, then a 38-year-old University of Pittsburgh researcher and professor.
- **First All-Aluminum Building** – August 1953. The first aluminum-faced skyscraper was the Alcoa Building, a 30-story, 410-foot structure with thin stamped aluminum panels forming the exterior walls. Alcoa is currently headquartered in a striking new building (also made of aluminum) close to Interspeech 2006 on the North Shore of the Allegheny River.
- **First Zippo Lighter** – 1932. George G. Blaisdell invented the Zippo lighter in 1932 in Bradford, Pennsylvania. The name Zippo was chosen by Blaisdell because he liked the sound of the word “zipper,” which was patented around the same time in nearby Meadville.
- **First Bingo Game** – early 1920s. Hugh J. Ward first came up with the concept of bingo in Pittsburgh and began running the game at carnivals in the early 1920s, taking it nationwide in 1924. He secured a copyright on the game and wrote a book of Bingo rules in 1933.
- **First US Commercial Radio Station** – 1920. Dr. Frank Conrad, Assistant Chief Engineer of Westinghouse Electric, first constructed a transmitter and installed it in a garage near his home in Wilkinsburg in 1916. The station was licensed as 8XK. At 6 pm on November 20, 8XK became KDKA Radio and began broadcasting at 100 watts from a make-shift shack atop one of the Westinghouse manufacturing buildings in East Pittsburgh.
- **Daylight Savings Time** – March 18, 1919. A Pittsburgh city councilman during the first World War, Robert Garland devised the nation’s first daylight savings plan, instituted in 1918.
- **The First Gas Station** – December, 1913. The first automobile service station, built by Gulf Refining Company, opened in Pittsburgh at Baum Boulevard and St. Clair Street in East Liberty.
- **First Motion Picture Theatre** – 1905. The first theater in the world devoted to the exhibition of motion pictures was the “Nickelodeon,” opened by Harry Davis on Smithfield Street in Pittsburgh.
- **The First World Series** – 1903. The Boston Pilgrims defeated the Pittsburgh Pirates five games to three in baseball’s first modern World Series in 1903.
- **Long-Distance Electricity** – 1885. Westinghouse Electric developed alternating current, allowing long-distance transmission of electricity for the first time.
Maps of the Conference Rooms at the Westin

Second floor:

Third floor:
General Interspeech 2006 Information

REGISTRATION

The registration desk is on the second floor in the main foyer. It is open throughout the conference from 0800 through 1900. The desk is also open all day and evening on Sunday September 17, and it is highly recommended that you on Sunday if at all possible to avoid long lines at the start of the conference.

CONQUEST: A SPOKEN INTERACTIVE DIALOG SYSTEM FOR CONFERENCE INFORMATION

ConQuest provides schedule information for the Interspeech 2006 conference and allows callers to learn details of the technical program.

Conquest is a telephone-based spoken dialog system developed by members of the CMU Speech Group (www.speech.cs.cmu.edu). The system is based on the RavenClaw-Olympus dialog system architecture (www.ravenclaw-olympus.org). It uses the SPHINX speech recognition system together with a RavenClaw-based dialog manager, template-based natural language generation, and the Cepstral Swift synthesizer.

Following the conference, the entire system, as well as the data collected during the conference, will be freely available to the research community. For more information about the system and the underlying components please visit the ConQuest web-site at http://www.conquest-dialog.org/

During the conference, ConQuest can be reached at the following phone number: 1-412-268-8817. To facilitate access, a telephone is available at the registration desk.

INTERNET CAFÉ

In addition to wireless internet being available in the guest rooms and all public areas of the conference, an internet café is available in the Crawford Room, behind the elevators on the third floor, across from the Allegheny Ballroom. The café provides computers and ethernet connectivity for those requiring access.

Speakers are required to upload their slides using the facilities in the Internet Café, where conference staff will provide assistance and verify that your slides display properly from the podium machines. Presenters will not be permitted to use their own machines for presentations unless there is a compelling reason to do so which is approved by the conference staff in the internet café in advance of the session.

MESSAGE AND ANNOUNCEMENT BOARD

A message and announcement board is available in the second floor foyer opposite the registration desk.
LIABILITY AND INSURANCE

While Interspeech 2006 and ISCA will make every effort to ensure that your participation in Interspeech 2006 is both enjoyable and safe, they cannot assume liability for personal injuries sustained by conference attendees, or for loss or damage to property belonging to conference participants, either during or as a result of attendance at the conference. It is, therefore, advised that participants arrange for their own personal health, accident, and travel insurance.

OFFICIAL PROGRAM

Opening Ceremony

The Opening Ceremony for Interspeech 2006 – ICSLP will take place Monday September 18 at 8:30 in the main Allegheny Ballroom on the third floor of the Westin Hotel.

Plenary Sessions

Plenary sessions will take place in the main Allegheny Ballroom from 9:00 to 10:00, Monday through Thursday September 18-21.

ISCA General Assembly

A meeting of the ISCA General Assembly will take place Tuesday September 19 from 18:30 to 19:30 in Allegheny Ballroom II.

Closing Ceremony

The closing ceremony will take place on Thursday September 21 at 16:30, also in the main Allegheny Ballroom.

SOCIAL PROGRAM

Technology Demonstrations from Carnegie Mellon, the University of Pittsburgh, and local corporate sponsors

Researchers in academia and industry in the Pittsburgh area will demonstrate their latest speech and language technology in the Westmoreland Room Sunday evening September 17 from 17:30 to 19:30. Refreshments will be provided under the sponsorship of the CMU Language Technologies Institute.

Student Panel Discussion

A special student panel discussion entitled “How to get your dream job? What are they looking for?” has been organized by the ISCA Student Advisory Committee (ISCA-SAC). It will take place in the Cambria Room Sunday evening September 17 from 18:30 to 20:00.
Welcome Reception

The welcome reception will be held in Pittsburgh Symphony’s Heinz Hall on Monday evening starting at 19:00. Heinz Hall is a short walk from the Westin Hotel, on the corner of Penn Avenue and Sixth Street, as shown in the map of downtown Pittsburgh earlier in this book. Heinz Hall for the Performing Arts, dedicated in 1971, and renovated in 1995, is the cornerstone of the Cultural District of Pittsburgh. Its 2,661-seat hall is home to the Pittsburgh Symphony Orchestra. We are grateful to Google, Inc. and Nuance Communications for sponsoring this event.

Student Reception

Be prepared for a night of jazz, as the Interspeech 2006 student reception will be held at Dowe's on Ninth on Tuesday evening starting at 20:00, after the ISCA General Assembly. Dowe's on Ninth is an eclectic, international jazz bar in downtown Pittsburgh. This year’s student reception will provide a buffet dinner, and the evening will include music by local jazz trumpeter extraordinaire Sean Jones and his band Mission Statement. If you enjoy jazz, this is an event not to be missed. Dowe's on Ninth is a short walk from the Westin Hotel, at 121 Ninth Street, as shown in the map of downtown Pittsburgh earlier in this book. We are grateful to IBM and Microsoft for sponsoring this event. More information on Dowe’s on Ninth and Sean Jones may be found at http://www.doweson9th.com and http://www.seanjonesjazz.com.

Banquet

The Interspeech 2006 banquet will be held in the Senator John Heinz Pittsburgh Regional History Center, starting at 19:00 on Wednesday September 20. The Regional History Center is also a short walk from the Westin Hotel, on Smallman Street between Twelfth Street and Thirteenth Streets, as shown in the map of downtown Pittsburgh earlier in this book. The Heinz History Center is a museum that relates and interprets 250 years of history in the Western Pennsylvania area. See the exhibits depicting the early struggles between the French and English, Pittsburgh’s role as the Steel Center of the World, and of course those ubiquitous 57 Varieties for which Heinz is so well known.
### Interspeech 2006 Schedule

#### Session Code (Key)

<table>
<thead>
<tr>
<th>Day</th>
<th>Time Slot</th>
<th>Room Code</th>
<th>Session Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sun - Sunday</td>
<td>1 - 10:30 – 12:30</td>
<td>A1 - Allegheny Ballroom, left</td>
<td>O - Oral</td>
</tr>
<tr>
<td>Mon - Monday</td>
<td>2 - 14:00 – 16:00</td>
<td>A2 - Allegheny Ballroom, center</td>
<td></td>
</tr>
<tr>
<td>Tue - Tuesday</td>
<td>3 - 16:30 – 18:30</td>
<td>A3 - Allegheny Ballroom, right</td>
<td>P - Poster</td>
</tr>
<tr>
<td>Wed - Wednesday</td>
<td></td>
<td>We - Westmoreland Room</td>
<td></td>
</tr>
<tr>
<td>Thu - Thursday</td>
<td></td>
<td>Bu - Butler Room</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ca - Cambria Room</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fo - Allegheny Ballroom Foyer</td>
<td>S - Special</td>
</tr>
</tbody>
</table>

The Allegheny Ballrooms and the Allegheny Foyer are on the third floor of the Westin Hotel. The Butler, Cambria, and Westmoreland Rooms are on the second floor of the hotel.

#### Tutorial Day: Sunday, September 17

<table>
<thead>
<tr>
<th>Time</th>
<th>Allegheny I</th>
<th>Allegheny II</th>
<th>Allegheny III</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00</td>
<td>Registration Opens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:00</td>
<td>T1 Microphone Array Processing and Source Separation for Speech Enhancement and Recognition Bhiksha Raj, Paris Smaragdis (MERL) and Michael L. Seltzer (Microsoft)</td>
<td>T2 Speech Under Stress John Hansen (University of Texas, Dallas)</td>
<td>T3 Speech and Language Processing Over the WWW Mazin Gilbert (AT&amp;T)</td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13:30</td>
<td>T4 A Survey Of Robust Speech Recognition Techniques Jasha Droppo (Microsoft)</td>
<td>T5 Nonspectral Features for Speech Processing B. Yegnanarayana (IIIT Hyderabad)</td>
<td>T6 Music Information Retrieval Thomas Kemp and Jana Eggink (Sony Germany)</td>
</tr>
<tr>
<td>17:00</td>
<td>Registration Continues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17:30</td>
<td>Demos from Carnegie Mellon, the University of Pittsburgh, and Pittsburgh-based companies; Westmoreland and Somerset Rooms</td>
<td></td>
<td>Student Panel Discussion “How to get your dream job? What are they looking for?”; Cambria Room</td>
</tr>
<tr>
<td>18:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Monday, September 18

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>07:30</td>
<td>Registration Opens</td>
</tr>
<tr>
<td>08:30</td>
<td><strong>Opening Ceremony, Allegheny Ballrooms I, II, III</strong></td>
</tr>
<tr>
<td>09:00</td>
<td>Plenary Speaker</td>
</tr>
<tr>
<td></td>
<td>John J. Ohala, ISCA Medalist, Allegheny Ballrooms I, II, III</td>
</tr>
<tr>
<td>10:00</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:30</td>
<td>Mon1A1O Language Modeling for Spoken Dialog Systems</td>
</tr>
<tr>
<td></td>
<td>Mon1A2O Feature Enhancement for Robust ASR</td>
</tr>
<tr>
<td></td>
<td>Mon1A3O Dialog and Discourse</td>
</tr>
<tr>
<td></td>
<td>Mon1WeS The Speech Separation Challenge</td>
</tr>
<tr>
<td></td>
<td>Mon1BuP Multilingual and Multi-accent Processing</td>
</tr>
<tr>
<td></td>
<td>Mon1CaP Corpora, Annotation, and Assessment Metrics I</td>
</tr>
<tr>
<td></td>
<td>Mon1FoP Speech Coding</td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch</td>
</tr>
<tr>
<td>14:00</td>
<td>Mon2A1O Speech Enhancement I</td>
</tr>
<tr>
<td></td>
<td>Mon2A2O ASR Other I</td>
</tr>
<tr>
<td></td>
<td>Mon2A3O Modeling Prosodic Features</td>
</tr>
<tr>
<td></td>
<td>Mon2WeO Spoken Information Retrieval</td>
</tr>
<tr>
<td></td>
<td>Mon2BuP Front-end Methods for ASR</td>
</tr>
<tr>
<td></td>
<td>Mon2CaP Language and Dialect Recognition</td>
</tr>
<tr>
<td></td>
<td>Mon2FoP Spoken Dialog Systems I</td>
</tr>
<tr>
<td>16:00</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>16:30</td>
<td>Mon3A1O Speaker Characterization and Recognition I</td>
</tr>
<tr>
<td></td>
<td>Mon3A2O System Combination</td>
</tr>
<tr>
<td></td>
<td>Mon3A3O Interpreting Prosodic Variation</td>
</tr>
<tr>
<td></td>
<td>Mon3WeS Articulatory Modeling</td>
</tr>
<tr>
<td></td>
<td>Mon3BuP Acoustic Modeling I – Training and Topologies</td>
</tr>
<tr>
<td></td>
<td>Mon3CaP Acoustic Signal Segmentation and Classification</td>
</tr>
<tr>
<td></td>
<td>Mon3FoP Linguistics, Phonology, and Phonetics I</td>
</tr>
<tr>
<td>18:30</td>
<td>Welcome Reception, Heinz Hall</td>
</tr>
<tr>
<td>19:00</td>
<td>Welcome Reception, Heinz Hall</td>
</tr>
<tr>
<td>20:00</td>
<td>Welcome Reception, Heinz Hall</td>
</tr>
</tbody>
</table>
## Tuesday, September 19

<table>
<thead>
<tr>
<th>Time</th>
<th>Event</th>
<th>Allegheny I</th>
<th>Allegheny II</th>
<th>Allegheny III</th>
<th>Westmoreland</th>
<th>Butler</th>
<th>Cambria</th>
<th>Foyer</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00</td>
<td>Registration Opens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:00</td>
<td>Plenary Speaker</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Michael Phillips, Allegheny Ballrooms I, II, III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>Tue1A1O Speech Translation</td>
<td>Tue1A2O Acoustic Modeling II – Adaptation</td>
<td>Tue1A3O Emotional Speech and Speaker State</td>
<td>Tue1WeS Speech and Language in Education</td>
<td>Tue1BuP Speech Perception I</td>
<td>Tue1CaP Speaker Characterization and Recognition II</td>
<td>Tue1FoP Speech Production, Physiology, and Pathology I</td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Tue2A1O Formant Estimation</td>
<td>Tue2A2O Language Processing Beyond and Below the Word-Level</td>
<td>Tue2A3O Spoken Dialog Systems II</td>
<td>Tue2WeO Corpora, Annotation, and Assessment Metrics II</td>
<td>Tue2BuP Robustness and Adaptation for ASR</td>
<td>Tue2CaP Multimodal, Translation and Information Retrieval</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>Tue3A1O Advances in Acoustic Segmentation</td>
<td>Tue3A2O Acoustic Modeling III – LVCSR</td>
<td>Tue3A3O Linguistics, Phonology, and Phonetics II</td>
<td>Tue3WeO Speech and Visual Processing</td>
<td>Tue3BuP TTS I</td>
<td>Tue3CaP Special Populations – Learners, Aged, Challenged</td>
<td>Tue3FoP Speech Enhancement II</td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td>ISCA General Assembly, Allegheny Ballroom II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20:00</td>
<td>Student Reception, Dowe’s on Ninth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Event</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>----------------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>08:00</td>
<td>Registration Opens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:00</td>
<td>Plenary Speaker Elissa Newport, Allegheny Ballrooms I, II, III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>Wed1A1O Speaker Characterization and Recognition III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed1A2O Robust ASR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed1A3O Speech Perception II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed1WeS Speech Summarization</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed1BuP Acoustic Modeling IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed1CaP Large Vocabulary Speech Recognition</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed1FoP Speech/Noise/Music Segmentation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Wed2A1O Pitch Estimation</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed2A2O Acoustic Modeling V – Novel Approaches</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed2A3O Corpus-Based Synthesis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed2WeS Spoken Dialog Technology R&amp;D</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed2BuP Modeling Speaker Emotional State</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed2CaP Language Modeling and ASR Applications</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed2FoP Spoken Language Understanding</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>Wed3A1O Segmentation and VAD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed3A2O Technologies for Specific Populations: Learners and Challenged</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed3WeS The Prosody of Turn-Taking and Dialog Acts</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed3BuP TTS II</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed3CaP Speaker Characterization and Recognition IV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wed3FoP Multichannel Speech Enhancement / Speech Perception</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18:30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19:00</td>
<td>Banquet, Heinz History Center</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Thursday, September 21

<table>
<thead>
<tr>
<th>Time</th>
<th>Allegheny I</th>
<th>Allegheny II</th>
<th>Allegheny III</th>
<th>Westmoreland</th>
<th>Butler</th>
<th>Cambria</th>
<th>Foyer</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00</td>
<td>Registration Opens</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>09:00</td>
<td>Plenary Speaker</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Raj Reddy, Allegheny Ballrooms I, II, III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10:30</td>
<td>Thu1A1O Diarization in ASR</td>
<td>Thu1A2O Language Model Adaptation, Refinement, and Evaluation</td>
<td>Thu1A3O Speech Production, Physiology and Pathology II</td>
<td>Thu1WeS From Ideas to Companies</td>
<td>Thu1BuP Voice Morphing</td>
<td>Thu1CaP ASR Other II</td>
<td>Thu1FoP Prosody</td>
</tr>
<tr>
<td>12:30</td>
<td>Lunch</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14:00</td>
<td>Thu2A1O Discriminative Training</td>
<td>Thu2A3O Speech Synthesis</td>
<td>Thu2WeO Multimodal Processing</td>
<td>Thu2BuP Speech Analysis</td>
<td>Thu2CaP Advances in Noisy ASR</td>
<td>Thu2FoP Source Separation and Localization</td>
<td></td>
</tr>
<tr>
<td>16:00</td>
<td>Coffee Break</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16:30</td>
<td>Closing Ceremony, Allegheny Ballrooms I, II, III</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Tutorials
Sunday, September 17, 2006

Microphone Array Processing and Source Separation for Speech Enhancement and Recognition
Bhiksha Raj (MERL), Paris Smaragdis (MERL), and Michael L. Seltzer (Microsoft)

Tutorial AM1 (9:00 – 12:30, Allegheny I)

Overview
This tutorial will address the topics of microphone array processing for signal enhancement and multi- and single-channel approaches to signal separation.

We will present a range of topics including:

1. Basic principles of microphone array processing, including issues of array geometry, spatial aliasing and the distinction between nearfield and farfield array responses, and source localization techniques
2. Beamforming algorithms for enhancing noisy signals or separating concurrent signals, including classical methods for fixed and adaptive beamforming, maximum-likelihood methods for beamforming
3. Independent component analysis (ICA) in time and frequency domains, and multiple-microphone methods for separating multiple concurrent signals when the number of sound sources exceeds the number of microphones
4. Single-channel source separation techniques including those based on estimation of spectrographic masks, non-negative matrix factorization and related techniques.
5. Automatic speech recognition in noisy and reverberant environments using microphone arrays, including the effectiveness of various array processing algorithms, feature compensation algorithms, and back-end model adaptation schemes.

Presenters

Bhiksha Raj joined MERL as a Staff Scientist. He completed his Ph.D. from Carnegie Mellon University (CMU) in May 2000. Dr. Raj works mainly on algorithmic aspects of speech recognition, with special emphasis on improving the robustness of speech recognition systems to environmental noise. His latest work has been on sound source separation and the use of novel secondary sensors for speech processing.

Paris Smaragdis is a member of the research staff in Mitsubishi Electric Research Laboratories in Cambridge MA, USA. Prior to that position he was at the Massachusetts Institute of Technology where he completed his graduate and postdoctoral training. His interests are computational audition, scene analysis and the intersection of machine learning with signal processing.

Michael L. Seltzer received the Sc.B. with honors from Brown University in 1996, and M.S. and Ph.D. degrees from Carnegie Mellon University in 2000 and 2003, respectively, all in electrical engineering. From 1998 to 2003, he was a member of the Robust Speech Recognition group at Carnegie Mellon University. In 2003, Dr. Seltzer joined the Speech Technology Group at Microsoft Research, Redmond, WA. His current research interests include speech enhancement, speech recognition in adverse acoustical environments, acoustic modeling, microphone array processing, and machine learning for speech and audio applications.
Speech Under Stress

John Hansen (University of Texas, Dallas)

Tutorial AM2 (9:00-12:30, Allegheny II)

Overview

The field of speech processing involves modeling of human production in order to formulate algorithms for effective systems in application areas such as speech recognition, speaker recognition, speech synthesis, speech coding, and speech enhancement. The variability brought on by stress causes significant variability which significantly impacts the performance of these algorithms. Stress in this context will refer to cognitive stress, physical stress, emotional stress (fear, anger, anxiety), and stress due to the presence of noise (known as Lombard effect).

This tutorial will provide a comprehensive overview to the field of speech under stress, and will include many audio/video examples to illustrate the range of speaker variability. This tutorial will include the following topics:

- **Historical Overview of Speech Under Stress**: we consider studies on analysis of speakers under stress (audio examples from Hindenburg disaster, aircraft pilots, 911 emergency calls, police communications, etc.).
- **Available Corpora on Speech Under Stress**
- **Analysis of Speech Under Stress**: we provide a comprehensive summary of the features and analysis results which show how speech production changes under stress.
- **Speech Recognition Under Stress**: we consider methods to improve Automatic Speech Recognition in stressful/noisy conditions on small vocabulary and trends towards large vocabulary systems and how they can be employed in mobile environments.
- **Detection of Speech Under Stress**: an overview of methods to detect stress in speech. These include linear features (pitch, glottal structure, spectral structure, duration and intensity, etc.), as well as nonlinear methods (based on the Teager Energy Operator).
- **Synthesis & Perception of Speech Under Stress**: summarize studies which consider human perception of speech under stress, and briefly consider methods for imparting stress on neutral speech for either speaker dependent synthesis or cross-speaker synthesis.

Presenter

**John H.L. Hansen** his Ph.D. and M.S. degrees in Electrical Engineering from Georgia Institute of Technology, Atlanta, Georgia, in 1988 and 1983, and B.S.E.E. degree from Rutgers University, New Brunswick, N.J. in 1982.

Hansen is presently the chair of the Electrical Engineering and Telecommunications Department at the University of Texas at Dallas, and holds a joint appointment as professor for Speech and hearing in the School of Behavioral and Brain Sciences at UTD. He is the Director and Co-founder of the Center for Robust Speech Systems. He was a co-founder of the Center for Speech and Language Research (CSLR) at the University of Colorado, Boulder, where he served as Associate Director from 1999-2003. He had previously been a faculty member at Duke University.

Hansen’s research interests include the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech pathology and voice assessment, speech enhancement and feature estimation in noise, robust speech recognition, and speech feature enhancement in hands-free environments for human-computer interaction.
Speech and Language Processing Over the WWW

Mazin Gilbert (AT&T)

Tutorial AM3 (9:00 – 12:30, Allegheny III)

Overview

Over the past decade, the World-Wide Web (WWW) has been evolving into a central communication hub for consumers and businesses to efficiently access and deliver multimedia information whether containing text, speech, graphics, audio or video. In this booming era of the internet, communication is evolving at an extraordinary pace, changing from voice over traditional landline phones to multimedia data across multiple multimodal devices, services and networks. Technological breakthroughs, which are making it possible for people to communicate more seamlessly and rapidly acquire information more efficiently, are revolutionizing the fields of speech and language processing and providing new research challenges and lucrative business opportunities that did not exist before. Combining speech and language technology with the wealth of information and knowledge that reside on the web may ultimately achieve the Artificial Intelligence (AI) dream - the dream of being able to converse naturally with a multilingual virtual agent that can understand and interact in spoken natural language. Although the lack of sufficiently advanced speech and language understanding technologies for processing highly unstructured web data would continue to pose technical challenges, the opportunities are endless and traditional and new businesses are booming in areas of communication, entertainment, and marketing, taking advantage of the internet revolution to expand and reinvent their business in ways that is never done before.

This tutorial is suitable for researchers and students who would like to acquire a better perspective on how the speech and language technologies are being transformed by the web, and how these technologies are impacting the way people and businesses communicate. The focus of the tutorial will be in the following areas:

1. Advances in Speech and Language Technology: Theoretical and practical perspective of the speech and language technologies including speech synthesis and recognition, speaker recognition, language understanding, question/answering, natural language processing, semantic classification, and machine translation.

2. The Web Transformation: The web transformation in areas of communication and information mining including document analysis and understanding, webpage analysis, information search and retrieval, trend analysis and tracking, and Services Over IP.

3. Emergence of a New Era: Technical challenges and lucrative business opportunities that are being created by the web revolution. The emergence of new applications and services for multimodal and multimedia information mining, web-based interactive virtual agents, and surveillance and intelligence gathering from blogs and web multimedia contents.

Presenter

Mazin Gilbert received the Ph.D. degree in Electrical Engineering and Electronics from the University of Liverpool, England. He is the Director of the Spoken Language Understanding and Dialog Division at AT&T Labs Research. His research lies in the areas of speech and language processing, machine learning, speech data mining, and spoken and multimodal dialog systems. Mazin has been instrumental in driving the speech and language research and software tools that led to the AT&T VoiceTone® offer in 2003. He has over eighty technical publications and is the author of the book “Artificial Neural Networks for Speech Analysis/Synthesis” (London: Chapman and Hall, 1994). He holds 11 US patents and is a recipient of several national and international awards. He is a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE) and the Chair of the IEEE Speech and Language Technical Committee.
A Survey of Robust Speech Recognition Techniques

Jasha Droppo (Microsoft)

Tutorial PM1 (13:30 – 17:00, Allegheny I)

Overview

The major obstacle to deploying a successful mobile speech recognition system, such as in a kiosk, automobile, or mobile phone, is environmental noise. Unless properly compensated for, noise can render any speech recognition device unusable.

Despite years of research, noise robustness remains an active area of interest. This tutorial is designed to provide attendees with an understanding of prevailing techniques, how they work, and how they relate to each other.

The course consists of three complementary sections: an analysis of how noise distorts speech recognition features, a summary of basic techniques that should be included in any system, and a cohesive taxonomy of current work in the field.

To understand how current techniques address the problem of additive noise, it is crucial to first know how additive noise distorts the speech recognition features. The first part of the tutorial is dedicated to analyzing the types of noises characteristic of deployed systems, how they affect the speech spectrogram and cepstral features, and how they affect overall system accuracy.

There are several well-proven techniques that address general non-informative signal variability, but are seldom included in published baselines. The second part of the tutorial presents these algorithms, which are in general easy to implement, use little resources, and provide considerable improvements in recognition accuracy. No production system or research baseline should be without them.

The final, and longest, part of the tutorial covers recent work in the field of robust speech recognition. The field is currently fractured across several lines, including between auditory scene analysis and data-driven probabilistic approaches, between feature-based and model-based techniques, between feature enhancement and feature normalization, and between maximum likelihood and discriminative training criteria. This tutorial covers each of these areas, describes the related techniques, and highlights the similarities of diverse approaches.

Presenter

Jasha Droppo received his B.S.E.E. degree (1994) from Gonzaga University in Spokane, Washington, and his M.S. and Ph.D. degrees (in 1996 and 2000, respectively) from the University of Washington in the field of Electrical Engineering. Dr. Droppo joined the Speech Technology Group at Microsoft Research in the spring of 2000. His work there has concentrated on improving the robustness of speech recognition systems to environmental noise. He has advanced the state of the art with his work on SPLICE, model based feature enhancement, uncertainty decoding, multi-sensory speech enhancement, and discriminatively trained speech recognition features. He is the author on more than 50 journal articles, conference papers, and patents in these above areas. His current interests include robust speech recognition, speech and acoustic digital signal processing, discriminative training algorithms, and non-stationary signal modeling.
Nonspectral Features for Speech Processing

B. Yegnanarayana (IIIT Hyderabad, India)

Tutorial PM2 (13:30 – 17:00, Allegheny II)

Overview

For most speech applications, the speech information is captured through short-time (10-30ms) spectral features. But the speech signal contains significant information in various other components, such as the phase of the short-time Fourier transform, the subsegmental (1-3ms) features, and the suprasegmental features. Several linear and nonlinear methods will be discussed to extract the information present in these components. The importance of these nonspectral features will be demonstrated in applications such as speaker verification, speech enhancement, speech synthesis, speech recognition, and language identification. The following is the description of contents and course material:

1. Speech information in sub-segmental, segmental and suprasegmental features
2. Some facts in short-time spectrum analysis of speech, and the need for exploring the nonspectral features in speech
3. Significance of the phase of the short-time FT of speech, and methods to process the phase using group-delay functions
4. Analysis of excitation information in speech: Epoch extraction, Hilbert envelope and phase of the LP residual
5. Analysis of voiced speech: Periodic and aperiodic decomposition, and formant extraction over transition regions
6. Prosody manipulation using epochs and LP residual
7. Speaker-specific source features from LP residual
8. Speech enhancement using the excitation source information: Single channel, multichannel and multispeaker
9. Duration modeling using artificial neural networks: language and speaker identification
10. Intonation modeling using artificial neural networks: language and speaker identification

Presenter

Bayya Yegnanarayana is a professor and Microsoft Chair at the International Institute of Information Technology (IIIT) in Hyderabad, India since July 2006. Prior to joining IIIT Hyderabad, he was a professor at IIT Madras, India (1980-2006). He was a visiting associate professor of Computer Science at Carnegie-Mellon University in USA from 1977-1980. He was a member of faculty at the Indian Institute of Science (IISc), Bangalore, from 1966 to 1978. He did BE, ME and PhD from IISc Bangalore, in 1964, 1966, and 1974, respectively. His research interests are in speech, image processing and neural networks. He has published several papers in these areas in IEEE and other international journals. He is also the author of the book “Artificial Neural Networks”, published by Prentice-Hall of India in 1999. He was an Associate Editor of the IEEE Tran on Speech and Audio Processing from 2003-2006. He is a Fellow of the Indian National Academy of Engineering, a Fellow of the Indian National Science Academy, and a Fellow of the Indian Academy of Sciences.
Music Information Retrieval
Thomas Kemp and Jana Eggink (Sony Germany)

Tutorial PM3 (13:30 – 17:00, Allegheny III)

Overview

In this tutorial, we will give a comprehensive overview over the research areas in Music Information Retrieval (MIR), explain the fundamental algorithms that have been successfully employed, and point you to the relevant authors and publications.

Topics that will be covered by the tutorial range from early Query by Humming systems over music similarity computation, music mood detection and genre recognition to music identification (fingerprinting). Different aspects of music transcription will be covered, including beat tracking, instrument identification and melody extraction. The technology behind automatic generation of individualized play lists and of personalized recommendations for music will be explained.

The basic signal processing techniques commonly used for these tasks will be introduced, as well as relevant developments in the areas of pattern recognition and machine learning.

One additional focus point of the tutorial will be music identification. Systems such as mobile phone music identification services will be explained in detail, and compared to well-known services such as CD identification.

Presenters

**Thomas Kemp** received his M.S. in Physics in 1993 with a work on thermoelectric infrared detectors, and a Ph.D. in computer science in 1999 with a dissertation about a TV news indexing and retrieval system based on speech recognition. He joined Sony Germany research in 2000, and is currently responsible for the Music Metadata Extraction activities in the Sony Research Center in Stuttgart. Dr. Kemp is author of more than 20 journal and conference papers and holds 8 international patents. He has been working for the last 10 years in the areas of signal processing, speech recognition and music information retrieval.

**Jana Eggink** received her M.A. in Systematic Musicology in 2001 from the University of Hamburg, Germany, with a thesis focusing on applications of computational auditory scene analysis (CASA) to music analysis. In 2006 she received her Ph.D. in Computer Science from the University of Sheffield, UK, working on automatic musical instrument identification and melody extraction from polyphonic music. She joined the Sony Germany Research Center in Stuttgart in 2005, where she is working in the fields of music information retrieval and general audio classification.
Plenary Talk 1

What a Study of Sound Can Tell Us About Human Speech Perception
(and Suggest About Machine Recognition)
John J. Ohala, University of California, Berkeley
(ISCA Medalist)

Allegheny Ballrooms, Monday, September 18, 9:00 to 10:00 am

Overview

Sound change – variation in pronunciation over time, community, or the context in which speech sounds occur — has been known since ancient times as witnessed by writers noticing “accents” of people in various regions or noticing that ancient poetry meant to rhyme or have some other fixed prosodic structure has lost those aspects over the centuries. Documentation of sound change advanced rapidly in the 18th and 19th centuries, especially with the development of the comparative method, a quasi-rigorous way of establishing cognate sets of words in different languages whose different pronunciations could be explained by the systematic change of words’ constituent sounds derived from some to-be-reconstructed parent language (e.g., English father, French pere, Spanish padre, Sanskrit pitar). There is today a vast archive of documented sound changes in scores of language families involving hundreds of languages. This data can be mined for insight into the causes or mechanisms of sound change, especially when combined with studies of speech production and speech perception. The account that I will present of sound change reduces to the following two principles, each of them empirically supported: 1) there is variation and ambiguity in pronunciation, some of it caused by constraints of the speech production mechanism and some of it due to the fact that the mapping between some speech articulations and speech acoustics is a many-to-few mapping, 2) such variation or ambiguity occasionally causes a listener to misinterpret how a word is to be pronounced, such that when that listener in turn speaks, the result is a different pronunciation. I will suggest that this scheme allows us to gain some insight into the basic distinctive features used in speech perception as well as their temporal domain and that this may allow us to improve on attempts to recognize speech by machines.

Presenter

John Ohala is Professor Emeritus in the Department of Linguistics, University of California, Berkeley (UCB). He was also head of the Phonology Laboratory at UCB from 1975 until his retirement in 2004. He received his PhD in Linguistics from UCLA in 1969. He has had post-doctoral positions and other visiting research or teaching positions at the Research Institute of Logopedics and Phoniatrics, (University of Tokyo), Bell Laboratories, University of Copenhagen, and University of Alberta (Edmonton) and the City University of Hong Kong. He was chair of the 2nd ICSLP (1992) in Banff, Canada, and of the 14th International Congress of Phonetic Sciences in San Francisco (1999). He serves or has served on the editorial boards of several speech- and language-related journals, including *Phonetica, Journal of Phonetics, Language & Speech, Language, and Speech Communication*. He was section editor for Speech Science and Phonetics for the Pergamon *Encyclopedia of Language and Linguistics*, 1994. His research interests are in experimental phonology and phonetics and ethological aspects of communication, including speech perception, sound change, phonetic and phonological universals, psycholinguistic studies in phonology, and sound symbolism.
Overview

The speech processing industry has made great progress over the last few decades. There are now mainstream markets for speech recognition and text-to-speech in call centers, mobile devices, automobiles, games, and dictation. In each of these, the technology has progressed to the point where it can provide real value for the end users and there is an active industry of companies delivering a variety of solutions to the market.

But, none of these has gotten to the point of truly ubiquitous use – where speech interfaces are a common and expected means of interaction.

Why is this? Is it due to limitations of the technology, have we not found the right applications, or do users just not find enough value in the use of speech interfaces?

In this talk, I will discuss the current state of the speech industry in the various markets and will provide some thoughts about where speech interfaces may become ubiquitous and what we need to do to make this happen.

Presenter

Mike Phillips has been active in the speech technology world for over twenty years. In 1994, he founded SpeechWorks based on technology that he and others had developed at the Spoken Language Systems group at MIT. Over the next ten years, Mike and team grew SpeechWorks from a small startup in a new market into the market leader in the now established market for speech enabled call center solutions. In 2003, SpeechWorks was acquired by ScanSoft (now named Nuance). Mike stayed on as the CTO of ScanSoft for two years after the acquisition. After spending a year as a visiting scientist at the Spoken Language Systems group at MIT, Mike now has a new startup (called Mobeus) which is focused on multimodal interfaces for mobile devices.
Plenary Talk 3

Statistical language learning in human infants and adults

Elissa L. Newport, University of Rochester

Allegheny Ballrooms, Wednesday, September 20, 9:00 to 10:00 am

Overview

In collaboration with Richard Aslin, I have been developing a statistical approach to language acquisition and investigating the abilities of human learners to perform the computations that would be required for acquiring properties of natural languages by such a method. Our studies have shown that adults, infants, and even nonhuman primates are capable of performing such computations online and with remarkable speed, on both speech and nonspeech materials. Our recent work examines differences between adults, infants, and nonhuman primates in their computational capacities, how these differences may help to explain why these learners differ in their abilities to learn languages, and why languages have some of the properties they have.

Presenter

Elissa L. Newport is the department chair and the George Eastman Professor of Brain and Cognitive Sciences at the University of Rochester. Her primary research interest is in human language acquisition, with research projects including naturalistic studies of children learning their first languages, experimental studies of infants, adults, and non-human primates learning miniature languages in the lab, fieldwork on emerging sign languages, and fMRI research on language and the brain. Professor Newport received her Ph.D. in Psychology at the University of Pennsylvania and was a Sloan Fellow in Linguistics and Cognitive Science at Penn and MIT. She has been on the faculty at the University of California at San Diego, University of Illinois, and, since 1988, University of Rochester. Her research is funded by the NIH, NSF, the McDonnell Foundation, and the Packard Foundation, and for this research has received the Claude Pepper Award of Excellence from NIH. She currently is a series editor for MIT Press, serves on the Board on Behavioral, Cognitive, and Sensory Sciences of the National Academy of Sciences, and she is the Chair of Section J (Psychology) of the American Association for the Advancement of Science. She is a Fellow of the American Academy of Arts & Sciences, the Cognitive Science Society, the Society for Experimental Psychologists, and the American Association for the Advancement of Science, and is a member of the National Academy of Sciences.
Plenary Talk 4

Speech Recognition: The Unfinished Agenda
Raj Reddy, Carnegie Mellon University

Allegheny Ballrooms, Thursday, September 21, 9:00 to 10:00 am

Overview

After many years of intensive research and development, speech science and technology have now reached a stage at which practical speech-based solutions are now in use in a number of application areas. Nevertheless, much remains in making the fruits of speech and language available to all of the world's peoples. Universal access to and continued multilingual development of language technologies is crucial for social and economic development for many reasons. Language barriers can slow down economic growth significantly, and globalization depends on cross-border and cross-language communication. Fluid and effective multi-lingual speech-based interfaces can eliminate cultural and social barriers, enable access to rare (and potentially beneficial) knowledge to all people regardless of their native language, and foster the preservation of minority languages and the cultures and heritage that they represent.

This talk will discuss some of the key issues and opportunities that remain in extending the benefits of language technologies to the entire world, and research challenges of the unfinished agenda.

Presenter

Raj Reddy is the Mozah Bint Nasser University Professor of Computer Science and Robotics in the School of Computer Science at Carnegie Mellon University. He began his academic career as an Assistant Professor at Stanford in 1966. He has been a member of the Carnegie Mellon faculty since 1969. He served as the founding Director of the Robotics Institute from 1979 to 1991 and the Dean of School of Computer Science from 1991 to 1999.

Reddy's research interests include the study of human-computer interaction and artificial intelligence. His current research interests include Million Book Digital Library Project; a Multifunction Information Appliance that can be used by the uneducated; Fiber To The Village Project; Mobile Autonomous Robots; and Learning by Doing.

Raj Reddy is a member of the National Academy of Engineering and the American Academy of Arts and Sciences. He was president of the American Association for Artificial Intelligence from 1987 to 89. Dr. Reddy was awarded the Legion of Honor by President Mitterand of France in 1984. He was awarded the ACM Turing Award in 1994, the Okawa Prize in 2004, the Honda Prize in 2005, and the Vannevar Bush Award in 2006. He served as co-chair of the President’s Information Technology Advisory Committee (PITAC) from 1999 to 2001 under Presidents Clinton and Bush.
Special Sessions

The Speech Separation Challenge

**Date and time:** Monday, September 18, 10:30 to 12:30  
**Location:** Westmoreland Room  
**Organizers:** Martin Cooke, University of Sheffield and Te-Won Lee, UCSD

This session will review and discuss the results of the initial first-ever large-scale global comparison of techniques for separating and recognizing speech, which had been conducted by the session organizers. The task in the evaluation was to recognize speech from a target talker in the presence of either stationary noise or other speech. Participating sites were provided with training data and two test sets consisting of speech in the presence of speech-shaped noise at various SNRs, and speech in the presence of competing speech at various TMRs. Targets and maskers were recorded using only a single microphone.

Articulatory Modeling

**Date and time:** Monday, September 18, 16:30 to 18:30  
**Location:** Westmoreland Room  
**Organizer:** Eric Bateson, University of British Columbia

This session will cover a broad range of topics in articulatory modeling and its applications. Articulatory modeling offers a model of the speech signal and its acoustic details produced from physical constraints in the articulatory domain. Major research topics in the area include model parameter estimation from the speech signal or from sensor data, the use of articulatory features in speech recognition, and speech production from articulatory models.

Speech and Language in Education.

**Date and time:** Tuesday, September 19, 10:30 to 12:30  
**Location:** Westmoreland Room  
**Organizers:** Patti Price, pprice.com and Abeer Alwan, UCLA

The special session on Speech and Language Technology in Education is intended to help form a bridge between research and deployment of speech and language-based educational technology. Its goals are to inform researchers of commercial efforts and real-world conditions of use with a view toward inspiring new research strategies, and to inform commercial efforts and educational representatives of research directions that might be deployed.

The session will consist of a one-hour period of nine poster/demo presentations, followed by a panel discussion with four 10-minute presentations from panelists. The remaining time will be devoted to discussion and questions from the audience.
Special Sessions (continued)

Speech Summarization

**Date and time:** Wednesday, September 20, 10:30 to 12:30  
**Location:** Westmoreland Room  
**Organizers:** Jean Carletta, Edinburgh, Julia Hirschberg, Columbia, and Sadaoki Furui, Tokyo Institute of Technology

This session provides a general introduction to the area of speech summarization and brings together researchers working within it for information exchange and discussion. Papers were invited that report current research in speech summarization for any speech genre, including the summarization of meetings, lectures, broadcast news, and spoken dialogue, using output that is spoken, textual, or multimodal. Methods considered included sentence (or dialogue act) compression, sentence extraction, and text rewriting. Techniques for evaluating summaries will also be discussed.

Spoken Dialog Technology R&D

**Date and time:** Wednesday, September 20, 14:00 to 16:00  
**Location:** Westmoreland Room  
**Organizer:** Roberto Pieraccini, SpeechCycle, Inc.

The industry and market of spoken dialog systems are reaching a never-seen-before maturity. Today there are thousands of deployed voice-enabled systems interacting with millions of users. Small, medium, and large companies are specializing in the various aspects of the industry, such as core technology, platforms, tools, applications, professional services, and hosting. Academic research, which has been traditionally the innovative force behind speech technology, needs to keep its pace of long term ambitious goals in order to help the industry reach higher levels of technological advancements. In this spirit, we would like to foster the communication between longer term research and industrial development in the field of spoken language human-machine communication. This session has been created with the aim of bringing R&D experts from small and large companies to present issues and results to a broader research community.

This session is dedicated to papers on technological and research issues on the design, development, and deployment of commercial spoken dialog systems.
Special Sessions (continued)

The Prosody of Turn-Taking and Dialog Acts

**Date and time:** Wednesday, September 20, 16:30 to 18:30  
**Location:** Westmoreland Room  
**Organizers:** Nigel Ward, University of Texas, El Paso and Elizabeth Shriberg, SRI and ICSI

A recent focus of research interest is the use of prosody for improved utterance-type tagging and for the swift and accurate detection of dialog events such as turn ends, back-channel invitations, disfluencies, and clause and sentence boundaries. Applications include better language modeling, improved pragmatic accuracy in dialog systems, improved timing in dialog systems, and potential insights into the cognitive mechanisms and social dynamics of human interaction.

This workshop will provide a venue not only for research contributions but also for in-depth discussions of underlying issues. The aim is to bring together researchers working from different perspectives to share techniques and attain a clearer understanding of how prosodic markers to dialog acts fit into the larger ecology of prosodic phenomena and the larger picture of spoken language applications.

From Ideas to Companies (Panel discussion)

**Date and time:** Thursday, September 21, 10:30 to 12:30  
**Location:** Westmoreland Room  
**Organizer:** Janet M. Baker, Saras Institute/Dibner Institute MIT

Ever wonder what it's like to start and run your own speech company? What's it like to take your ideas, your technology, expertise, and drive to make speech products and services for other people to use? Where are the primary challenges, good experiences, and not-so-good? Risks? Rewards? Is it worthwhile? Is it for you? Well here's a chance to find out! Several very knowledgeable founders of speech technology companies will each give presentations on their personal experiences and insights into this challenging endeavor. You, the audience, are then invited to ask them specific questions, on issues ranging from R&D, sales and marketing, finance, personnel, intellectual property, to personal effects on them and their families!
# Session Table of Contents

<table>
<thead>
<tr>
<th>Session ID</th>
<th>Session Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon1A1O</td>
<td>Language Modeling for Spoken Dialog Systems</td>
<td>52</td>
</tr>
<tr>
<td>Mon1A2O</td>
<td>Feature Enhancement for Robust ASR</td>
<td>53</td>
</tr>
<tr>
<td>Mon1A3O</td>
<td>Dialog and Discourse</td>
<td>54</td>
</tr>
<tr>
<td>Mon1WeS</td>
<td>The Speech Separation Challenge</td>
<td>55</td>
</tr>
<tr>
<td>Mon1BuP</td>
<td>Multilingual and Multi-Accent Processing</td>
<td>56</td>
</tr>
<tr>
<td>Mon1CaP</td>
<td>Corpora, Annotation, and Assessment Metrics I</td>
<td>58</td>
</tr>
<tr>
<td>Mon1FoP</td>
<td>Speech Coding</td>
<td>60</td>
</tr>
<tr>
<td>Mon2A1O</td>
<td>Speech Enhancement I</td>
<td>62</td>
</tr>
<tr>
<td>Mon2A2O</td>
<td>ASR Other I</td>
<td>63</td>
</tr>
<tr>
<td>Mon2A3O</td>
<td>Modeling Prosodic Features</td>
<td>64</td>
</tr>
<tr>
<td>Mon2WeO</td>
<td>Spoken Information Retrieval</td>
<td>65</td>
</tr>
<tr>
<td>Mon2BuP</td>
<td>Front-End Methods for ASR</td>
<td>66</td>
</tr>
<tr>
<td>Mon2CaP</td>
<td>Language and Dialect Recognition</td>
<td>68</td>
</tr>
<tr>
<td>Mon2FoP</td>
<td>Spoken Dialog Systems I</td>
<td>70</td>
</tr>
<tr>
<td>Mon3A1O</td>
<td>Speaker Characterization and Recognition I</td>
<td>72</td>
</tr>
<tr>
<td>Mon3A2O</td>
<td>System Combination</td>
<td>73</td>
</tr>
<tr>
<td>Mon3A3O</td>
<td>Interpreting Prosodic Variation</td>
<td>74</td>
</tr>
<tr>
<td>Mon3WeS</td>
<td>Articulatory Modeling</td>
<td>75</td>
</tr>
<tr>
<td>Mon3BuP</td>
<td>Acoustic Modeling I — Training and Topologies</td>
<td>75</td>
</tr>
<tr>
<td>Mon3CaP</td>
<td>Acoustic Signal Segmentation and Classification</td>
<td>77</td>
</tr>
<tr>
<td>Mon3FoP</td>
<td>Linguistics, Phonology, and Phonetics I</td>
<td>79</td>
</tr>
<tr>
<td>Tue1A1O</td>
<td>Speech Translation</td>
<td>83</td>
</tr>
<tr>
<td>Tue1A2O</td>
<td>Acoustic Modeling II — Adaptation</td>
<td>84</td>
</tr>
<tr>
<td>Tue1A3O</td>
<td>Emotional Speech and Speaker State</td>
<td>85</td>
</tr>
<tr>
<td>Tue1WeS</td>
<td>Speech and Language in Education</td>
<td>86</td>
</tr>
<tr>
<td>Tue1BuP</td>
<td>Speech Perception I</td>
<td>87</td>
</tr>
<tr>
<td>Tue1CaP</td>
<td>Speaker Characterization and Recognition II</td>
<td>89</td>
</tr>
<tr>
<td>Tue1FoP</td>
<td>Speech Production, Physiology, and Pathology I</td>
<td>91</td>
</tr>
<tr>
<td>Tue2A1O</td>
<td>Formant Estimation</td>
<td>94</td>
</tr>
<tr>
<td>Tue2A2O</td>
<td>Language Processing Beyond and Below the Word-Level</td>
<td>95</td>
</tr>
<tr>
<td>Tue2A3O</td>
<td>Spoken Dialog Systems II</td>
<td>96</td>
</tr>
<tr>
<td>Tue2WeO</td>
<td>Corpora, Annotation, and Assessment Metrics II</td>
<td>97</td>
</tr>
<tr>
<td>Tue2BuP</td>
<td>Robustness and Adaptation for ASR</td>
<td>98</td>
</tr>
<tr>
<td>Tue2CaP</td>
<td>Multimodal, Translation and Information Retrieval</td>
<td>100</td>
</tr>
<tr>
<td>Tue3A1O</td>
<td>Advances in Acoustic Segmentation</td>
<td>103</td>
</tr>
<tr>
<td>Tue3A2O</td>
<td>Acoustic Modeling III — LVCSR</td>
<td>104</td>
</tr>
<tr>
<td>Tue3A3O</td>
<td>Linguistics, Phonology, and Phonetics II</td>
<td>105</td>
</tr>
<tr>
<td>Tue3WeO</td>
<td>Speech and Visual Processing</td>
<td>106</td>
</tr>
<tr>
<td>Tue3BuP</td>
<td>TTS I</td>
<td>107</td>
</tr>
<tr>
<td>Tue3CaP</td>
<td>Special Populations — Learners, Aged, Challenged</td>
<td>109</td>
</tr>
<tr>
<td>Tue3FoP</td>
<td>Speech Enhancement II</td>
<td>111</td>
</tr>
<tr>
<td>Wed1A1O</td>
<td>Speaker Characterization and Recognition III</td>
<td>114</td>
</tr>
<tr>
<td>Day</td>
<td>Session</td>
<td>Topic</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>Wed1A2O</td>
<td>Robust ASR</td>
<td></td>
</tr>
<tr>
<td>Wed1A3O</td>
<td>Speech Perception II</td>
<td></td>
</tr>
<tr>
<td>Wed1WeS</td>
<td>Speech Summarization</td>
<td></td>
</tr>
<tr>
<td>Wed1BuP</td>
<td>Acoustic Modeling IV</td>
<td></td>
</tr>
<tr>
<td>Wed1CaP</td>
<td>Large Vocabulary Speech Recognition</td>
<td></td>
</tr>
<tr>
<td>Wed1FoP</td>
<td>Speech/Noise/Music Segmentation</td>
<td></td>
</tr>
<tr>
<td>Wed2A1O</td>
<td>Pitch Estimation</td>
<td></td>
</tr>
<tr>
<td>Wed2A2O</td>
<td>Acoustic Modeling V — Novel Approaches</td>
<td></td>
</tr>
<tr>
<td>Wed2A3O</td>
<td>Corpus-Based Synthesis</td>
<td></td>
</tr>
<tr>
<td>Wed2WeS</td>
<td>Spoken Dialog Technology R&amp;D</td>
<td></td>
</tr>
<tr>
<td>Wed2BuP</td>
<td>Modeling Speaker Emotional State</td>
<td></td>
</tr>
<tr>
<td>Wed2CaP</td>
<td>Language Modeling and ASR Applications</td>
<td></td>
</tr>
<tr>
<td>Wed3A1O</td>
<td>Segmentation and VAD</td>
<td></td>
</tr>
<tr>
<td>Wed3A3O</td>
<td>Technologies for Specific Populations: Learners and Challenged</td>
<td></td>
</tr>
<tr>
<td>Wed3WeS</td>
<td>The Prosody of Turn-Taking and Dialog Acts</td>
<td></td>
</tr>
<tr>
<td>Wed3BuP</td>
<td>TTS II</td>
<td></td>
</tr>
<tr>
<td>Wed3CaP</td>
<td>Speaker Characterization and Recognition IV</td>
<td></td>
</tr>
<tr>
<td>Wed3FoP</td>
<td>Multichannel Speech Enhancement/Speech Perception</td>
<td></td>
</tr>
<tr>
<td>Thu1A1O</td>
<td>Diarization in ASR</td>
<td></td>
</tr>
<tr>
<td>Thu1A2O</td>
<td>Language Model Adaptation, Refinement, and Evaluation</td>
<td></td>
</tr>
<tr>
<td>Thu1A3O</td>
<td>Speech Production, Physiology, and Pathology II</td>
<td></td>
</tr>
<tr>
<td>Thu1BuP</td>
<td>Voice Morphing</td>
<td></td>
</tr>
<tr>
<td>Thu1CaP</td>
<td>ASR Other II</td>
<td></td>
</tr>
<tr>
<td>Thu1FoP</td>
<td>Prosody</td>
<td></td>
</tr>
<tr>
<td>Thu2A1O</td>
<td>Discriminative Training</td>
<td></td>
</tr>
<tr>
<td>Thu2A3O</td>
<td>Speech Synthesis</td>
<td></td>
</tr>
<tr>
<td>Thu2WeO</td>
<td>Multimodal Processing</td>
<td></td>
</tr>
<tr>
<td>Thu2BuP</td>
<td>Speech Analysis</td>
<td></td>
</tr>
<tr>
<td>Thu2CaP</td>
<td>Advances in Noisy ASR</td>
<td></td>
</tr>
<tr>
<td>Thu2FoP</td>
<td>Source Separation and Localization</td>
<td></td>
</tr>
</tbody>
</table>
Robust Interpretation in Dialogue by Combining Confidence Scores with Contextual Features
Matthew Purver\textsuperscript{1}, Florin Ratiu\textsuperscript{1}, Lawrence Cavedon\textsuperscript{2};
\textsuperscript{1}Stanford University, USA; \textsuperscript{2}National ICT Australia, Australia
Mon1A1O.1
We present an approach to dialogue management and interpretation that evaluates and selects amongst candidate dialogue moves based on features at multiple levels. Multiple interpretation methods can be combined, multiple speech recognition and parsing hypotheses tested, and multiple candidate dialogue moves considered to choose the highest scoring hypothesis overall. We integrate hypotheses generated from shallow slot-filling methods and from relatively deep parsing, using pragmatic information. We show that this gives more robust performance than using either approach alone, allowing n-best list reordering to correct errors in speech recognition or parsing.

A Clustering Approach to Semantic Decoding
Hui Ye, Steve Young; University of Cambridge, UK
Mon1A1O.2
This paper presents a novel algorithm for semantic decoding in spoken language understanding systems. Unlike conventional semantic parsers which either use hand-crafted rules or statistical models trained from fully annotated data, the proposed approach uses an unsupervised sentence clustering technique called Y-clustering to automatically select a set of exemplar sentences from a training corpus. These exemplars are combined with simple sentence-level semantic annotations to form templates which are then used for semantic decoding. The performance of this approach was evaluated in the travel domain using the ATIS corpus. Training is fast and cheap, and the results are significantly better than those achieved using HMM-based or stack-based statistical parsers.

A Bootstrapping Approach for Developing Language Model of New Spoken Dialogue Systems by Selecting Web Texts
Teruhisa Misu, Tatsuya Kawahara; Kyoto University, Japan
Mon1A1O.3
This paper proposes a bootstrapping method of constructing statistical language models for new spoken dialogue systems by collecting and selecting sentences from the World Wide Web (WWW). To make effective search queries that cover the target domain in full detail, we exploit the document set described about the target domain as seeding data. An important issue is how to filter the retrieved Web pages, since all of the retrieved Web texts are not necessarily suitable as training data. We induct an existing dialogue corpus of different domain to prefer the texts of spoken style. The proposed method was evaluated on two different tasks of software support and sightseeing guidance, and significant reduction of the word error rate was achieved. We show that it is vital to incorporate the dialogue corpus, though not relevant to the target domain, in the text selection phase.

Phoneme-to-Grapheme Mapping for Spoken Inquiries to the Semantic Web
Axel Horndasch\textsuperscript{1}, Elmar Nöth\textsuperscript{1}, Anton Batliner\textsuperscript{1}, Volker Warnke\textsuperscript{2}; \textsuperscript{1}Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany; \textsuperscript{2}Sympalog Voice Solutions GmbH, Germany
Mon1A1O.4
Automatic methods for grapheme-to-phoneme (G2P) and phoneme-to-grapheme (P2G) conversion have become very popular in recent years. Their performance has improved considerably, while at the same time these developments required less input from expert lexicographers. Continuing in this tradition we will present in this paper a data-driven, language-independent approach called MASSIVE with which it is possible to create efficient online modules for automatic symbol mapping. Our framework is solely based on statistical methods for training and run-time and has been optimized for P2G conversion in the context of spoken inquiries to the Semantic Web, an issue researched in the SmartWeb project. MASSIVE systems can be trained using a pronunciation lexicon, the output of a phone recognizer or any other suitable set of corresponding symbol strings. Successful tests have been performed on German and English data sets.

Bootstrapping Language Models for Dialogue Systems
Karl Weilhammer, Matthew N. Stuttle, Steve Young; University of Cambridge, UK
Mon1A1O.5
We report results on rapidly building language models for dialogue systems. Our base line is a recogniser using a grammar network. We show that we can almost halve the word error rate (WER) by combining language models generated from a simple task grammar with a standard speech corpus and data collected from the web using a sentence selection algorithm based on relative perplexity. This model compares very well to a language model using "in-domain" data from a Wizard Of Oz (WOZ) collection. We strongly advocate the use of statistical language models (SLMs) in speech recognisers for dialogue systems and show that costly WOZ data collections are not necessary to build SLMs.

Question Answering with Discriminative Learning Algorithms
Junlan Feng; AT&T Labs Research, USA
Mon1A1O.6
In this paper, we describe a discriminative learning approach for question answering. Our training corpus consists of 2 million Frequently Asked Questions (FAQs) and their corresponding answers that we mined from the World Wide Web. This corpus is used to train the lexical and semantic association model between questions and answers. We evaluate our approach on two question answering tasks: 2003 Text Retrieval Conference Question Answering Task, and finding answers to FAQs. In both cases, the proposed approach achieved significant improvements over the results for an information retrieval based question answering model.
Feature Normalization Using Smoothed Mixture Transformations

Patrick Kenny, Vishwa Gupta, G. Boulianne, Pierre Ouellet, Pierre Dumouchel; Centre de Recherche Informatique de Montréal, Canada

We propose a method for estimating the parameters of SPLICE-like transformations from individual utterances so that this type of transformation can be used to normalize acoustic feature vectors for speech recognition on an utterance-by-utterance basis in a similar manner to cepstral mean normalization. We report results on an in-house French language multi-speaker database collected while deploying an automatic closed-captioning system for live broadcast news. An unusual feature of this database is that there are very large amounts of training data for the individual speakers (typically several hours) so that it is very difficult to improve on multi-speaker modeling by using standard methods of speaker adaptation. We found that the proposed method of feature normalization is capable of achieving a 6% relative improvement over cepstral mean normalization on this task.

Stochastic Vector Mapping-Based Feature Enhancement Using Prior Model and Environment Adaptation for Noisy Speech Recognition

Chia-Hsin Hsieh, Chung-Hsien Wu, Jun-Yu Lin; National Cheng Kung University, Taiwan

This paper presents an approach to feature enhancement for noisy speech recognition. Three prior models are introduced to characterize clean speech, noise and noisy speech respectively using sequential noise estimation based on noise-normalized stochastic vector mapping. Environment adaptation is also adopted to reduce the mismatch between training data and test data. For AURORA2 database, the experimental results indicate that a 0.77% digit accuracy improvement for multi-condition training and 0.29% digit accuracy improvement for clean speech training were achieved without stereo training data compared to the SPLICE-based approach with recursive noise estimation. For MAT-BN Mandarin broadcast news database, a 2.6% syllable accuracy improvement for anchor speech and 4.2% syllable accuracy improvement for field report speech were obtained compared to the MCE-based approach.

A Framework for Robust MFCC Feature Extraction Using SNR-Dependent Compression of Enhanced Mel Filter Bank Energies

Babak Nasersharif, Ahmad Akbari; Iran University of Science & Technology, Iran

We previously proposed a noise compensation method for Mel filter bank energies and so MFCC features. This compensation method includes two steps: Mel sub-band spectral subtraction and then compression of Mel-Sub-band energies. In the compression step, we propose a sub-band SNR-dependent compression function. We use this function instead of logarithm function in conventional MFCC feature extraction in presence of additive noise. Experimental results show that the proposed method significantly improves MFCC features performance in noisy conditions where it decreases word error rate about 70% in SNR value of 0 dB for different types of additive noise.
A Stochastic Approach for Dialog Management Based on Neural Networks

Lluís F. Hurtado, David Griol, Encarna Segarra, Emilio Sanchis; Universitat Politècnica de València, Spain

In this article, we present an approach for the construction of a stochastic dialog manager, in which the system answer is selected by means of a classification procedure. In particular, we use neural networks for the implementation of this classification process, which takes into account the data supplied by the user and the last system turn. The stochastic model is automatically learnt from training data which are labeled in terms of dialog acts. An important characteristic of this approach is the introduction of a partition in the space of sequences of dialog acts in order to deal with the scarcity of available training data. This system has been developed in the DIHANA project, whose goal is the design and development of a dialog system to access a railway information system using spontaneous speech in Spanish. An evaluation of this approach is also presented.

Discourse Structure and Speech Recognition Problems

Mihai Rotaru, Diane J. Litman; University of Pittsburgh, USA

We study dependencies between discourse structure and speech recognition problems (SRP) in a corpus of speech-based computer tutoring dialogues. This analysis can inform us whether there are places in the discourse structure prone to more SRP. We automatically extract the discourse structure by taking advantage of how the tutoring information is encoded in our system. To quantify the discourse structure, we extract two features for each system turn: depth of the turn in the discourse structure and the type of transition from the previous turn to the current turn. The χ² test is used to find significant dependencies. We find several interesting interactions which suggest that the discourse structure can play an important role in several dialogue related tasks: automatic detection of SRP and analyzing spoken dialogues systems with a large state space from limited amounts of available data.

A TextTiling Based Approach to Topic Boundary Detection in Meetings

Satyanjeev Banerjee, Alexander I. Rudnicky; Carnegie Mellon University, USA

Our goal is to automatically detect boundaries between discussions of different topics in meetings. Towards this end we adapt the TextTiling algorithm [1] to the context of meetings. Our features include not only the overlapped words between adjacent windows, but also overlaps in the amount of speech contributed by each meeting participant. We evaluate our algorithm by comparing the automatically detected boundaries with the true ones, and computing precision, recall and f-measure. We report average precision of 0.85 and recall of 0.59 when segmenting unseen test meetings. Error analysis of our results shows that although the basic idea of our algorithm is sound, it breaks down when participants stray from typical behavior (such as when they monopolize the conversation for too long).

An User-Centered Development of an Intuitive Dialog Control for Speech-Controlled Music Selection in Cars

Stefan Schulz, Hilko Donker; Dresden University of Technology, Germany

During the last years speech dialog systems have proven valuable for controlling infoainment systems in cars. The increasing popularity of mobile MP3 players and their capability to store more and more MP3, suggest the development of a speech control function for car MP3 players. With technical advantages in hardware and speech recognition it is possible to focus on the development of an intuitive speech dialog, which ensures that the advantages of speech control can be utilized while disadvantages are minimal at the same time. This paper describes the user-centered development of such an intuitive dialog, its special challenges and the reached quality in usability and ease of use. Also the generalizability of the results is discussed.

Doing Research on a Deployed Spoken Dialogue System: One Year of Let’s Go! Experience

Antoine Raux, Dan Bohus, Brian Langner, Alan W. Black, Maxine Eskenazi; Carnegie Mellon University, USA

This paper describes our work with Let’s Go, a telephone-based bus schedule information system that has been in use by the Pittsburgh population since March 2005. Results from several studies show that while task success correlates strongly with speech recognition accuracy, other aspects of dialogue such as turn-taking, the set of error recovery strategies, and the initiative style also significantly impact system performance and user behavior.

Detecting Question-Bearing Turns in Spoken Tutorial Dialogues

Jackson Liscombe, Jennifer J. Venditti, Julia Hirschberg; Columbia University, USA

Current speech-enabled Intelligent Tutoring Systems do not model student question behavior the way human tutors do, despite evidence indicating the importance of doing so. Our study examined a corpus of spoken tutorial dialogues collected for development of ITSpeak, an Intelligent Tutoring Spoken Dialogue System. The authors extracted prosodic, lexical, syntactic, and student and task dependent information from student turns. Results of running 5-fold cross validation machine learning experiments using AdaBoosted C4.5 decision trees show prediction of student question-bearing turns at a rate of 79.7%. The most useful features were prosodic, especially the pitch slope of the last 200 milliseconds of the student turn. Student pre-test score was the most-used feature. Findings indicate that using turn-based units is acceptable for incorporating question detection capability into practical Intelligent Tutoring Systems.
Mon1WeS : The Speech Separation Challenge
Special Session, Westmoreland Room, 10:30 – 12:30
Chair: Te-Won Lee, University of California at San Diego, USA

A Computational Auditory Scene Analysis System for Robust Speech Recognition
Soundararajan Srinivasan, Yang Shao, Zhaozhong Jin, Deliang Wang; Ohio State University, USA

Mon1WeS.1
We present a computational auditory scene analysis system for separating and recognizing target speech in the presence of competing speech or noise. We estimate, in two stages, the ideal binary time-frequency (T-F) mask which retains the mixture in a local T-F unit if and only if the target is stronger than the interference within the unit. In the first stage, we use harmonicity to segregate the voiced portions of individual sources in each time frame based on multipitch tracking. Additionally, unvoiced portions are segmented based on an onset/offset analysis. In the second stage, speaker characteristics are used to group the T-F units across time frames. The resulting T-F masks are used in conjunction with missing-data methods for recognition. Systematic evaluations on a speech separation challenge task show significant improvement over the baseline performance.

CASA Based Speech Separation for Robust Speech Recognition
Runqiang Han, Pei Zhao, Qin Gao, Zhiping Zhang, Hao Wu, Xihong Wu; Peking University, China

Mon1WeS.2
This paper introduces a speech separation system as a front-end processing step for automatic speech recognition (ASR). It employs computational auditory scene analysis (CASA) to separate the target speech from the interference speech. Specifically, the mixed speech is preprocessed based on auditory peripheral model. Then a pitch tracking is conducted and the dominant pitch is used as a main cue to find the target speech. Next, the time frequency (T-F) units are merged into many segments. These segments are then combined into streams via CASA initial grouping. A regrouping strategy is employed to refine these streams via amplitude modulate (AM) cues, which are finally organized by the speaker recognition techniques into corresponding speakers. Finally, the output streams are reconstructed to compensate the missing data in the abovementioned processing steps by a cluster based feature reconstruction. Experimental results of ASR show that at low TMR (<-6dB) the proposed method offers significantly higher recognition accuracy.

Enhancement of Harmonic Content of Speech Based on a Dynamic Programming Pitch Tracking Algorithm
Mark R. Every, Philip J.B. Jackson; University of Surrey, UK

Mon1WeS.3
For pitch tracking of a single speaker, a common requirement is to find the optimal path through a set of voiced or voiceless pitch estimates over a sequence of time frames. Dynamic programming (DP) algorithms have been applied before to this problem. Here, the pitch candidates are provided by a multi-channel autocorrelation-based estimator, and DP is extended to pitch tracking of multiple concurrent speakers. We use the resulting pitch information to enhance harmonic content in noisy speech and to obtain separations of target from interfering speech.

Recent Advances in Speech Fragment Decoding Techniques
Jon Barker, André Coy, Ning Ma, Martin Cooke; University of Sheffield, UK

Mon1WeS.4
This paper addresses the problem of recognising speech in the presence of a competing speaker. We employ a speech fragment decoding technique that treats segregation and recognition as coupled problems. Data-driven techniques are used to segment a spectro-temporal representation into a set of spectro-temporal fragments, such that each fragment is dominated by one or other of the speech sources. A speech fragment decoder is used which employs missing data techniques and clean speech models to simultaneously search for the set of fragments and the word sequence that best matches the target speaker model. The paper reports recent advances in this technique, and presents an evaluation based on artificially mixed speech utterances. The fragment decoder produces significantly lower error rates than a conventional recogniser, and mimics the pattern of human performance whereby performance increases as the target-masker ratio is reduced below -3 dB.

Speech Recognition Using Factorial Hidden Markov Models for Separation in the Feature Space
Tuomas Virtanen; Tampere University of Technology, Finland

Mon1WeS.5
This paper proposes an algorithm for the recognition and separation of speech signals in non-stationary noise, such as another speaker. We present a method to combine hidden Markov models (HMMs) trained for the speech and noise into a factorial HMM to model the mixture signal. Robustness is obtained by separating the speech and noise signals in a feature domain, which discards unnecessary information. We use mel-cepstral coefficients (MFCCs) as features, and estimate the distribution of mixture MFCCs from the distributions of the target speech and noise. A decoding algorithm is proposed for finding the state transition paths and estimating gains for the speech and noise from a mixture signal. Simulations were carried out using speech material where two speakers were mixed at various levels, and even for high noise level (9 dB above the speech level), the method produced relatively good (60% word recognition accuracy) results. Audio demonstrations are available at www.cs.tut.fi/~tuomasv.

Combining Missing-Feature Theory, Speech Enhancement and Speaker-Dependent/-Independent Modeling for Speech Separation
Ji Ming1, Timothy J. Hazen2, James R. Glass2; 1 Queen’s University Belfast, UK; 2 Massachusetts Institute of Technology, USA

Mon1WeS.6
This paper considers the recognition of speech given in the form of two mixed sentences, spoken by the same talker or by two different talkers. The database published on the ICSLP’2006 website for Two-Talker Speech Separation Challenge is used in the study. A system that recognizes and reconstructs both sentences from the given mixture is described. The system involves a combination of several different techniques, including a missing-feature approach for improving crosstalk/noise robustness, Wiener filtering for speech restoration, HMM-based speech reconstruction, and speaker-dependent/-independent modeling for speaker/speech recognition. For clean speech recognition, the system obtained a word accuracy rate of 96.7%. For the two-talker speech separation challenge task, the system obtained 81.4% at 6 dB TMR (target-to-masker ratio) and 34.1% at -9 dB TMR.

Super-Human Multi-Talker Speech Recognition: The IBM 2006 Speech Separation Challenge System
T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, Ramesh Gopinath; IBM T.J. Watson Research Center, USA

Mon1WeS.7
We describe a system for model based speech separation which achieves super-human recognition performance when two talkers...
speak at similar levels. The system can separate the speech of two
speakers from a single channel recording with remarkable results.
It incorporates a novel method for performing two-talker speaker
identification and gain estimation. We extend the method of model
based high resolution signal reconstruction to incorporate temporal
dynamics. We report on two methods for introducing dynamics;
the first uses dynamics in the acoustic model space, the second
incorporates dynamics based on sentence grammar. The addition of
temporal constraints leads to dramatic improvements in the separa-

tion performance. Once the signals have been separated they are then
recognized using speaker dependent labeling.

**Modified Phase Opponency Based Solution to the
Speech Separation Challenge**

*Om D. Deshmukh, Carol Y. Espy-Wilson; University of
Maryland, USA*

**Mon1BuP.1**

In this work, we present a single-channel speech enhancement
technique called the Modified Phase Opponency (MPO) model as a
solution to the Speech Separation Challenge. The MPO model is based
on a neural model for detection of tones-in-noise called the Phase
Opponency (PO) model. Replacing the noisy speech signals by the

*Mon1BuP.2**

**Automatic Speech Recognition of Cantonese-English
Code-Mixing Utterances**

*Joyce Y.C. Chan, P.C. Ching, Tan Lee, Houwei Cao;
Chinese University of Hong Kong, China*

**Mon1BuP.3**

This paper describes our recent work on the development of a large-

vocabulary, speaker-independent, continuous speech recognition
system for Cantonese-English code-mixing utterances. The details of
both acoustic modeling and language modeling will be discussed. For
acoustic modeling, Cantonese accents in English words are handled
by applying cross-lingual acoustic units, as well as modifications in
pronunciation dictionary. Statistic language models are built from
a small amount of text data, as there are many limitations on data
collection. Language boundary detection based on language identifi-
cation algorithms is applied, and it offers a slight improvement to the
overall accuracy. The recognition accuracy for Chinese characters and
English lexicons in the code-mixing utterances is 56.37% and 52.99%,
respectively.

**The ICSI+ Multilingual Sentence Segmentation System**

*M. Zimmerman1, Dilek Hakkani-Tür1, J. Fung1, N.
Mirghafori1, L. Gottlieb1, Elizabeth Shriberg2, Yang Liu3;*

*1International Computer Science Institute, USA; 2SRI
International, USA; 3University of Texas at Dallas, USA*

**Mon1BuP.4**

The ICSI+ multilingual sentence segmentation with results for English
and Mandarin broadcast news automatic speech recognizer transcrip-

tions represents a joint effort involving ICSI, SRI, and UT Dallas.
Our approach is based on using hidden event language models for
exploiting lexical information, and maximum entropy and boosting
classifiers for exploiting lexical, as well as prosodic, speaker change
and syntactic information. We demonstrate that the proposed method-
ology including pitch- and energy-related prosodic features performs
significantly better than a baseline system that uses words and simple
pause features only. Furthermore, the obtained improvements are
consistent across both languages, and no language-specific adaptation
of the methodology is necessary. The best results were achieved by
combining hidden event language models with a boosting-based
classifier that to our knowledge has not previously been applied for
this task.

**Cross-Language Evaluation of Voice-to-Phoneme
Conversions for Voice-Tag Application in Embedded
Platforms**

*Yan Ming Cheng, Changxue Ma, Lynette Melnar; Motorola
Labs, USA*

**Mon1BuP.5**

Previously, we proposed two voice-to-phoneme conversion algorithms
for speaker-independent voice-tag creation specifically targeted at
applications on embedded platforms, an environment sensitive to CPU and memory resource consumption [1]. These two algorithms (batch mode and sequential) were applied in a same-language context, i.e., both acoustic model training and voice-tag creation and application were performed on the same language.

In this paper, we investigate the cross-language application of these two voice-to-phone conversion algorithms, where the acoustic models are trained on a particular source language while the voice-tags are created and applied on a different target language. Here, both algorithms create phonetic representations of a voice-tag of a target language based on the speaker-independent acoustic models of a distinct source language. Our experiments show that recognition performance of these voice-tags vary depending on the source-target language pair, with the variation reflecting the predicted phonological similarity between the source and target languages. Among the most similar languages, performance nears that of the native-trained models and surpasses the native reference baseline.

A Multi-Space Distribution (MSD) Approach to Speech Recognition of Tonal Languages

Huanliang Wang1, Yao Qian2, Frank K. Soong2, Jian-Lai Zhou2, Jiqing Han1, Jian-Lai Zhou2, Harbin Institute of Technology, China; 2Microsoft Research Asia, China

Mon1BuP.6

Tone plays an important role in recognizing spoken tonal languages like Chinese. However, the F0 contour discontinuity between voiced and unvoiced segments has traditionally been a bottleneck in modeling tone contour for automatic speech recognition and synthesis. Various heuristic approaches were proposed to get around the problem. The Multi-Space Distribution (MSD) was proposed by Tokuda et al. and applied to HMM-based speech synthesis, which models the two probability spaces, discrete for unvoiced region and continuous for voiced F0 contour, in a linearly weighted mixture. We extend the MSD to tone modeling for speech recognition applications. Specifically, modeling tones in speaker-independent, spoken Chinese is formulated and tested in a Mandarin speech database. The tone features and spectral features are further separated into two streams and stream-dependent models are built to cluster the two features into separated decision trees. The recognition results show that the ultimate performance of tonal syllable error rate can be improved from toneless baseline system to the MSD based stream-dependent system, 50.5% to 36.1% and 46.3% to 35.1%, for the two systems resulted from using two different phone sets. The absolute tonal syllable error rate improvement of the new approach is 5.5% and 6.1%, comparing with the conventional tone modeling.

Comparison of Acoustic Modeling Techniques for Vietnamese and Khmer ASR

Viet Bac Le, Laurent Besacier; CLIPS, France

Mon1BuP.7

This paper presents a comparison of some different acoustic modeling strategies for under-resourced languages. When only limited speech data are available for under-resourced languages, we propose some crosslingual acoustic modeling techniques. We apply and compare these techniques in Vietnamese ASR. Since there is no pronunciation dictionary for some under-resourced languages, we investigate grapheme-based acoustic modeling. Some initialization techniques for context independent modeling and some question generation techniques for context dependent modeling are applied and compared for Khmer ASR.

Multi-Accent Chinese Speech Recognition

Yi Liu, Pascale Fung; Hong Kong University of Science & Technology, China

Mon1BuP.8

Multiple accents are often present in spontaneous Chinese Mandarin speech as most Chinese have learned Mandarin as a second language. We propose a method to handle multiple accents as well as standard speech in a speaker-independent system by merging auxiliary accent decision trees with standard trees and reconstruct the acoustic model. In our proposed method, tree structures and shape are modified according to accent-specific data while the parameter set of the baseline model remains the same. The effectiveness of this approach is evaluated on Cantonese and Wu accented, as well as standard Mandarin speech. Our method yields a significant 4.4% and 3.3% absolute word error rate reduction without sacrificing the performance on standard Mandarin speech.

Comparative Analysis of Formants of British, American and Australian Accents

Seyed Ghorshi, Saeed Vaseghi, Qin Yan; Brunel University, UK

Mon1BuP.9

This paper compares and quantifies the differences between formants of speech across accents. The cross entropy information measure is used to compare the differences between the formants of the vowels of three major English accents namely British, American and Australian. An improved formant estimation method, based on a linear prediction (LP) model feature analysis and a hidden Markov model (HMM) of formants, is employed for estimation of formant trajectories of vowels and diphthongs. Comparative analysis of the formant space of the three accents indicates that these accents are mostly conveyed by the first two formants. The third and fourth formants exhibit some significant differences across accents for only a few phonemes most notably the variants of vowel ‘r’ in the American (rhotic) accent compared to British (non-rhotic accent). The issue of speaker variability versus accent variability is examined by comparing the cross-entropies of speech models trained on different groups of speakers within and across the accents.

Automatic Initial/Final Generation for Dialectal Chinese Speech Recognition

Linquan Liu, Thomas Fang Zheng, Wenhui Wu; Tsinghua University, China

Mon1BuP.10

Phonetic differences always exist between any Chinese dialect and standard Chinese (Putonghua). In this paper, a method, named automatic dialect-specific Initial/Final (IF) generation, is proposed to deal with the issue of phonemic difference which can automatically produce the dialect-specific units based on model distance measure. A dialect-specific decision tree regrowing method is also proposed to cope with the tri-IF expansion due to the introduction of dialect-specific IFs (DIFs). In combination with a certain adaptation technique, the proposed methods can achieve a syllable error rate (SER) reduction of 18.5% for Shanghai-accented Chinese compared with the Putonghua-based baseline while the use of the DIF set only can lead to an SER reduction of 5.5%.

Maximum Entropy Modeling for Diacritization of Arabic Text

Ruhi Sarikaya1, Ossama Enam2, Imed Zitouni1, Yuqing Gao1, IBM T.J. Watson Research Center, USA; 2IBM Egypt, Egypt

Mon1BuP.11

We propose a novel modeling framework for automatic diacritization of Arabic text. The framework is based on Markov modeling where each grapheme is modeled as an state emitting a diacritic (or none) from the diacritic space. This space is exactly defined using 13 diacritics and a null-diacritic and covers all the diacritics used in any Arabic text. The state emission probabilities are estimated using maximum entropy (MaxEnt) models. The diacritization process is formulated as a search problem where the most likely diacritization realization is assigned to a given sentence. We also propose a diacritization parse tree (DPT) for Arabic that allows joint representation of diacritics,
We used such a system in a study where human subjects perform the drawbacks of both prototype-based and Wizard of Oz simulations. Transcripts of any desired Word Error Rate (WER), thus overcoming an Automatic Speech Recognition (ASR) system that generates realistic broadcast archive. In this paper, we proposed a procedure for prototyping archives is the lack of text transcripts of the audio channel of the webcast archive. One of the hurdles users face when browsing and skimming through webcasts are archived and accessed numerous times retrospectively. Most of the most important issues. This paper proposes a new technique of applying a statistical multidimensional scaling method to analyze the reusability of a speech corpus. In the experiment using six speech corpora, which contains isolated words and short sentences used in car navigation system, an effect of the proposed method is evaluated by a usual approach of cross task recognition. Furthermore, the relationship among those speech corpora is clearly shown by the proposed method.

We describe some aspects of the new lexicon we are producing, Com-bilex, whose structure and implementation is specifically designed to reduce redundancy and improve the representation of productive elements of English. Most importantly, many English words are predictable derivations of baseforms, or compounds. Storing the lexicon as a combination of baseforms and derivational rules speeds up lexicon development, and improves coverage and maintainability.

It is indispensable to establish an objective test methodology for noise-reduced speech. This paper proposes a new methodology which estimates word intelligibility of the noise-reduced speech from PESQ MOS (subjective MOS estimated by the PESQ). To evaluate the effectiveness of the proposed methodology, a word intelligibility test of the noise-reduced speech was performed by using four noise reduction algorithms and word lists which take word difficulty into account, and then the word intelligibility was estimated by the proposed methodology. The results confirmed that the word intelligibility can be estimated well from the PESQ MOS without distinguishing the noise reduction algorithms and the noise types.

The increased availability of broadband connections has recently led to an increase in the use of Internet broadcasting (webcasting). Most webcasts are archived and accessed numerous times retrospectively. One of the hurdles users face when browsing and skimming through archives is the lack of text transcripts of the audio channel of the webcast archive. In this paper, we proposed a procedure for prototyping an Automatic Speech Recognition (ASR) system that generates realistic transcripts of any desired Word Error Rate (WER), thus overcoming the drawbacks of both prototype-based and Wizard of Oz simulations. We used such a system in a study where human subjects perform question-answering tasks using archives of webcast lectures, and showed that their performance and perception of transcript quality is linearly affected by WER, and that transcripts of WER equal or less than 25% would be acceptable for use in webcast archives.

In order to develop a target speech recognition system with less cost of time and money, reusability of existing speech corpora is becoming one of the most important issues. This paper proposes a new technique of applying a statistical multidimensional scaling method to analyze the reusability of a speech corpus. In the experiment using six speech corpora, which contains isolated words and short sentences used in car navigation system, an effect of the proposed method is evaluated by a usual approach of cross task recognition. Furthermore, the relationship among those speech corpora is clearly shown by the proposed method.

We describe some aspects of the new lexicon we are producing, Com-bilex, whose structure and implementation is specifically designed to reduce redundancy and improve the representation of productive elements of English. Most importantly, many English words are predictable derivations of baseforms, or compounds. Storing the lexicon as a combination of baseforms and derivational rules speeds up lexicon development, and improves coverage and maintainability.

It is indispensable to establish an objective test methodology for noise-reduced speech. This paper proposes a new methodology which estimates word intelligibility of the noise-reduced speech from PESQ MOS (subjective MOS estimated by the PESQ). To evaluate the effectiveness of the proposed methodology, a word intelligibility test of the noise-reduced speech was performed by using four noise reduction algorithms and word lists which take word difficulty into account, and then the word intelligibility was estimated by the proposed methodology. The results confirmed that the word intelligibility can be estimated well from the PESQ MOS without distinguishing the noise reduction algorithms and the noise types.

The increased availability of broadband connections has recently led to an increase in the use of Internet broadcasting (webcasting). Most webcasts are archived and accessed numerous times retrospectively. One of the hurdles users face when browsing and skimming through archives is the lack of text transcripts of the audio channel of the webcast archive. In this paper, we proposed a procedure for prototyping an Automatic Speech Recognition (ASR) system that generates realistic transcripts of any desired Word Error Rate (WER), thus overcoming the drawbacks of both prototype-based and Wizard of Oz simulations. We used such a system in a study where human subjects perform question-answering tasks using archives of webcast lectures, and showed that their performance and perception of transcript quality is linearly affected by WER, and that transcripts of WER equal or less than 25% would be acceptable for use in webcast archives.
distributed recordings in high bandwidth quality can be performed efficiently and reliably. The paper presents the infrastructure developed at BAS for WWW-based speech recordings, it discusses the strategies employed to get schools to participate in the project, and it presents preliminary analyses of the speech database.

A New Single-Ended Measure for Assessment of Speech Quality

Timothy Murphy, Dorel Picovici, Abdulhussain E. Mahdi; University of Limerick, Ireland

Mon1Cap.7

This paper proposes a new non-intrusive measure for objective speech quality assessment in telephony applications and evaluates its performance. The measure is based on estimating perception-based subjective auditory distances between voiced parts of the degraded speech under test and an appropriately formulated artificial reference model of clean speech signals. The reference model is extracted from one or many pre-formulated speech reference books. The reference books are formed by optimally clustering large number of parametric speech vectors extracted from a database of clean speech signals, using an efficient K dimensional tree structure. The measured auditory distances are then mapped into objective listening quality scores. Reported evaluation results show that the proposed measure offers sufficiently accurate and low-complexity assessment method of speech quality, making it suitable for real time applications.

Speech Technology for Minority Languages: The Case of Irish (Gaelic)

Ailbhe Ñ Í Chasaide, John Wogan, Brian Ó Raghaíllaigh, Áine Í Bhriain, Eric Zoerner, Harald Berthelsen, Christer Gobl; Trinity College Dublin, Ireland

Mon1Cap.8

The development of speech technology could play an important role in the maintenance and preservation of minority languages, especially where the population of native speakers are dwindling. This paper outlines the efforts within the WISPR project, to develop annotated spoken corpora along with some of the prerequisites for the synthesis of Irish (Gaelic). It details the particular challenges that have confronted us as well as the strategies adopted to overcome them. It highlights the need for gearing our methodologies to these constraints and to maximise the reusability of resources. Our long-term goal is not only to develop these resources for Irish, but also, in parallel, to develop methodologies that will enable the technology to be flexible and suitable to the envisaged end users, e.g., more flexible kinds of synthesisers, with expressive capabilities and multiple voices, including children’s. It is therefore a major consideration to develop resources in such a way that they are in some sense independent of any single methodology (unit selection vs. other modalities for synthesis development).

Further Investigations on the Relationship Between Objective Measures of Speech Quality and Speech Recognition Rates in Noisy Environments

Francisco José Fraga 1, Carlos Alberto Ynoguti 2, André Godoi Chiovato 2; 1 Federal University of ABC, Brazil; 2 National Institute of Telecommunications, Brazil

Mon1Cap.9

The relationship between an objective measure of speech quality (PESQ) and the recognition rate of a given speech recognition system was already investigated by other researchers. In this paper, we present a further investigation on such a relationship. In our research, the speech recognition tests were performed on a wider class of signals and SNR. The experimental setup as well as the speech recognition systems now evaluated were based on the directions given by the Aurora project. Moreover, a new parametric modeling approach for the PESQ-MOS versus speech recognition rate curve, based on the logistic function, is proposed. This new modeling allows some meaningful interpretations of the parameters of the logistic function in terms of system robustness, and permits to make inferences in the regions outside the experimental measures. Furthermore, the PESQ versus SNR characteristic was used to group types of noise, leading to a much better fit of the logistic function over the data points.

Non-Intrusive Speech Quality Assessment with Low Computational Complexity

Volodya Grancharov, David Y. Zhao, Jonas Lindblom, W. Bastiaan Kleijn; KTH, Sweden

Mon1Cap.10

We describe an algorithm for monitoring subjective speech quality without access to the original signal that has very low computational and memory requirements. The features used in the proposed algorithm can be computed from commonly used speech-coding parameters. Reconstruction and perceptual transformation of the signal are not performed. The algorithm generates quality assessment ratings without explicit distortion modeling. The simulation results indicate that the proposed non-intrusive objective quality measure performs better than the ITU-T P.563 standard despite its very low computational complexity.

Using Speech Recognition Technique for Constructing a Phonetically Transcribed Taiwanese (Min-Nan) Text Corpus

Min-Siong Liang 1, Ren-Yuan Lyu 1, Yuang-Chin Chiang 2; 1 Chang Gung University, Taiwan; 2 National Tsing Hua University, Taiwan

Mon1Cap.11

Collection of Taiwanese text corpus with phonetic transcription suffers from the problems of multiple pronunciation variation. By augmenting the text with speech, and using automatic speech recognition with a sausage searching net constructed from the multiple pronunciations of the text corresponding to its speech utterance, we are able to reduce the effort for phonetic transcription. By using the multiple pronunciation lexicon, the error rate of transcription 13.94% was achieved. Further improvement can be achieved by adapting the pronunciation lexicon with pronunciation variation (PV) rules derived from a manual corrected speech corpus. The PV rules can be categorized into two kinds: the knowledge-based and data-driven rules. By incorporating the PV rules, the error rate reduction 13.63% could be achieved. Although the technique was developed for Taiwanese speech, it could also be adapted easily to be applied in the other similar “minority” Chinese spoken languages.

SloParl — Slovenian Parliamentary Speech and Text Corpus for Large Vocabulary Continuous Speech Recognition

Andrej Žgank, Tomaž Rotovnik, Matej Grašič, Marko Kos, Damjan Vlaj, Zdravko Kačič; University of Maribor, Slovenia

Mon1Cap.12

This paper present a novel Slovenian language resource — SloParl database. It consists from debates acquired in the Slovenian Parliament. The main goal of the project was to cost-effectively collect a new Slovenian language resource that could be used to augment the available Slovenian speech corpora for developing a large vocabulary continuous speech recognition system. The SloParl speech corpus has a total length of 100 hours. The selected sessions between years 2000 – 2005 were incorporated in it. This speech corpus will be used for lightly supervised or unsupervised acoustic models’ training. In accordance with this, the accompanying transcriptions were prepared. The second part of the SloParl database is the text corpus, which covers text of all debates from period 1996 – 2005. It consists of 23M words.
An Annotation Scheme for Agreement Analysis
Siew Leng Toh, Fan Yang, Peter A. Heeman; Oregon Health & Science University, USA
Mon1CaP.13
To accomplish a task that requires collaboration, people would first agree on a strategy and then together carry it out [1]. Our research interest lies in understanding how people explore different strategies and reach an agreement in conversation. We began by examining two-person dialogues in a very limited domain, in which we could just focus on the agreement process. In this paper, we describe an annotation scheme of coding the conversants' behaviors of exploring possible strategies, suggesting and accepting the optimal one, and then maintaining it. We report the inter-coder reliability of the annotation scheme on three expert annotators and two non-experts.

Mon1FoP : Speech Coding
Poster Session, Allegheny Ballroom Foyer, 10:30 – 12:30
Chair: Keiichi Tokuda, Nagoya Institute of Technology, Japan

Signal Modification Incorporating Perceptual Weighting Filter
Joon-Hyuk Chang1, Woohyang Lim2, Nam Soo Kim2; 1Inha University, Korea; 2Seoul National University, Korea
Mon1FoP.1
In this paper, an improved preprocessor for low-bit-rate speech coding employing the perceptual weighting filter is proposed. Speech modification in the proposed approach is performed according to a criterion which makes a compromise between the modification and perceptual weighted quantization errors. For this, the perceptual weighting filter is expressed in terms of a transform domain matrix. The proposed approach is effective in enhancing the speech signal at coder-decoder (CODEC) output through a number of listening tests.

Enhanced Dynamic Codebook Reordering for Advanced Quantizer Structures
Jani Nurminen; Nokia Research Center, Finland
Mon1FoP.2
Conventional dynamic codebook reordering can often significantly enhance the achievable compression efficiency in simple one-stage vector quantization. When applied to more advanced quantizer structures, such as multi-stage vector quantizers, the performance of the technique becomes worse. This paper describes in detail an enhanced approach for dynamic codebook reordering that improves the performance by taking into account the whole quantizer structure. The significant efficiency improvements provided by the proposed approach are demonstrated in practical experiments. Though the results presented in this paper relate to a speech storage system, the proposed approach can also be employed more widely in compression applications that keep the encoder and the decoder in synchrony.

An Efficient Segment-Based Speech Compression Technique for Hand-Held TTS Systems
Chang-Heon Lee1, Sung-Kyo Jung2, Thomas Eriksson2, Won-Suk Jun3, Hong-Goo Kang1; 1Yonsei University, Korea; 2Chalmers University of Technology, Sweden; 3Voiceware Co. Ltd., Korea
Mon1FoP.3
This paper proposes a novel segment-based speech coding algorithm to efficiently compress the database for concatenative text-to-speech (TTS) systems. To achieve a high compression ratio and meet the fundamental requirements of concatenative TTS synthesizers, i.e. partial segment decoding and random access capability, we adopt a modified analysis-by-synthesis scheme. The spectral coefficients are quantized by a length-based interpolation method and excitation signals are modeled with both non-predictive and predictive approaches. Considering that pitch pulse waveforms of a specific speaker show low intra-variation, the conventional adaptive codebook for pitch prediction is replaced by a speaker dependent pitch-pulse codebook. By applying the proposed algorithm to a hand-held Korean TTS system, we verify that the proposed coder provides a compression ratio of about 1/13, a low complexity of around 1.2 WMOPS, and random access capability.

An Unified Unit-Selection Framework for Ultra Low Bit-Rate Speech Coding
Ramasubramanian V.1, Harish D.2; 1Siemens Information Systems Ltd., India; 2International Institute of Information Technology, India
Mon1FoP.4
We propose a unified framework for segment quantization of speech at ultra low bit-rates of 150 bits/sec based on unit-selection principle using a modified one-pass dynamic programming algorithm. The algorithm handles both fixed- and variable- length units in a unified manner, thereby providing a generalization over two existing unit selection methods, which deal with 'single-frame' and 'segmental' units differently. The proposed algorithm performs unit-selection based quantization directly on the units of a continuous codebook, thereby not incurring any of the sub-optimality of the existing 'segmental' algorithm. Moreover, the existing 'single-frame' algorithm becomes a special case of the proposed algorithm. Based on the rate-distortion performance on a multi-speaker database, we show that fixed length units of 6-8 frames perform significantly better than single-frame units and offer similar spectral distortions as variable-length phonetic units, thereby circumventing expensive segmentation and labeling of a continuous database for unit selection based low bit-rate coding.

Efficient VQ Techniques and General Noise Shaping in Noise Feedback Coding
Jes Thyssen, Juin-Hwey Chen; Broadcom Corporation, USA
Mon1FoP.5
This paper describes new efficient Vector Quantization (VQ) techniques that enable low complexity implementations of VQ-based Noise Feedback Coding (NFC). These methods offer mathematical equivalence to higher complexity methods. Furthermore, the paper presents efficient codec structures for general noise shaping as used in BroadVoice 16 — a new SCTE (Society of Cable Telecommunications Engineers) and PacketCable speech coding standard for Cable Telephony.

Classified Comfort Noise Generation for Efficient Voice Transmission
Yasheng Qian, Wei-Shou Hsu, Peter Kabal; McGill University, Canada
Mon1FoP.6
Comfort noise insertion during speech pause has been applied to Voice-over-IP and wireless networks for increasing bandwidth efficiency. We present two classified comfort noise generation (CCNG) schemes using Gaussian Mixture classifiers (GMM-C). Our first scheme employs a classified prototype background noise codebook with the prototype noise waveform chosen using a GMM-C. The second scheme utilizes a classified enhanced excitation codebook. The new CCNG algorithms provide better comfort noise during speech pauses and a smaller classification rate. We have retrofitted the scheme into existing speech transmission system, such as ITU-T G.711/Appendix II and G.723.1/Annex A. The perceived quality of a voice conversation of the novel system has been noticeably enhanced for car and babble.
noise. For the G.711 system, a large improvement is obtained for car noise while the largest amelioration is for babble noise in the G.723.1 system.

Integration of a CELP Coder in the ARDOR Universal Sound Codec

Balázs Kövesi, Dominique Massaloux, David Virette, Julien Bensa; France Télécom, France

Mon1FoP.7

This paper describes the CELP coding module within the Adaptive Rate-Distortion Optimized sound codeR (ARDOR). The ARDOR codec combines coding techniques of different nature using a rate-distortion control mechanism, and is able to adapt to a large range of signal characteristics and system constraints. The implemented CELP codec is derived from the 3GPP AMR-WB codec. Adaptations were necessary to match the ARDOR structure constraints and several new features have been added to improve the codec performance in this context. Listening test results are given to illustrate the behavior of the final codec compared to state-of-the-art coders.

Two Stage Transform Vector Quantization of LSFs for Wideband Speech Coding

Saikat Chatterjee, T.V. Sreenivas; Indian Institute of Science, India

Mon1FoP.8

We investigate the use of a two stage transform vector quantizer (TSTVQ) for coding of line spectral frequency (LSF) parameters in wideband speech coding. The first stage quantizer of TSTVQ, provides better matching of source distribution and the second stage quantizer provides additional coding gain through using an individual cluster specific decorrelating transform and variance normalization. Further coding gain is shown to be achieved by exploiting the slow time-varying nature of line spectra and thus using inter-frame cluster continuity (ICC) property in the first stage of TSTVQ method. The proposed method saves 3–4 bits and reduces the computational complexity by 58–66%, compared to the traditional split vector quantizer (SVQ), but at the expense of 1.5–2.5 times of memory.

Comparison of Prediction Based LSF Quantization Methods Using Split VQ

Saikat Chatterjee, T.V. Sreenivas; Indian Institute of Science, India

Mon1FoP.9

Further improvement in performance, to achieve near transparent quality LSF quantization, is shown to be possible by using a higher order two dimensional (2-D) prediction in the coefficient domain. The prediction is performed in a closed-loop manner so that the LSF reconstruction error is the same as the quantization error of the prediction residual. We show that an optimum 2-D predictor, exploiting both inter-frame and intra-frame correlations, performs better than existing predictive methods. Computationally efficient split vector quantization technique is used to implement the proposed 2-D prediction based method. We show further improvement in performance by using weighted Euclidean distance.

High-Rate Data Embedding in Unvoiced Speech

Konrad Hofbauer 1, Gernot Kubin 2; 1 EEC, France; 2 Graz University of Technology, Austria

Mon1FoP.10

We propose a blind speech watermarking algorithm which allows high-rate embedding of digital side information into speech signals. We exploit the fact that the well-known LPC vocoder works very well for unvoiced speech. Using an autocorrelation based pitch tracking algorithm, a voiced/unvoiced segmentation is carried out. In the unvoiced segments, the linear prediction residual is replaced by a data sequence. This substitution does not cause perceptual degradation as long as the residual's power is matched. The signal is resynthesized using the unmodified LPC filter coefficients. The watermark is decoded by a linear prediction analysis of the received signal and the information is extracted from the sign of the residual. The watermark is nearly imperceptible and provides a channel capacity of up to 2000 bit/s in an 8 kHz-sampled speech signal.

Pitch Resynchronization While Recovering from a Late Frame in a Predictive Speech Decoder

Kyle D. Anderson, Philippe Gournay; VoiceAge Corporation, Canada

Mon1FoP.11

The concealment procedure used by CELP speech decoders to regenerate lost frames introduces an error that propagates into the following frames. Within the context of voice transmission over packet networks, some packets arrive too late to be decoded and must also be concealed. Once they arrive however, those packets can be used to update the internal state of the decoder, which stops error propagation. Yet, care must be taken to ensure a smooth transition between the concealed frame and the following “updated” frame computed with properly updated internal states. During voiced or quasi-periodic segments, the pitch phase error that is generally introduced by the concealment procedure makes it difficult and detrimental to quality to use the traditional fade-in, fade-out approach. This paper presents a method to handle that pitch phase error. Specifically, the transition is done in such a way that the natural pitch periodicity of the speech signal is not broken.
Mon2A1O : Speech Enhancement I

Oral Session, Allegheny Ballroom, left, 14:00 – 16:00
Chair: Mark A. Clements, Georgia Institute of Technology, USA

A Novel Environment-Dependent Speech Enhancement Method with Optimized Memory Footprint

Suhadi Suhadi¹, Sorel Stan¹, Tim Fingscheidt²; ¹BenQ Mobile, Germany; ²Braunschweig Technical University, Germany

Mon2A1O.1

Data-driven speech enhancement (Fingscheidt and Suhadi [1]) aims at improving speech quality for voice calls in a specific noise environment. The essence of the method are a set of frequency-dependent weighting rules, indexed by a priori and a posteriori SNRs, which are learned from clean speech and background noise training data. The weighting rules must be stored for each frequency bin separately and take up about 400 kBytes memory, which makes DSP implementations relatively expensive.

In this paper we propose an alternative definition of the weighting rules which requires only 27 kBytes memory. That is 6.7% of the memory consumption of the original algorithm, with virtually no loss in performance measured in terms of speech distortion and noise attenuation. Our approach is to redefine the weighting rules on the Bark scale and store their parametric representation obtained by polynomial curve fitting.

Weighted Codebook Mapping for Noisy Speech Enhancement Using Harmonic-Noise Model

Esfandiar Zavarehei, Saeed Vaseghi, Qin Yan; Brunel University, UK

Mon2A1O.2

Most noisy speech enhancement methods result in partial suppression and distortion of speech spectrum. At instances when the local signal-to-noise ratio at a frequency band is very low speech partials are often obliterated. In this paper a method for enhancement and restoration of noisy speech based on a harmonic-noise model (HNM) is introduced. A HNM imposes a temporal-spectral structure that may reduce processing artifacts. The restoration process is enhanced through incorporation of a priori HNM of clean speech stored in a pre-trained codebook. The restored speech is a SNR-dependent combination of the de-noised observation and the speech obtained from weighted codebook mapping. The additional improvements of speech quality resulting from the proposed method in comparison to conventional and modern speech enhancement systems are evaluated. The results show that the proposed method improves the quality of noisy speech and restores much of the information lost to noise.

MMSE Estimation of Complex-Valued Discrete Fourier Coefficients with Generalized Gamma Priors

J. Jensen, R.C. Hendriks, J.S. Erkelens, R. Heusdens; Delft University of Technology, The Netherlands

Mon2A1O.3

We consider DFT based techniques for single-channel speech enhancement. Specifically, we derive minimum mean-square error estimators of clean speech DFT coefficients based on generalized gamma prior probability density functions. Our estimators contain as special cases the well-known Wiener estimator and the more recently derived estimators based on Laplacian and two-sided gamma priors. Simulation experiments with speech signals degraded by various additive noise sources verify that the estimator based on the two-sided gamma prior is close to optimal amongst all the estimators considered in this paper.

Automatic Removal of Typed Keystrokes from Speech Signals

Amarnag Subramanya¹, Michael L. Seltzer², Alex Acero²; ¹University of Washington, USA; ²Microsoft Research, USA

Mon2A1O.4

Laptop computers are increasingly being used as recording devices to capture meetings, interviews, and lectures using the laptop’s local microphone. In these scenarios, the user frequently also uses the same laptop to make notes. Because of the close proximity of the laptop’s microphone to its keyboard, the captured speech signal is significantly corrupted by the impulsive sounds the user’s keystrokes generate. In this paper we propose an algorithm to automatically detect and remove keystrokes from a recorded speech signal. The detection and removal stages both operate by exploiting the natural correlations present in speech signals, but do so in different ways. The proposed algorithm is computationally efficient, requires no user-specific training or enrolment, and results in significantly enhanced speech. The proposed keystroke removal algorithm was evaluated through user listening tests and speech recognition experiments on speech recordings made in a realistic environment.

Lattice LP Filtering for Noise Reduction in Speech Signals

Erhard Rank, Gernot Kubin; Graz University of Technology, Austria

Mon2A1O.5

We present a simple yet effective algorithm for noise reduction of speech signals using a lattice LP filter. Based on previous investigations and a theoretical analysis of the lattice filter parameter estimation we introduce an improved parameter estimation algorithm that takes into account the non-stationary nature of speech and expected noise signals, yielding a good suppression of stationary and slowly time-varying noise. The algorithm has zero delay for the speech signal, promoting its application for telephony or hearing aids. No additional or explicit noise estimation algorithm is needed.

Speech Enhancement Using Modified Phase Opponency Model

Om D. Deshmukh, Carol Y. Espy-Wilson; University of Maryland, USA

Mon2A1O.6

We previously presented a single-channel speech enhancement technique called the Modified Phase Opponency (MPO) model. The MPO model is based on a neural model called the Phase Opponency (PO) model. The efficacy of the MPO model was demonstrated on speech signals corrupted by additive white noise. In the present work, we extend the MPO model to perform efficiently on speech signals corrupted by additive colored noise with time varying spectral characteristics and amplitude levels. The MPO enhancement scheme outperforms many of the statistical and signal-theoretic speech enhancement techniques when evaluated using three different objective quality measures on a subset of the Aurora database. The superiority of the MPO speech enhancement scheme in enhancing speech signals when the amplitude level and the spectral characteristics of the background noise are fluctuating is also demonstrated.
Mon2A2O : ASR Other I
Oral Session, Allegheny Ballroom, center, 14:00 – 16:00
Chair: Qiang Huo, University of Hong Kong, China

Computer-Assisted Closed-Captioning of Live TV Broadcasts in French
G. Boulianne, J.-F. Beaumont, M. Boisvert, J. Brousseau, P. Cardinal, C. Chapdelaine, M. Comeau, Pierre Ouellet, F. Osterrath; Centre de Recherche Informatique de Montréal, Canada
Mon2A2O.1
Growing needs for French closed-captioning of live TV broadcasts in Canada cannot be met only with stenography-based technology because of a chronic shortage of skilled stenographers. Using speech recognition for live closed-captioning, however, requires several specific problems to be solved, such as the need for low-latency real-time recognition, remote operation, automated model updates, and collaborative work. In this paper we describe our solutions to these problems and the implementation of a live captioning system based on the CRIM speech recognizer. We report results from field deployment in several projects. The oldest in operation has been broadcasting real-time closed-captions for more than 2 years.

On the Use of Morphological Analysis for Dialectal Arabic Speech Recognition
Mohamed Affiy, Ruhi Sarikaya, Hong-Kwang Jeff Kuo, Laurent Besacier, Yuqing Gao; IBM T.J. Watson Research Center, USA
Mon2A2O.2
Arabic has a large number of affixes that can modify a stem to form words. In automatic speech recognition (ASR) this leads to a high out-of-vocabulary (OOV) rate for typical lexicon size, and hence a potential increase in WER. This is even more pronounced for dialects of Arabic where additional affixes are often introduced and the available data is typically sparse. To address this problem we introduce a simple word decomposition algorithm which only requires a text corpus and a predefined list of affixes. Using this algorithm to create the lexicon for Iraqi Arabic ASR results in about 10% relative improvement in word error rate (WER). Also using the union of the segmented and unsegmented vocabularies and interpolating the corresponding language models results in further WER reduction. The net WER improvement is about 13% relative.

Recognition of Classroom Lectures in European Portuguese
Isabel Trancoso¹, Ricardo Nunes², Luís Neves², Céu Viana³, Helena Moniz², Diamantino Caseiro¹, Ana Isabel Mata²;¹L2F INESC-ID/IST, Portugal; ²L2F INESC-ID, Portugal; ³CILUL, Portugal
Mon2A2O.3
Classroom lectures may be very challenging for automatic speech recognizers, because the vocabulary may be very specific and the speaking style very spontaneous. Our first experiments using a recognizer trained for Broadcast News resulted in word error rates near 60%, clearly confirming the need for adaptation to the specific topic of the lectures, on one hand, and for better strategies for handling spontaneous speech. This paper describes our efforts in these two directions: the different domain adaptation steps that lowered the error rate to 45%, with very little transcribed adaptation material, and the exploratory study of spontaneous speech phenomena in European Portuguese, namely concerning filled pauses.

Investigating Automatic Decomposition for ASR in Less Represented Languages
Thomas Pellegrini, Lori Lamel; LIMSI, France
Mon2A2O.4
This paper addresses the use of an automatic decomposition method to reduce lexical variety and thereby improve speech recognition of less well-represented languages. The Amharic language has been selected for these experiments since only a small quantity of resources are available compared to well-covered languages. Inspired by the Harris algorithm, the method automatically generates plausible affixes, that combined with decomposing can reduce the size of the lexicon and the OOV rate. Recognition experiments are carried out for four different configurations (full-word and decompounded) and using supervised training with a corpus containing only two hours of manually transcribed data.

Automatic Transcription of Somali Language
Abdillahi Nimaan, Pascal Nocéra, Jean-François Bonastre; LIA, France
Mon2A2O.5
Most African countries follow an oral tradition system to transmit their cultural, scientific and historic heritage through generations. This ancestral knowledge accumulated during centuries is today threatened of disappearing. Automatic transcription and indexing tools seem potential solution to preserve it. This paper presents the first steps of automatic speech recognition (ASR) of Djibouti languages in order to index the Djibouti cultural heritage. This work is dedicated to process Somali language, which represents half of the targeted Djiboutian audio archives. We describe the principal characteristics of audio (10 hours) and textual (3M words) training corpora collected and the first ASR results of this language. Using the specificities of the Somali language, (words are composed of a concatenation of sub-words called “roots”) in this paper, we improve the obtained results. We also discuss future ways of research like roots indexing of audio archives.

Analysis of Overlaps in Meetings by Dialog Factors, Hot Spots, and Collection Site: Insights for Automatic Speech Recognition
Özgür Çetin¹, Elizabeth Shriberg²;¹International Computer Science Institute, USA; ²SRI International, USA
Mon2A2O.6
In previous work we found that automatic speech recognition (ASR) results on meetings show interesting patterns with respect to speaker overlaps, including a robust asymmetry in word error rates (WERs) before and after overlaps. The paradigm used allowed us to infer that these correlations are not due to crosstalk itself but to changes in how a person speaks around overlap regions. To better understand these ASR and perplexity results, we analyze speaker overlaps with respect to various factors, including collection site, speakers, dialog acts, and hot spots.

We examine a total of 101 meetings from the ICSI meeting corpus and the NIST meeting transcription evaluations of the last four years. We find that overlaps tend to occur at high-perplexity regions in the foreground talker’s speech. We also find that overlap regions tend to have higher perplexity than those in nonoverlaps, if trigrams or 4-grams are used, but unigram perplexity within overlaps is considerably lower than that of nonoverlaps. These appear to be robust findings, because they hold in general across meetings from different collection sites, even though meeting style and absolute rates of overlap vary by site. Further analyses of overlap with respect to speakers and meeting content reveal interesting relationships between overlap and dialog acts, as well as between overlap and “hot spots” (points of increased participant involvement). Finally, results from the ICSI meeting corpus show that individual speakers have widely varying rates of being overlapped.
Mon2A3O : Modeling Prosodic Features
Oral Session, Allegheny Ballroom, right, 14:00 – 16:00
Chair: Yoshinori Sagisaka, NICT, Japan

Combining Acoustic, Lexical, and Syntactic Evidence for Automatic Unsupervised Prosody Labeling
Sankaranarayanan AnanthaKrishnan, Shrikanth Narayanan; University of Southern California, USA
Mon2A3O.1

Automatic labeling of prosodic events in speech has potentially significant implications for spoken language processing applications, and has received much attention over the years, especially after the introduction of annotation standards such as ToBI. Current labeling techniques are based on supervised learning, relying on the availability of a corpus that is annotated with the prosodic labels of interest in order to train the system. However, creating such resources is an expensive and time-consuming task. In this paper, we examine an unsupervised labeling algorithm for accent (prominence) and prosodic phrase boundary detection at the linguistic syllable level, and evaluate their performance on an standard, manually annotated corpus. We obtain labeling accuracies of 77.8% and 88.5% for the accent and boundary labeling tasks, respectively. These figures compare well against previously reported performance levels for supervised labelers.

On the Correlation Between Energy and Pitch Accent in Read English Speech
Andrew Rosenberg, Julia Hirschberg; Columbia University, USA
Mon2A3O.2

In this paper, we describe a set of experiments that examine the correlation between energy and pitch accent. We tested the discriminative power of the energy component of frequency subbands with a variety of frequencies and bandwidths on read speech spoken by four native speakers of Standard American English, using an analysis by classification approach. We found that the frequency region most robust to speaker differences is between 2 and 20 bark. Across all speakers, using only energy features we were able to predict pitch accent in read speech with accuracy of 81.9%.

Corpus-Based Generation of Fundamental Frequency Contours Using Generation Process Model and Considering Emotional Focuses
Keikichi Hirose, Yasufumi Asano, Nobuaki Minematsu; University of Tokyo, Japan
Mon2A3O.3

We formerly conducted emotional speech synthesis using our corpus-based method of generating fundamental frequency ($F_0$) contours from text. The method predicts command values of $F_0$ contour generation process model instead of directly predicting $F_0$ value of each time frame. A better control of $F_0$ contours was realized by taking the emotional level of each bunsetsu into account: adding information on which bunsetsu(s) the emotion is especially placed to the command predictor inputs. In the case of anger, $F_0$ contours closer to the target contours are obtained by adding emotional levels. Speech synthesis was conducted by generating $F_0$ contours in two ways: using commands predicted by taking emotional levels into account and those not. The result of perceptual experiment indicated that emotion was conveyed well by adding emotional levels.

Mon2A3O.4

Prosodic Boundaries in Czech: An Experiment Based on Delexicalized Speech
Tomáš Duběda; Charles University in Prague, Czech Republic

The present experiment attempts to determine the role of prosody in the identification of stress unit boundaries in Czech, using three types of delexicalized stimuli evaluated by native speakers. After describing the relative contribution of intonation and duration to boundary perception, an analysis of misplaced boundaries is provided. Identified patterns concern especially the relationship between tonal structure and boundary salience, the order of preference between intonation and duration, and the tendency towards perceptual filling of accent lapses.

Totally Data-Driven Intonation Prediction Model Using a Novel F0 Contour Parametric Representation
Lifu Yi, Jian Li, Xiaoyan Lou, Jie Hao; Toshiba China R&D Center, China
Mon2A3O.5

This paper proposes a novel parametric representation of mandarin intonation based on orthogonal polynomial approximation. The polynomial is a simplified representation of Parallel Encoding and Target Approximation (PENTA) intonation model that includes a target component and an approximation component. We also propose predicting the polynomial parameters from linguistic and phonetic attributes by generalized linear models (GLM). The optimal attributes are automatically selected by stepwise regression method. Thus both model structures and model coefficients are optimized in a totally data-driven manner. In addition, speaking rate is introduced as a new attribute for prediction. When the method is applied to intonation prediction of Mandarin speech, it achieves F0 RMSE of 30.21 Hz and correlation coefficients of 0.85 in open test. Informal perceptual experiments show that the predicted intonation is quite appropriate and natural.

A Comparison of Inter-Transcriber Reliability for Two Systems of Prosodic Annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices)
Laura Dilley1, Mara Breen2, Marti Bolivar2, John Kraemer2, Edward Gibson2; 1Ohio State University, USA; 2Massachusetts Institute of Technology, USA
Mon2A3O.6

Agreement was investigated among five labelers for the use of two prosodic annotation systems: the ToBI (Tones and Break Indices) system [1,2] and the RaP (Rhythm and Pitch) system [3]. Each system permits the labeling of pitch accents and two levels of phrasal boundaries; RaP also permits labeling of speech rhythm and distinguishes multiple levels of prominence on syllables. After training with computerized materials and getting expert feedback, coders applied each system to a corpus of read and spontaneous speech (36 minutes for ToBI and 19 for RaP). Inter-coder reliability was computed according to two metrics: transcriber-syllable-pairs and the kappa statistic. High agreement was obtained for both systems for pitch accent presence, pitch accent type, boundary presence, boundary type, and, for RaP, presence and strength of metrical prominences. Agreement levels for ToBI were similar to those of previous studies [4,5], indicating that participants were proficient coders. Moreover, the high level of agreement demonstrated for the RaP system indicates that RaP is a viable alternative to ToBI for prosodic labeling of large speech corpora.
Saliency Parsing for Automated Directory Assistance

Issac Alphonso, Shuangyu Chang; Tellme Networks Inc., USA

Mon2WeO.1
In a statistical language model based automated directory assistance system, extracting the salient information from the recognition output can significantly increase the accuracy of the backend listing database search. In this paper, we describe a Hidden Markov model (HMM) based saliency parser that was developed to accurately and efficiently identify salient words from the recognition output by modeling both the syntactic structure as well as the lexical distribution. The parser can be trained using a relatively small data set with coarse syntactic class labels, without the need for detailed syntactic knowledge or a treebank-like corpus. Experimental results on a research corpus of directory assistance utterances betoken the parser’s importance within the automated system. The results demonstrate that the proposed saliency parser can significantly improve the overall automation rate without increasing the error rate.

Open-Vocabulary Spoken Document Retrieval Based on New Subword Models and Subword Phonetic Similarity

Kohei Iwata1, Yoshiaki Itoh1, Kazunori Kajima1, Masaaki Ishigame1, Kazuyo Tanaka2, Shi-wook Lee3; 1Iwate Prefectural University, Japan; 2University of Tsukuba, Japan; 3AIST, Japan

Mon2WeO.2
A new type of video retrieval system is proposed that identifies a target video section by searching for a word passage submitted as a quoted speech or text query. The proposed system has two unique characteristics. The first characteristic is that it is based on subword models such as phonemes, syllables, and morphemes so the system is able to deal with any type of query, including new words and personal names. The second characteristic is that the system relies on acoustic similarity between subword models. Furthermore, new subword models were constructed for the retrieval system to improve performance. The new models were based on two concepts: context-dependent models and more sophisticated in the time axis than phone models. Through experimentation, the effectiveness and scope of the proposed spoken document retrieval system were confirmed, and suitable subword models for the proposed method discussed.

Improved Topic Classification Over Maximum Entropy Model Using K-Norm Based New Objectives

Xiang Li, Ea-Ee Jan, Cheng Wu, David Lubensky; IBM T.J. Watson Research Center, USA

Mon2WeO.3
Maximum Entropy (MaxEnt) model has been proven to be a very effective approach in the topic classification task, where a specific topic from a pre-defined topic set will be assigned to each sentence. Although it is originally developed based on the maximization of the conditional probability entropy under certain constraints, MaxEnt model is indeed an exponential distribution model that maximizes the log-likelihood of the training data. This log-likelihood criterion bears similarity with the classification accuracy criterion, which is the ultimate performance measure of a topic classifier. But these two criterion still differ from each other, and their discrepancy consequently reduces the benefit of optimization in improving classification accuracy. In this paper we propose to use different objective functions, which are closer to the classification accuracy criterion, to replace the log-likelihood objective used in the MaxEnt model estimation process. Specifically, we propose a **Summation-Log K-norm objective** and a **Summation K-norm objective**. Our experiments conducted on two large volume topic classification dataset prove the effectiveness of our new objectives in improving topic classification performance on top of the state-of-art MaxEnt model.

Efficient Interactive Retrieval of Spoken Documents with Key Terms Ranked by Reinforcement Learning

Yi-cheng Pan, Jia-yu Chen, Yen-shin Lee, Yi-sheng Fu, Lin-shan Lee; National Taiwan University, Taiwan

Mon2WeO.4
Unlike written documents, spoken documents are difficult to display on the screen; it is also difficult for users to browse these documents during retrieval. It has been proposed recently to use interactive multi-modal dialogues to help the user navigate through a spoken document archive to retrieve the desired documents. This interaction is based on a topic hierarchy constructed by the key terms extracted from the retrieved spoken documents. In this paper, the efficiency of the user interaction in such a system is further improved by a key term ranking algorithm using Reinforcement Learning with simulated users. Significant improvements in retrieval efficiency, which are relatively robust to the speech recognition errors, are observed in preliminary evaluations.

Discriminative Named Entity Recognition of Speech Data Using Speech Recognition Confidence

Katsuhito Sudoh, Hajime Tsukada, Hideki Isozaki; NTT Corporation, Japan

Mon2WeO.5
This paper presents a method for the named entity recognition (NER) of speech data that uses automatic speech recognition (ASR) confidence as a feature that indicates whether each word is correctly recognized. An NER model is trained using ASR results with named entity (NE) labels to include an ASR confidence feature as well as corresponding transcriptions with NE labels. Experiments using support vector machines (SVMs) and speech data from Japanese newspaper articles show that the proposed method achieves higher F-measure in NER than a simple application of text-based NER to ASR results.

Using Latent Semantic Indexing for Morph-Based Spoken Document Retrieval

Ville T. Turunen, Mikko Kurimo; Helsinki University of Technology, Finland

Mon2WeO.6
Previously, phone-based and word-based approaches have been used for spoken document retrieval. The former suffers from high error rates and the latter from limited vocabulary of the recognizer. Our method relies on unlimited vocabulary continuous speech recognizer that uses morpheme-like units discovered in an unsupervised manner. The morpheme-like units, or “morphs” for short, have been successfully used also as index terms. One problem using morphs as index terms is that the segmentation does not always separate the same stem for different inflected forms of the same word. This resembles the problem of synonyms. In this paper, we apply latent semantic indexing to morph based retrieval. The idea is to project morphs that correspond to the same word, as well as other semantically related terms, to the same dimension. The results show clear improvements in Finnish spoken document retrieval performance.
Feature Combination Using Linear Discriminant Analysis and Its Pitfalls
Ralf Schlüter, András Zolnay, Hermann Ney; RWTH Aachen University, Germany
Mon2BuP.1
In this paper, Linear Discriminant Analysis (LDA) is investigated with respect to the combination of different acoustic features for automatic speech recognition. It is shown that the combination of acoustic features using LDA does not consistently lead to improvements in word error rate. A detailed analysis of the recognition results on the Verbmobil (VM) and on the English portion of the European Parliament Plenary Sessions (EPIS) corpus is given. This includes an independent analysis of the effect of the dimension of the input to LDA, the effect of strongly correlated input features, as well as a detailed numerical analysis of the generalized eigenvalue problem underlying LDA. Relative improvements in word error rate of up to 5% were observed for LDA-based combination of multiple acoustic features.

Discriminant Linear Processing of Time-Frequency Plane
Fabio Valente, Hynek Hermansky; IDIAP Research Institute, Switzerland
Mon2BuP.2
Extending previous works done on considerably smaller data sets, the paper studies linear discriminant analysis of about 30 hours of phoneme-labeled speech data in the time-frequency domain. Analysis is carried both independently in time and frequency and jointly. Data driven spectral basis show similar frequency sensitivity as human hearing. LDA-derived temporal FIR filters are consistent with temporal lateral inhibition. Considerable improvement is obtained using first temporal discriminant.

Automatic Speech Recognition Experiments with Articulatory Data
Esmeralda Uraga, Thomas Hain; University of Sheffield, UK
Mon2BuP.3
In this paper we investigate the use of articulatory data for speech recognition. Recordings of the articulatory movements originate from the MOCHA corpus, a database which contains speech, EGG, EMA and EPG recordings. It was found that in a Hidden Markov Model (HMM) based recognition framework careful processing of these signals can yield significantly better performance than that obtained by decoding of the acoustic signals. We present detailed results on the processing of the signals and the associated performance of monophone and triphone systems. Experimental evidence shows that acoustic-signal-to-word mappings and articulatory-signal-to-word mappings are equally complex. However, for the latter, evidence of short-comings of standard HMM based modelling is visible and should be addressed in future systems.

Speech Recognition with Phonological Features: Some Issues to Attend
Frederik Stouten, Jean-Pierre Martens; Ghent University, Belgium
Mon2BuP.4
It is often argued that acoustic-phonetic or articulatory features could be beneficial to automatic speech recognition because they provide a convenient interface between the acoustic and the linguistic level. Former research has shown that a combination of acoustic and articulatory information can lead to improved ASR. However there exists no purely articulatory driven ASR system that outperforms state-of-the-art systems driven by acoustic features. In this paper we propose a novel method for improving ASR on the basis of articulatory features. It is designed to take account of (1) the correlations between articulatory features and (2) the fact that not all articulatory features are relevant for the description of a certain phonetic unit. We also investigate to what extend an acoustic and an articulatory feature driven system make different errors.

Multi-Source Far-Distance Microphone Selection and Combination for Automatic Transcription of Lectures
Matthias Wölfel, Christian Fügner, Shahjith Ikbal, John W. McDonough; Universität Karlsruhe, Germany
Mon2BuP.5
In this work, we present our progress in multi-source far field automatic speech-to-text transcription for lecture speech. In particular, we show how the best of several far field channels can be selected based on a signal-to-noise ratio criterion, and how the signals from multiple channels can be combined at either the waveform level using blind channel combination or at the hypothesis level using confusion network techniques to improve the accuracy of a far field lecture transcription system. Using the techniques described here, we ran a series of experiments on the test set used by the US National Institute of Standards and Technologies for the RT-05S evaluation. For the multiple distant microphones (MDM) task of RT-05S, our system achieved a word error rate of 38.3% which represents an improvement of over 13% absolute compared to the best reported results in the RT-05S evaluation.

Statistical Analysis and Performance of DFT Domain Noise Reduction Filters for Robust Speech Recognition
Colin Breithaupt, Rainer Martin; Ruhr-Universität Bochum, Germany
Mon2BuP.6
Noise reduction frontends have been developed independently for speech communication and speech recognition purposes with the result that one and the same algorithm does not perform well in both application domains. In this paper we show that noise reduction filters based on the discrete Fourier transform (DFT) which are used in speech communication can also perform well in robust automatic speech recognition (ASR) experiments if some form of feature smoothing is applied.

We analyse the statistics of the Mel frequency cepstral coefficients (MFCCs) that are used as speech features and describe the effects on recognition results if the mean and variance of these features change. It is shown that recognizers are more sensitive to an increase in variance of enhanced features than to errors in their mean values. We present a method that compensates for the increased variance of DFT-based noise reduction frontends by means of using prior knowledge and smoothing. We achieve high segmental SNR improvements as well as recognition results close to those of the Advanced Frontend (AFE) of the European Telecommunications Standards Institute (ETSI) for all noise types.

Normalization of the Inter-Frame Information Using Smoothing Filtering
L. García, José C. Segura, Carmen Benítez, Javier Ramírez, Ángel de la Torre; Universidad de Granada, Spain
Mon2BuP.7
A filter that introduces inter-frame information into the voice features set is proposed in this paper. The filter adds the autocorrelations of the cepstral coefficients to the set of characteristics used for training and recognition. Those autocorrelations should not depend on the
Optimization of Class Weights for LDA Feature Transformations
Andrej Ljolje; AT&T Labs Research, USA

Mon2BuP.11

One popular feature type in speech recognition is based on linear transformations of sequences of cepstral feature vectors. In general the transformation is generated in two steps: first a transformation like linear discriminant analysis (LDA) or heteroscedastic linear discriminant analysis (HLDA) is used to maximize separation between classes and reduce the dimensionality, followed by a decorrelating transformation. Here we investigate the weighting of classes when using the LDA transformation. In particular we are concerned with the special status of silence, for which the data can be arbitrarily long, and which can be represented by more than one silence/noise model. The special case of our acoustic models for commercial applications, which consist of several sub-models for each type of application, like general English, digits, names, alphabet, etc., creates a conflict when using a transformation like LDA to improve the separability of states which correspond to the same phoneme, but used within a different type of task. We also evaluate replacing sample counts with error/accuracy counts and cross-task LDA transformation estimation. The results show that it is important to take these conditions into account and demonstrate accuracy/speed improvements when appropriate care is taken in computing the LDA transformations.

LDA Based Feature Estimation Methods for LVCSR
Janne Pylkkönen; Helsinki University of Technology, Finland

Mon2BuP.12

Features that model temporal aspects of phonemes are important in speech recognition. One method is to use linear discriminant analysis (LDA) to find discriminative features from a spectro-temporal input formed by concatenating consecutive frames of short-time spectrum features. Others use e.g. neural networks to process longer span spectral segments to improve recognition accuracy. Still the most widely used method for including temporal cues is to augment the short-time spectral features with simple time derivatives.

In this paper a new feature estimation method based on pairwise linear discriminants is presented. We compare it and some of its variants to traditional MFCC features and to LDA estimated features in a large vocabulary continuous speech recognition (LVCSR) task. The features obtained with the new estimation method show significant improvements in recognition accuracy over MFCC and LDA features.

Robust Feature Extraction Based on Spectral Peaks of Group Delay and Autocorrelation Function and Phase Domain Analysis
G. Farahani, S.M. Ahadi, M. Mehdi Homayounpour; Amirkabir University of Technology, Iran

Mon2BuP.13

This paper presents a new robust feature set for noisy speech recognition in phase domain along with spectral peaks obtained from group delay and autocorrelation functions. The group delay domain is appropriate for formant tracking and autocorrelation domain is well-known for its pole preserving and noise separation properties. In this paper, we report on appending spectral peaks obtained in either group delay or autocorrelation domains to the feature vectors extracted originally in phase domain to create a new feature set.

We tested our features on the Aurora 2 noisy isolated-word task and found that it led to improvements over other group delay-based and autocorrelation-based methods that use magnitude instead of phase for feature extraction.

Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition
Hynek Bořil, Petr Fousek, Petr Pollák; Czech Technical University in Prague, Czech Republic

Mon2BuP.10

Adverse environment conditions. Because they should only depend on the information to recognize, a normalization of that inter-frame information is convenient. The filter defined implements this normalization by transforming the autocorrelations into a normalized domain defined with clean adaptation data. This temporal processing of the features is added to the Histogram Equalization of the cepstral coefficients (HEQ) used to normalize the MFCCs. An analysis is done about the most effective domain (original MFCCs or equalized MFCCs) on which the temporal processing should be executed. Performance results for the proposed algorithm are presented for AURORA2 and AURORA4 databases.

Comparative Study on Contributions of Pitch-Synchronization and Peak-Amplitude Towards Robustness Issue of ASR
Mahammad Ghulam, Junsei Horikawa, Tsuneo Nitta; Toyohashi University of Technology, Japan

Mon2BuP.8

We proposed previously a novel pitch-synchronous peak-amplitude (PS-PA) based feature extraction method, which achieved significant recognition accuracy for robust ASR [1]. It is well-known that an auditory neuron has pitch detection mechanism that can be useful for speech detection, and also peak-amplitudes in temporal pattern are robust to noise. In this paper, we conduct several experiments to find out relative contributions of pitch-synchronization (PS) and peak-amplitudes (PA) on recognition accuracy of robust ASR. Experiments include methods with fixed and pitch-synchronous frame lengths, and that with traditional peak-amplitudes and pitch-synchronous peak-amplitudes. The experimental results show that both PS and PA have strong contributions towards robust ASR and the effect of PS is higher than that of PA.

Phoneme Recognition Based on Fisher Weight Map to Higher-Order Local Auto-Correlation
Yasuo Ariki, Shunsuke Kato, Tetsuya Takiguchi; Kobe University, Japan

Mon2BuP.9

In this paper, we propose a new feature extraction method based on higher-order local auto-correlation (HLAC) and Fisher weight map (FWM). Widely used MFCC features lack temporal dynamics. To solve this problem, 35 types of local auto-correlation features are computed in two-dimensional local regions. These local features are accumulated over more global regions by weighting high scores on the discriminative areas where the typical features among all phonemes are well expressed. This score map is called Fisher weight map. We verified the effectiveness of the HLAC and FWM through vowel recognition and total phoneme recognition.

Transformations
Optimization of Class Weights for LDA Feature Transformations
Andrej Ljolje; AT&T Labs Research, USA

Mon2BuP.11

One popular feature type in speech recognition is based on linear transformations of sequences of cepstral feature vectors. In general the transformation is generated in two steps: first a transformation like linear discriminant analysis (LDA) or heteroscedastic linear discriminant analysis (HLDA) is used to maximize separation between classes and reduce the dimensionality, followed by a decorrelating transformation. Here we investigate the weighting of classes when using the LDA transformation. In particular we are concerned with the special status of silence, for which the data can be arbitrarily long, and which can be represented by more than one silence/noise model. The special case of our acoustic models for commercial applications, which consist of several sub-models for each type of application, like general English, digits, names, alphabet, etc., creates a conflict when using a transformation like LDA to improve the separability of states which correspond to the same phoneme, but used within a different type of task. We also evaluate replacing sample counts with error/accuracy counts and cross-task LDA transformation estimation. The results show that it is important to take these conditions into account and demonstrate accuracy/speed improvements when appropriate care is taken in computing the LDA transformations.

LDA Based Feature Estimation Methods for LVCSR
Janne Pylkkönen; Helsinki University of Technology, Finland

Mon2BuP.12

Features that model temporal aspects of phonemes are important in speech recognition. One method is to use linear discriminant analysis (LDA) to find discriminative features from a spectro-temporal input formed by concatenating consecutive frames of short-time spectrum features. Others use e.g. neural networks to process longer span spectral segments to improve recognition accuracy. Still the most widely used method for including temporal cues is to augment the short-time spectral features with simple time derivatives.

In this paper a new feature estimation method based on pairwise linear discriminants is presented. We compare it and some of its variants to traditional MFCC features and to LDA estimated features in a large vocabulary continuous speech recognition (LVCSR) task. The features obtained with the new estimation method show significant improvements in recognition accuracy over MFCC and LDA features.

Robust Feature Extraction Based on Spectral Peaks of Group Delay and Autocorrelation Function and Phase Domain Analysis
G. Farahani, S.M. Ahadi, M. Mehdi Homayounpour; Amirkabir University of Technology, Iran

Mon2BuP.13

This paper presents a new robust feature set for noisy speech recognition in phase domain along with spectral peaks obtained from group delay and autocorrelation functions. The group delay domain is appropriate for formant tracking and autocorrelation domain is well-known for its pole preserving and noise separation properties. In this paper, we report on appending spectral peaks obtained in either group delay or autocorrelation domains to the feature vectors extracted originally in phase domain to create a new feature set.

We tested our features on the Aurora 2 noisy isolated-word task and found that it led to improvements over other group delay-based and autocorrelation-based methods that use magnitude instead of phase for feature extraction.

Data-Driven Design of Front-End Filter Bank for Lombard Speech Recognition
Hynek Bořil, Petr Fousek, Petr Pollák; Czech Technical University in Prague, Czech Republic

Mon2BuP.10

Adverse environment conditions. Because they should only depend on the information to recognize, a normalization of that inter-frame information is convenient. The filter defined implements this normalization by transforming the autocorrelations into a normalized domain defined with clean adaptation data. This temporal processing of the features is added to the Histogram Equalization of the cepstral coefficients (HEQ) used to normalize the MFCCs. An analysis is done about the most effective domain (original MFCCs or equalized MFCCs) on which the temporal processing should be executed. Performance results for the proposed algorithm are presented for AURORA2 and AURORA4 databases.
Frequency Warping by Linear Transformation of Standard MFCC

Sankaran Panchapagesan; University of California at Los Angeles, USA
Mon2Cap.14
A novel linear transform (LT) is proposed for frequency warping (FW) with standard filterbank based MFCC features. Here, we use the idea of spectral interpolation of [9] to perform a continuous warping in the log filterbank output domain, and incorporate both interpolation and warping into a single warped IDCT matrix. The new transformation matrix is thus mathematically simpler than in [9], and no modification of standard MFCC feature extraction is required like the previous approach. In VTLN experiments with maximum likelihood score (MLS) estimation of the FW parameter, the new LT outperformed regular VTLN implemented by warping the Mel filterbank. In speaker adaptation experiments using the new LT to transform HMM means, the results were significantly better than MLLR for limited adaptation data and comparable to those in [8], while using the computationally simpler MLS FW estimation.

Mon2Cap : Language and Dialect Recognition
Poster Session, Cambria Room, 14:00 – 16:00
Chair: Joseph P. Campbell, Massachusetts Institute of Technology, USA

Automatic Language Identification Using Wavelets
Ana Lilia Reyes-Herrera, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez; INAOE, Mexico
Mon2Cap.1
Spoken language identification consists in recognizing a language based on a sample of speech from an unknown speaker. The traditional approach for this task mainly considers the phonotactic information of languages. However, for marginalized languages — languages with few speakers or oral languages without a fixed writing standard —, this information is practically not at hand and consequently the usual approach is not applicable. In this paper, we present a method that only considers the acoustic features of the speech signal and does not use any kind of linguistic information. This method applies a wavelet transform to extract the acoustic features of the speech signal. The experimental results on a pairwise discrimination task among nine languages demonstrated that this approach considerably outperforms other previous methods based on the sole use of acoustic features.

Minimum Classification Error Training of Hidden Markov Models for Acoustic Language Identification
Josef G. Bauer, Ekaterina Timoshenko; Siemens AG, Germany
Mon2Cap.2
The goal of Acoustic Language Identification (LID) is to identify the language of spoken utterances. The described system is based on parallel Hidden Markov Model (HMM) phoneme recognizers. The standard approach for parameter learning of Hidden Markov Model parameters is Maximum Likelihood (ML) estimation which is not directly related to the classification error rate. Based on the Minimum Classification Error (MCE) parameter estimation scheme we introduce Minimum Language Identification Error (MLIDE) training that results in HMM model parameters (mean vectors) that give minimum classification error on the training data. Using a large telephone speech corpus with 7 languages achieve a language classification error rate of 4.7% which is a 40% reduction of error rate compared with a baseline system using ML trained HMMs. Even if the system trained on fixed network telephone speech is applied to mobile network speech data MLIDE can greatly improve the system performance.

Unsupervised Adaptation for Acoustic Language Identification
Ekaterina Timoshenko, Josef G. Bauer; Siemens AG, Germany
Mon2Cap.3
Our system for automatic language identification (LID) of spoken utterances is performed with language dependent parallel phoneme recognition (PPR) using Hidden Markov Model (HMM) phoneme recognizers and optional phoneme language models (LMs). Such a LID system for continuous speech requires many hours of orthographically transcribed data for training of language dependent HMMs and LMs as well as phonetic lexica for every considered language (supervised training). To avoid the time consuming process of obtaining the orthographically transcribed training material we propose an algorithm for automatic unsupervised adaptation that requires only raw audio data as training material covering the requested language and acoustic environment. The LID system was trained and evaluated using fixed and mobile network databases (DBs) from the SpeechDat II corpus. The baseline system — based on supervised training using fixed network databases and covering 4 languages — achieved a LID error rate of 6.7% for fixed data and 19.5% for mobile data. Using unsupervised adaptation of the HMMs trained on fixed network data the error rate for mobile DBs database mismatch is reduced to 10.6%. Exploring a situation when orthographically transcribed training data is not available at all multilingual HMMs were unsupervised adapted to fixed and mobile DBs and perform at 10.8% and 12.4% error rate respectively.

Low Complexity LID Using Pruned Pattern Tables of LZW
S.V. Basavaraja, T.V. Sreenivas; Indian Institute of Science, India
Mon2Cap.4
We present two discriminative language modelling techniques for Lempel-Ziv-Welch (LZW) based LID system. The previous approach to LID using LZW algorithm was to directly use the LZW pattern tables for language modelling. But, since the patterns in a language pattern table are shared by other language pattern tables, confusability prevailed in the LID task. For overcoming this, we present two pruning techniques (i) Language Specific (LS-LZW) — in which patterns common to more than one pattern table are removed. (ii) Length-Frequency product based (LF-LZW) — in which patterns having their length-frequency product below a threshold are removed. These approaches reduce the classification score (Compression Ratio [LZW-CR] or the weighted discriminant score [LZW-WDS]) for non native languages and increases the LID performance considerably. Also the memory and computational requirements of these techniques are much less compared to basic LZW techniques.

Improved Language Identification Using Support Vector Machines for Language Modeling
Xi Yang1, Lu-Feng Zhai1, Manhong Sui1, Herbert Gish2; 1Hong Kong University of Science & Technology, China; 2BBN Technologies, USA
Mon2Cap.5
Automatic language identification (LID) decisions are made based on scores of language models (LM). In our previous paper [1], we have shown that replacing n-gram LMs with SVMs significantly improved performance of both the PPRLM and GMM-tokenization-based LID systems when tested on the OGI-TS corpus. However, the relatively small corpus size may limit the general applicability of the findings. In this paper, we extend the SVM-based approach on the larger CallFriend corpus evaluated using the NIST 1996 and 2003 evaluation sets. With more data, we found that SVM is still better than n-gram models. In addition, back-end processing is useful with SVM scores in Call-Friend which differs from our observation in the OGI-TS corpus. By combining the SVM-based GMM and phonotactic systems, our LID system attains
an ID error of 12.1% on NIST 2003 evaluation set which is more than 4% (25% relatively) better than the baseline n-gram system.

Recent Advances in Phonotactic Language Recognition Using Binary-Decision Trees
Jiří Navrátil; IBM T.J. Watson Research Center, USA

Mon2CaP.6

Binary decision trees are an effective model structure in language recognition. This paper presents several related algorithmic steps to address data sparseness issues and computational complexity. In particular, a tree-adaptation step, a recursive bottom-up smoothing step, and two variants of the Flip-Flop approximation algorithm are introduced to language detection and studied in the context of the NIST Language Recognition Evaluation task.

Fusion of Phonotactic and Prosodic Knowledge for Language Identification
Chi-Yueh Lin, Hsiao-Chuan Wang; National Tsing Hua University, Taiwan

Mon2CaP.7

Over the last few decades, language identification systems based on different kinds of linguistic knowledge had been studied by many researchers. Most of systems utilize one kind of linguistic knowledge only, i.e. phonotactic, phonetic repertoire, or prosody. It is possible to get the improvement by combining several linguistic knowledge. However, the combination of two systems based on different kinds of linguistic knowledge is not a trivial task. This paper presents a method where local identification results made by two individual systems, i.e. prosody-based and phonotactic-based systems, are fused in a Bayesian framework. Under this framework, local decisions, the associated false-alarm and miss probabilities are fused via Bayesian formulation to make the final decision. Experiments conducted on OGI/TTS corpus demonstrate the effectiveness of this decision-level fusion strategy.

Vector-Based Spoken Language Recognition Using Output Coding
Haizhou Li, Bin Ma, Rong Tong; Institute for Infocomm Research, Singapore

Mon2CaP.8

The vector-based spoken language recognition approach converts a spoken utterance into a high dimensional vector, also known as a bag-of-sounds vector, that consists of n-gram statistics of acoustic units. Dimension reduction would better prepare the bag-of-sounds vectors for classifier design. We propose projecting the bag-of-sounds vectors onto a low dimensional SVM output coding space, where each dimension represents a decision hyperplane between a pair of spoken vectors for classifier design. We propose projecting the bag-of-sounds vectors onto a low dimensional SVM output coding space, where each dimension represents a decision hyperplane between a pair of spoken vectors for classifier design. We propose projecting the bag-of-sounds vectors onto a low dimensional SVM output coding space, where each dimension represents a decision hyperplane between a pair of spoken vectors for classifier design. We propose projecting the bag-of-sounds vectors onto a low dimensional SVM output coding space, where each dimension represents a decision hyperplane between a pair of spoken vectors for classifier design. The trends also indicate that listeners’ decisions are influenced by conceptual representation of prototypical accent characteristics, such as “people from New York talk fast.” These observations suggest that listeners use both bottom-up processing, based on the acoustic input, and top-down processing, based on their conceptual representation of prototypical accent characteristics. Those processes are multi-dimensional in that listeners use utterance content (e.g., meaning or comprehensibility) as well as accent characteristics in the acoustic input even though our experiment focuses on recognition of individual accents. This study word selections that are dialect dependent. These findings contribute to a deeper understanding of the cognitive aspects of accent variation, and its applications for speech technology, such as accent classification for speaker identification or speech recognition.

Perceptual Identification and Phonetic Analysis of 6 Foreign Accents in French
Bianca Vieru-Dimulescu, Philippe Boula de Mareüil; LIMS, France

Mon2CaP.11

A perceptual experiment was designed to determine to what extent native French listeners are able to identify foreign accents in French: Arabic, English, German, Italian, Portuguese and Spanish. They succeed in recognizing the speaker’s mother tongue in more than 50% of cases (while rating their degree of accentedness as average). They perform best with Arabic speakers and worst with Portuguese speakers. The Spanish/Italian and English/German accents are the most mistaken ones. Phonetic analyses were conducted; clustering and scaling techniques were applied to the results, and were related to the listeners’ reactions that were recorded during the test. They support the idea that differences in the vowel realization (especially concerning the phoneme /y/) seem to outweigh rhythmic cues.
are talking spontaneously. A technique which aims to find the dialect dependence in the untranscribed audio by selecting the most discriminative Gaussian mixtures and selecting the most discriminative frames of speech is proposed. The Gaussian Mixture Model (GMM) based classifier is retrained after the dialect dependence information is identified. Both the MS-GMM (GMM trained with Mixture Selection) and FS-GMM (GMM trained with Frame Selection) classifiers improve dialect classification performance significantly. Using 122 speakers across three dialects of Spanish with 3.3 hours of speech, the relative error reduction is 30.4% and 26.1% respectively.

**Mon2FoP : Spoken Dialog Systems I**
Poster Session, Allegheny Ballroom Foyer, 14:00 – 16:00
Chair: Alexander I. Rudnick, Carnegie Mellon University, USA

**Dynamic Extension of a Grammar-Based Dialogue System: Constructing an All-Recipes Knowing Robot**
Petra Gieselmann, Alex Waibel; Universität Karlsruhe, Germany

In the upcoming field of humanoid and human-friendly robots, the ability of the robot for simple, unconstrained and natural communication with its users is of central importance. The basis for appropriate actions of the robot is the correct understanding of the user utterances. To be able to cover all the entities a user might talk about, we enhanced our dialogue manager with an ability for dynamic vocabulary generation out of information found across the internet. As a test case, we chose an internet recipe database integrated in the dialogue manager of our household robot so that it can understand several thousand recipes and ingredients now.

**Scalable and Portable Web-Based Multimodal Dialogue Interaction with Geographical Databases**
Alexander Gruenstein, Stephanie Seneff, Chao Wang; Massachusetts Institute of Technology, USA

We describe work towards developing a scalable and portable framework for enabling map-based multimodal dialogue interaction over the web. Working in the context of a restaurant-guide system, we show how large information databases harvested from the web can be accommodated in our speech recognizer, parser, and web-based GUI. We compare two dynamic language modeling techniques, which calculate context-dependent weights for the large sets of proper nouns associated with geographical entities such as restaurants and streets. We show that the more fine-grained approach results in a 7.8% reduction in concept error rate.

**System- versus User-Initiative Dialog Strategy for Driver Information Systems**
Chantal Ackermann, Marion Libossek; LMU München, Germany

In this paper we examine the advantages of a system-initiative approach for the voice user interface of a driver information system (DIS). The problems of user-initiative systems are the steep learning curve and the high demand on memory to recall the correct voice command. This is especially true in a car environment where the main, and most important task is to drive the car. In a Wizard of Oz experiment, we compared this approach to one that uses a more system-initiative form of interaction. Furthermore, a context sensitive help prompt was included in the new system, instead of just a context sensitive list of commands. The results show that, for novice users, the error rate, the number and time of task completion, the mental workload, and the subjective *joy of use* (as measured by a semantic differential) are all better for the proposed system. Nevertheless, the possibility to use shortcuts remains. Thus, an expert user could still skip the supermenus and jump into the given submenu by saying the right voice command.

**Have We Met? MDP Based Speaker ID for Robot Dialogue**
Filip Krsmovonic, Curtis Spencer, Daniel Jurafsky, Andrew Y. Ng; Stanford University, USA

We present a novel approach to speaker identification in robot dialogue that allows a robot to recognize people during natural conversation and address them by name. Our Stanford AI Robot (STAIR) dialogue system attempts to mirror the human speaker identification process. We model the robot dialogue problem as a Markov Decision Process (MDP) and apply a reinforcement learning algorithm to try to learn the optimal dialogue actions. The MDP model works in conjunction with a traditional statistical cluster based speaker identification subsystem. Our approach also addresses open-set speaker identification, dynamically adding new speaker profiles as well as continuously updating known profiles.

**Prominent Words as Anchors for TRP Projection**
R.J.H. van Son, Wienieke Wesseling, Louis C.W. Pols; University of Amsterdam, The Netherlands

The effect of the position of the last accented word on the projection of TRPs was investigated with two RT experiments. Subjects were asked to respond with minimal responses to prerecorded dialogs and impoverished versions of these dialogs, containing either only intonation and pause information, *humm*ed stimuli, or no periodic component at all, *whispered* stimuli. The distribution of these elicited response delays was comparable to that of natural turn switches. It is shown that the presence of non-prominent words before a TRP reduces the delays of elicited and natural responses alike, even in impoverished speech. This suggests that the presence of an prominent, informative, word starts the projection of a possible upcoming TRP. The availability of non-prominent, predictable, speech then allows listeners to improve their predictions of the exact timing of the TRP.

**Learning Multi-Goal Dialogue Strategies Using Reinforcement Learning with Reduced State-Action Spaces**
Heriberto Cuayhuitl, Steve Renals, Oliver Lemon, Hiroshi Shimodaira; University of Edinburgh, UK

Learning dialogue strategies using the reinforcement learning framework is problematic due to its expensive computational cost. In this paper we propose an algorithm that reduces a state-action space to one which includes only valid state-actions. We performed experiments on full and reduced spaces using three systems (with 5, 9 and 20 slots) in the travel domain using a simulated environment. The task was to learn multi-goal dialogue strategies optimizing single and multiple confirmations. Average results using strategies learnt on reduced spaces reveal the following benefits against full spaces: 1) less computer memory (94% reduction), 2) faster learning (93% faster convergence) and better performance (8.4% less time steps and 7.7% higher reward).

**Pitch Range and Pause Duration as Markers of Discourse Hierarchy:**
Jörg Mayer, Ekaterina Jasinskaja, Ulrike Kölsh; University of Potsdam, Germany

Discourse structure is reflected by a number of global prosodic parameters, like for example pause duration and pitch range. Dis-
course structure is also known to affect the accessibility/salience of antecedents of anaphoric expressions. Assuming these generalizations are correct, one can ask whether listeners use the information encoded in pauses and pitch range to resolve anaphoric references in ambiguous contexts. To examine this, we conducted a series of perception experiments with ambiguous discourses, where the pitch range of the sentences and the pause duration between sentences was manipulated. The results of our experiments corroborate our main research hypothesis that global prosodic parameters influence the resolution of anaphoric pronouns. The direction of the observed effect is clearly in accordance with the predictions of the existing theories of discourse anaphora and the current state of research on discourse prosody.

Radiobot-CFF: A Spoken Dialogue System for Military Training

Antonio Roque, Anton Leuski, Vivek Rangarajan, Susan Robinson, Ashish Vaswani, Shrikanth Narayanan, David Traum; University of Southern California, USA

Mon2FoP.8

We describe a spoken dialogue system which can engage in Call For Fire (CFF) radio dialogues to help train soldiers in proper procedures for requesting artillery fire missions. We describe the domain, an information-state dialogue manager with a novel system of interactive information components, and provide evaluation results.

Is Voice Quality Enough? — Study on How the Situation and User’s Awareness Influence the Utterance Features

Shinya Yamada, Toshihiko Itoh, Kenji Araki; Hokkaido University, Japan

Mon2FoP.9

This paper presents the characteristic differences of linguistic and acoustic features observed in different spoken dialogue situations and with different dialogue partners: human-human vs. human-machine interactions. And it also presents influences of awareness of users on those characteristics. We compare the linguistic and acoustic features of the user’s speech to a spoken dialogue system and to a human operator in several goal setting and destination database searching tasks for a car navigation system. Because it is not clear enough whether different dialogue situations and different dialogue partners cause any differences of linguistic or acoustic features on one’s utterances in a speech interface system, we have performed experiments in several dialogue situations[4]. However, in these experiments the conditions such as voice quality and awareness of users such as impressions on the partner and prejudices against a system have not been considered. And so we collected a set of spoken dialogues in new dialogue situations. To investigate influence of voice quality, we also prepare recorded voice for response of dialogue partners and compared the influences of voice (natural voice, synthetic voice and recorded voice). We also made users answer questionnaire before and after the experiments and investigated characteristic differences caused by awareness of users. Additionally, in order to confirm the usefulness of the results of all experiments, we actually applied acoustic features of users’ utterances and identified the utterances made to a system.

Development of Slovak GALAXY/VoiceXML Based Spoken Language Dialogue System to Retrieve Information from the Internet

Jozef Juhár1, Stanislav Ondas1, Anton Čižmár1, Milan Rusko2, Gregor Rozinaj3, Roman Jarina4; 1Technical University of Košice, Slovak Republic; 2Slovak Academy of Science, Slovak Republic; 3Slovak University of Technology, Slovak Republic; 4University of Žilina, Slovak Republic

Mon2FoP.10

In this paper we describe the research and development of the first Slovak spoken language dialogue system. The dialogue system is based on the DARPA Communicator architecture. The proposed system consists of the Galaxy hub and telephony, automatic speech recognition, text-to-speech, backend, transport and VoiceXML dialogue management modules. The SLDS enables multi-user interaction in the Slovak language. The functionality of the SLDS is demonstrated and tested via two pilot applications, “Weather forecast for Slovakia” and “Timetable of Slovak Railways”. The required information is retrieved from Internet resources in multi-user mode through PSTN, ISDN, GSM and/or VoIP network.

LINTest: A Development Tool for Testing Dialogue Systems

Lars Degerstedt, Arne Jönsson; Linköpings Universitet, Sweden

Mon2FoP.11

In this paper we present a development tool for testing dialogue systems. Testing software through the specification is important for software development in general and should be as automated as possible. For dialogue systems, the corpus can be seen as one part of the specification and the dialogue system should be tested on available corpora on each new build. The testing tool is inspired from work on agile software development methods, test driven development and unit testing, and can be used in two modes and during various phases of development.
Mon3A1O : Speaker Characterization and Recognition I

Oral Session, Allegheny Ballroom, left, 16:30 - 18:30
Chair: Sadaoki Furui, Tokyo Institute of Technology, Japan

Improving the Characterization of the Alternative Hypothesis via Kernel Discriminant Analysis for Likelihood Ratio-Based Speaker Verification
Yi-Hsiang Chao1, Wei-Ho Tsai2, Hsin-Min Wang1, Ruei-Chuan Chang1,1; Academia Sinica, Taiwan; 2National Taipei University of Technology, Taiwan

Mon3A1O.1

The performance of a likelihood ratio-based speaker verification system is highly dependent on modeling of the target speaker’s voice (the null hypothesis) and characterization of non-target speakers’ voices (the alternative hypothesis). To better characterize the ill-defined alternative hypothesis, this study proposes a new likelihood ratio measure based on a composite-structure Gaussian mixture model, the so-called GMM2. Motivated by the combined use of a variety of background models to represent the alternative hypothesis, GMM2 is designed with an inner set of mixture weights connected to the significance of each individual Gaussian density, and an outer set of mixtures connected to the significance of each individual background model. Through the use of kernel discriminant analysis namely, Kernel Fisher Discriminant (KFD) or Support Vector Machine (SVM), GMM2 is trained in such a manner that the utterances of the null hypothesis can be optimally separated from those of the alternative hypothesis.

A Discriminative Method for Speaker Verification Using the Difference Information
Zhenchun Lei, Yingchun Yang, Zhaohui Wu; Zhejiang University, China

Mon3A1O.2

In this paper, a discriminative method is proposed for speaker verification. An utterance can be mapped into a matrix by computing the difference to a codebook, and then expand the mapped matrix to a vector as the input of support vector machines for speaker verification. The Gaussian mixture model-based method is also constructed by utilizing its nature. The mapped vector indicates the utterance’s fitness to the codebook. Compared with the derivative operation in the famous fisher kernel the difference operation is used in our method. Experiments were run on the YOHO database in the text-independent scenario. The new method is superior to the conventional GMM for speaker verification.

A Multiclass Framework for Speaker Verification Within an Acoustic Event Sequence System
Nicolas Scheffer, Jean-François Bonastre; LIA, France

Mon3A1O.3

Building acoustic events and their sequence analysis (AES system) is a method that proved its efficiency in [1]. Indeed, the methodology combines the power of the world model GMM, used in state-of-the-art speaker detection systems, for extracting speaker independent events with an analysis of these event sequences via tools usually used in so-called High Level Speaker Detection systems. The efficiency of this system has been validated at the last NIST evaluation campaign. This paper aims at proposing a new framework by applying an AES system on multiple classes, C-AES. The originality of this work is to consider that intraclass sequence analysis can bring more information than a global analysis on the whole speaker utterance. This paper also proposes a method to take into account the priori knowledge of the classes within the scoring process. The results support the fact that intraclass information is discriminant for speaker verification, as a combination with a state-of-the-art GMM brings a 12% relative gain at the DCF.

Speaker Cluster Based GMM Tokenization for Speaker Recognition
Bin Ma, Donglai Zhu, Rong Tong, Haizhou Li; Institute for Infocomm Research, Singapore

Mon3A1O.4

We present a speaker recognition system with multiple GMM tokenizers as the front-end, and vector space modeling as the back-end classifier. GMM tokenizer captures the acoustic and phonetic characteristics of a speaker from the speech signals without the need of phonetic transcription. To enhance the speaker characteristics coverage and provide more discriminative information, a speaker clustering algorithm is proposed to build multiple GMM tokenizers that are arranged in parallel. For an input utterance, each of the tokenizers outputs a token sequence, which is then represented by a vector of n-gram probabilities. Multiple vectors are concatenated to form a composite vector. Finally the Support Vector Machine (SVM) is used as the back-end classifier of the composite vectors. We use the 2002 NIST Speaker Recognition Evaluation (SRE) corpus for training GMM tokenizers and background modeling, and evaluate on the 2001 NIST SRE corpus.

Intra-Speaker Variability Compensation in Speaker Verification with Limited Enrolling Data
Claudio Garreton, Nestor Becerra Yoma, Carlos Molina, Fernando Huenupan; Universidad de Chile, Chile

Mon3A1O.5

In this paper a compensation method is proposed to address the problem of limited enrolling data in speaker verification. Instead of adapting the client HMM, the technique presented here modifies the speaker’s prior distribution by maximizing the a posteriori p.d.f. in order to optimize the reduction in intra-speaker variability. The proposed approach can lead to reductions of 38.9% and 61.8% in EER and in the integral below the false-acceptation / false-rejection ROC curve, respectively.

Speaking Faces for Face-Voice Speaker Identity Verification
Girija Chetty, Michael Wagner; University of Canberra, Australia

Mon3A1O.6

In this paper, we describe an approach for an animated speaking face synthesis and its application in modeling impostor/replay attack scenarios for face-voice based speaker verification systems. The speaking face reported here learns the spatiotemporal relationship between speech acoustics and MPEG4 compliant facial animation points. The influence of articulatory, perceptual, and prosodic acoustic features along with auditory context on prediction accuracy was examined. The results are indicative of vulnerability of audiovisual identity verification systems to impostor/replay attacks using synthetic faces. The level of vulnerability depends on several factors, such as the type of audiovisual features, the fusion techniques used for the audio and video features and their relative robustness. Also, the success of the synthetic impostor depends on the type of co-articulation models and acoustic features used for the audiovisual mapping of speaking face synthesis.
A Study on Lattice Rescoring with Knowledge Scores for Automatic Speech Recognition
Sabato Marco Siniscalchi, Jinyu Li, Chin-Hui Lee; Georgia Institute of Technology, USA
Mon3A2O.1
We study lattice rescoring with knowledge scores for automatic speech recognition. Frame-based log likelihood ratio is adopted as a score measure of the goodness-of-fit between a speech segment and the knowledge sources. We evaluate our approach in two different applications: phone recognition, and connected digit continuous recognition. By incorporating knowledge scores obtained from 15 attribute detectors for place and manner of articulation, we reduced phone error rate from 40.52% to 35.16% using monophone models. The error rate can be further reduced to 33.42% for triphone models. The same lattice rescoring algorithm is extended to connected digit recognition using the TIDIGITS database, and without using any digit-specific training data. We observed the digit error rate can be effectively reduced to 4.03% from 4.54% which was obtained with the conventional Viterbi decoding algorithm with no knowledge scores.

Cross-System Adaptation and Combination for Continuous Speech Recognition: The Influence of Phoneme Set and Acoustic Front-End
Sebastian Stüker1, Christian Fügen1, Susanne Burger2, Matthias Wölfl1,1 Universität Karlsruhe, Germany; 2 Carnegie Mellon University, USA
Mon3A2O.2
Cross-system adaptation and system combination methods, such as ROVER and confusion network combination, are known to lower the word error rate of speech recognition systems. They require the training of systems that are reasonably close in performance but at the same time produce output that differs in its errors. This provides complementary information which leads to performance improvements. In this paper we demonstrate the gains we have seen with cross-system adaptation and system combination on the English EPPS and RT0-05S lecture meeting task. We obtained the necessary varying systems by using different acoustic front-ends and phoneme sets on which our models are based. In a set of contrastive experiments we show the influence that the exchange of the components has on adaptation and system combination.

Generating Complementary Systems for Speech Recognition
C. Breslin, M.J.F. Gales; University of Cambridge, UK
Mon3A2O.3
Large Vocabulary Continuous Speech Recognition (LVCSR) systems often use a multi-pass recognition framework where the final output is obtained from a combination of multiple models. Previous systems within this framework have normally built a number of independently trained models, before performing multiple experiments to determine the optimal combination. For two models to give improvements upon combination, it is clear that they must be complementary, i.e. they must make different errors. While independently trained models often do give improvements when they are combined, it is not guaranteed that they will be complementary. This paper presents a new algorithm, Minimum Bayes Risk Leveraging (MBRL), for explicitly generating systems that are complementary to each other. This algorithm is based on Minimum Bayes Risk training, but within a boosting-like iterative framework. Experimental results are reported on a Broadcast News Mandarin task. These experiments show small but consistent gains when combining complementary systems using confusion network combination.
Towards an Integrated Understanding of Speaking Rate in Conversation

Jiahong Yuan, Mark Liberman, Christopher Cieri; University of Pennsylvania, USA

Mon3A3O.1
We investigate factors that affect speaking rate in conversation, using large corpora of conversational telephone speech in English and Chinese. We find that speaking rate as a function of "turn" length rises rapidly for turns from one to seven words; remains level (when final words are included) or falls gradually (if final words are excluded) for turns of medium length; and rises slowly for longer turns. When talking with strangers or discussing certain topics, people tend to use longer turns but slower speech rates. In general older people have a slower speech, and males tend to speak slightly faster than females. Finally, we find that the effect of L1 (native language) on L2 (second language) speaking rate is L1 dependent.

Prosody of Interrogative and Affirmative Sentences in Vietnamese Language: Analysis and Perceptive Results

Minh Quang Vă, Đỗ Nhật Trân, Eric Castelli; MICA, Vietnam

Mon3A3O.2
This paper presents a new study on the prosody of Vietnamese language. Sentence pairs containing one interrogative sentence and one affirmative sentence, which have the same tones and the same number of syllables to avoid the effects of lexical tones and of co-articulation, are recorded in order to analyze their prosody evolution. Comparisons allow us to characterize differences between interrogative and affirmative sentences at sentence prosody level. Our work is completed by a perceptual study on re-synthesized sound where all syllables of the sentence are replaced by the vowel /a/ to hide lexical meaning, while the prosody of the sentence is kept unchanged. Our goal is to see if such a prosody carries any information about sentence nature characteristics, and then whether it enables listeners to classify sentence type (in this case interrogative and affirmative), despite the complex form of this prosody in tonal languages. The obtained results show that information on sentence type is present at the end of the second half of the last syllable and that about 70% of sentences are properly classified for female synthesis voice.

Intonational Cues to Student Questions in Tutoring Dialogs

Jennifer J. Venditti, Julia Hirschberg, Jackson Liscombe; Columbia University, USA

Mon3A3O.3
Successful Intelligent Tutoring Systems (ITSs) must be able to recognize when their students are asking a question. They must identify question form as well as function in order to respond appropriately. Our study examines whether intonational features, specifically, F0 height and rise range, are useful cues to student question type in a corpus of 643 American English questions. Results show a quantitative effect of both form and function. In addition, among clarification-seeking questions, we observed differences based on the type of clarification being sought.

Testing the Effect of Audiovisual Cues to Prominence via a Reaction-Time Experiment

Emiel Krahmer, Marc Swerts; Tilburg University, The Netherlands

Mon3A3O.4
This article discusses a perception experiment to investigate the relation between auditory and visual cues for marking prosodic prominence. The methodology makes use of a reaction-time experiment. For this experiment, recordings of a sentence with 3 accents were systematically manipulated in such a way that auditory and visual markers of prominence were either congruent (occurring on the same word) or incongruent (in that the auditory and the visual cues were positioned on different words). Subjects were instructed to indicate as fast as possible which word they perceived as the most prominent one. Classification results show first of all that subjects’ responses were much more dependent on auditory than on visual cues. In addition, however, we found that incongruent stimuli lead to slower reaction times than congruent stimuli, showing that visual cues do have an impact on the cognitive processing of prosodic prominence.

Effect of Genre, Speaker, and Word Class on the Realization of Given and New Information

Agustin Gravano, Julia Hirschberg; Columbia University, USA

Mon3A3O.5
There is much evidence in the literature that speakers tend to deaccent discourse-given entities, while accenting new ones. However, speakers do not always follow this simple strategy and the causes for such variation are not yet well understood. In this paper, we describe several new forms of variability in the relationship between given/new information and accenting behavior, variation due to individual differences and to word class. We present results indicating that different speakers have different strategies for making new words prominent. We analyze two word-classes — nouns and verbs — in a corpus of spontaneous and read direction-giving monologues, and show that speakers use different combinations of pitch, intensity and inter-word pauses to distinguish between given and new information. Most interestingly, we find that in both genres all speakers tend to produce given verbs with higher intensity than new verbs.

Word Order and Tonal Shape in the Production of Focus in Short Finnish Utterances

Martti Vainio1, Juhani Järviäki2, Stefan Werner3; 1University of Helsinki, Finland; 2University of Turku, Finland; 3University of Joensuu, Finland

Mon3A3O.6
This paper presents results from a study on the production of Finnish prosody. The effect of word order and the tonal shape in the production of Finnish prosody was studied as produced by 8 native Finnish speakers. Predictions formulated with regard to results from an earlier study pertaining to the perception of prominence were tested. These predictions had to do with the tonal shape of the utterances in the form of a flat hat pattern and the effect of word order on the so called top-line declination within an adverbial phrase in the utterances. The results from the experiment give support to the following claims: the temporal domain of prosodic focus is the whole utterance, word order reversal from unmarked to marked has an effect on the production of prosody, and the production of the tonal aspects of focus in Finnish follows a basic flat hat pattern. That is the prominence of a word can be produced by an f0 rise or a fall, depending on the location of the word in an utterance. The basic accentual shape of a Finnish word is then not a pointed rise/fall hat shape as claimed before since it can vary depending on the syllable structure and the position within an utterance.
Modeling Sensory-to-Motor Mappings Using Neural Nets and a 3D Articulatory Speech Synthesizer

Bernd J. Kröger, Peter Birkholz, Jim Kannampuzha, Christiane Neuschafer-Rube, University Hospital Aachen, Germany; University of Rostock, Germany

A comprehensive neural model of speech motor control including a three dimensional articulatory speech synthesizer as a front-end device is described in detail in this paper. The training of the sensory-to-motor mappings — which can be interpreted as the prelinguistic phase of speech acquisition — is described in detail for quasi-static as well as for dynamic articulation.

Semi-Automatic Extraction of Vocal Tract Movements from Cineradiographic Data

Julie Fontecave, Frédéric Berthommier; ICP, France

Since high speed X-ray films still provide the best dynamic view of the entire vocal tract, large existing databases have been preserved and are available for the speech research community. We propose a new technique for facilitating the extraction of the vocal tract shape and the movements of the articulators from complete sequences of these databases. The method was first developed for the extraction of the tongue movements in “Wioland”. It has been adapted to a sequence of the ATR database, Laval43 (provided by Rochette). The method, based on the retromarking algorithm, combines the human expertise applied for marking a small number of key images, and the automatic processing of the video data. It has been extended to other articulators (lips, velum) in order to obtain the shape and the sections of the complete vocal tract. Quantitative evaluations of the estimate error and a comparison with Thimm and Luettin (1999) are achieved.

Towards Continuous Speech Recognition Using Surface Electromyography

Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, Alex Waibel; Carnegie Mellon University, USA; Universität Karlsruhe, Germany

We present our research on continuous speech recognition of the surface electromyographic signals that are generated by the human articulatory muscles. Previous research on electromyographic speech recognition was limited to isolated word recognition because it was very difficult to train phoneme-based acoustic models for the electromyographic speech recognizer. In this paper, we demonstrate how to train the phoneme-based acoustic models with carefully designed electromyographic feature extraction methods. By decomposing the signal into different feature space, we successfully keep the useful information while reducing the noise. Additionally, we also model the anticipatory effect of the electromyographic signals compared to the speech signal. With a 108-word decoding vocabulary, the experimental results show that the word error rate improves from 86.8% to 32.0% by using our novel feature extraction methods.

A Trajectory Mixture Density Network for the Acoustic-Articulatory Inversion Mapping

Korin Richmond; University of Edinburgh, UK

This paper proposes a trajectory model which is based on a mixture density network trained with target features augmented with dynamic features together with an algorithm for estimating maximum likelihood trajectories which respects constraints between the static and derived dynamic features. This model was evaluated on an inversion mapping task. We found the introduction of the trajectory model successfully reduced root mean square error by up to 7.5%, as well as increasing correlation scores.

Articulatory Features for “Meeting” Speech Recognition

Florian Metze; Technische Universität Berlin, Germany

“Meeting” speech, for example from the RT-045 task, contains a mixture of different speaking styles that leads to word error rates higher than 25% even when close-talking microphones are being used. The problem is even more serious, as word error rates are particularly high when speakers use a clear speaking mode, for example because they want to stress an important point. Previous work showed that an approach that combines standard phone-based acoustic models with models detecting the presence or absence of “Articulatory Features” such as “Rounded” or “Voiced” can improve ASR performance particularly for these cases. This paper presents a discriminative approach to automatically computing from training or adaptation data the feature stream weights needed for the above approach, therefore presenting a framework for integrating articulatory features into existing automatic speech recognition systems. We find a 7% relative improvements on top of our best RT-045 system using discriminative adaptation.

Training of Coarticulation Models Using Dominance Functions and Visual Unit Selection Methods for Audio-Visual Speech Synthesis

Zdeněk Kriouil, Miloš Železný, Laděk Müller, Jakub Kanis; University of West Bohemia, Czech Republic

This paper presents results of training of coarticulation models for Czech audio-visual speech synthesis. Two approaches for solution of coarticulation in audio-visual speech synthesis were used, coarticulation based on dominance functions and visual unit selection. For both approaches, coarticulation models were trained. Models for unit selection approach were trained by visually clustered data. These data were obtained using decision tree algorithm. Outputs of audio-visual speech synthesis for both approaches were assessed and compared objectively.

Phone Recognition Analysis for Trajectory HMM

Le Zhang, Steve Renals; University of Edinburgh, UK

The trajectory HMM has been shown to be useful for model-based speech synthesis where a smoothed trajectory is generated using temporal constraints imposed by dynamic features. To evaluate the performance of such model on an ASR task, we present a trajectory decoder based on tree search with delayed path merging. Experiment on a speaker-dependent phone recognition task using the MOCHA-TIMIT database shows that the MLE-trained trajectory model, while retaining attractive properties of being a proper generative model, tends to favour over-smoothed trajectory among competing hypotheses, and does not perform better than a conventional HMM. We use this to build an argument that models giving better fit on training data may suffer a reduction of discrimination by being too faithful to training data. This partially explains why alternative acoustic models that try to explicitly model temporal constraints do not achieve significant improvements in ASR.
Discriminative Kernel-Based Phoneme Sequence Recognition
Joseph Keshet1, Shai Shalev-Shwartz2, Samy Bengio2, Yoram Singer1, Dan Chazan3; 1Hebrew University, Israel; 2IDIAP Research Institute, Switzerland; 3Technion, Israel
Mon3BuP.2
We describe a new method for phoneme sequence recognition given a speech utterance, which is not based on the HMM. In contrast to HMM-based approaches, our method uses a discriminative kernel-based training procedure in which the learning process is tailored to the goal of minimizing the Levenshtein distance between the predicted phoneme sequence and the correct sequence. The phoneme sequence predictor is devised by mapping the speech utterance along with a proposed phoneme sequence to a vector-space endowed with an inner-product that is realized by a Mercer kernel. Building on large margin techniques for predicting whole sequences, we are able to devise a learning algorithm which distills to separating the correct phoneme sequence from all other sequences. We describe an iterative algorithm for learning the phoneme sequence recognizer and further describe an efficient implementation of it. We present initial encouraging experimental results with the TIMIT and compare the proposed method to an HMM-based approach.

Combining Phonetic Attributes Using Conditional Random Fields
Jeremy Morris, Eric Fosler-Lussier; Ohio State University, USA
Mon3BuP.3
A Conditional Random Field is a mathematical model for sequences that is similar in many ways to a Hidden Markov Model, but is discriminative rather than generative in nature. In this paper, we explore the application of the CRF model to ASR processing of discriminative phonetic features by building a system that performs first-pass phonetic recognition using discriminatively trained phonetic features. With this system, we show that this CRF model achieves an accuracy level in a phone recognition task that is superior to a similarly trained HMM model.

Discriminative MLE Training Using a Product of Gaussian Likelihoods
T. Nagarajan, Douglas O’Shaughnessy; Université du Québec, Canada
Mon3BuP.4
In this paper, we describe a discriminative technique to determine an optimal HMM topology for the each of the models in a continuous speech recognition system such that the word error rate (WER) is minimized. In conventional model selection techniques such as Bayesian information criterion (BIC), the model complexity is determined without considering the other classes in a system. In our work, an optimal model topology is selected by considering how well a given model can discriminate examples of other classes from its own. By doing so, the estimated model parameters indirectly make sure that class separability is increased. In an earlier work [1], we have proposed this technique and experiments were carried out on an E-set. Presently, we extend it for building a syllable-based continuous speech recognition system. Preliminary experiments carried out on the TIMIT corpus show that a considerable reduction in WER can be achieved using the proposed technique over the BIC-based technique for model selection.

State-Level Variable Modeling for Phoneme Classification
Hao-Zheng Li, Douglas O’Shaughnessy; Université du Québec, Canada
Mon3BuP.5
In HMM-based pattern recognition, the structure of the HMM is often predetermined according to some prior knowledge. In the recognition process, we usually make our judgment based on the maximum likelihood of the HMM, without considering the time-varying property of state-level variables, which unfortunately may lead to incorrect results. In this paper, we analyze the property of state-level variables in the HMM and show it is possible to significantly enhance the performance of speech recognition systems when using the state-level variable time-varying property. We propose four methods to model state-level variable trajectories and then test them on a phoneme classification task on the TIMIT speech corpus, 11.95% error rate reduction is achieved and some empirical conclusions are drawn.

A Time-Synchronous Phonetic Decoder for a Long-Contextual-Span Hidden Trajectory Model
Xiaolong Li, Li Deng, Dong Yu, Alex Acero; Microsoft Research, USA
Mon3BuP.6
A novel time-synchronous decoder, designed specifically for a Hidden Trajectory Model (HTM) whose likelihood score computation depends on long-span phonetic contexts, is presented. HTM is a recently developed acoustic model aimed to capture the underlying dynamic structure of speech coarticulation and reduction using a compact set of parameters. The long-span nature of the HTM had posed a great technical challenge for developing efficient search algorithms for full evaluation of the model. Taking on the challenge, the decoding algorithm is developed to deal effectively with the exponentially increased search space by HTM-specific techniques for hypothesis representation, word-ending recombination, and hypothesis pruning. Experimental results obtained on the TIMIT phonetic recognition task are reported, extending our earlier HTM evaluation paradigms based on N-best and A* lattice rescoring.

Analysis of HMM Temporal Evolution for Automatic Speech Recognition and Utterance Verification
Marta Casar, José A.R. Fonollosa; Universitat Politècnica de Catalunya, Spain
Mon3BuP.7
This paper proposes a double layer speech recognition and utterance verification system based on the analysis of the temporal evolution of HMM’s state scores. For the lower layer, it uses standard HMM-based acoustic modeling followed by a Viterbi grammar-based search step which provides us with the state scores of the acoustic models. In the second layer, these state scores are added to the regular set of acoustic parameters, building a new set of expanded HMMs. Using this expanded set of HMMs for speech recognition a significant improvement in performance is achieved. Next, we will use this new architecture for utterance verification in a “second opinion” framework. We will consign to the second layer evaluating the reliability of decoding using the acoustic models from the first layer. An outstanding improvement in performance versus a baseline verification algorithm has been achieved.

Improvements to Bucket Box Intersection Algorithm for Fast GMM Computation in Embedded Speech Recognition Systems
Min Tang, Aravind Ganapathiraju; Conversay, USA
Mon3BuP.8
Real-time performance is a very important goal for embedded speech recognition systems, where the evaluation of likelihoods for Gaussian mixture models (GMM) usually dominates the computation of a continuous density hidden Markov model (CDHMM) based system. The Bucket Box Intersection (BBI) algorithm is an optimization technique that uses a K-Dimensional binary tree to speed up the score computation of GMM without significantly hurting the recognition accuracy. In this paper, we propose three improvements to the traditional BBI algorithm. First, we define the optimal dividing hyper-plane as the plane
that generates minimum expected number of mixture evaluations instead of the median hyper-plane. The size of BBI tree is reduced largely because of that. Second, we refine the location of dividing plane as the one that has the same Mahalanobis distance to the closest dividing mixture pair, instead of the boundary of Gaussian box. By doing this, we are able to improve recognition accuracy. Finally, we introduce the dividing planes which run across two dimensions to boost the range of dividing plane candidates and thus bring more speed-ups. We evaluated these techniques using Conversay’s speech engine CASSI in 2 different domains. The experimental results of new BBI algorithm show significant performance improvement over traditional BBI algorithm. Compared to the baseline system with no BBI algorithm implementation, we were able to speed up Gaussian computations by 50% with a less than 5% relative increase in word error rate.

**Forward-Backwards Training of Hybrid HMM/BN Acoustic Models**

**Konstantin Markov, Satoshi Nakamura; NICT, Japan**

Mon3BnP.9

In this paper, we describe an application of the Forward-Backwards (F-B) algorithm for maximum likelihood training of hybrid HMM/Bayesian Network (BN) acoustic models. Previously, HMM/BN parameter estimation was based on a Viterbi training algorithm that requires two passes over the training data: one for BN learning and one for updating HMM transition probabilities. In this work, we first analyze the F-B learning for a conventional HMM and show that the state PDF parameter estimation is analogous to weighted-data classifier training. The gamma variable of the Forward-Backwards algorithm plays the role of the data weight. From this perspective, it is straightforward to apply F-B-based training to the HMM/BN models since the BN learning algorithm allows training with weighted data. Experiments on stressed speech (American, British and Australian English) show that F-B training outperforms the previous Viterbi learning approach and that the HMM/BN model achieved better performance than the conventional HMM.

**A Comparative Study of Gaussian Selection Methods in Large Vocabulary Continuous Speech Recognition**

**Dirk Gehrig1, Thomas Schaaf2; 1Universität Karlsruhe, Germany; 2 Carnegie Mellon University, USA**

Mon3BuP.10

Gaussian mixture models are the most popular probability density used in automatic speech recognition. During decoding, often many Gaussians are evaluated. Only a small number of Gaussians contribute significantly to probability. Several promising methods to select relevant Gaussians are known. These methods have different properties in terms of required memory, overhead and quality of selected Gaussians. Projection search, bucket box intersection, and Gaussian clustering are investigated in a broadcast news system with focus on adaptation (MLLR).

**A Successive State and Mixture Splitting for Optimizing the Size of Models in Speech Recognition**

**Soo-Young Suk1, Seong-Jun Hahm2, Ho-Youl Jung2, Hyun-Yeol Chung2; 1AIST, Japan; 2 Yeungnam University, Korea**

Mon3BuP.11

A Successive State and Mixture Splitting (SSMS) algorithm for optimizing the size of models used in speech recognition for small size of mobile devices is proposed in this paper. The proposed algorithm employs essentially Continuous Hidden Markov Model (CHMM) structure and this CHMM consists of variable parameter topology in order to minimize the number of model parameters and to reduce recognition time. SSMS splits the Gaussian Output Probability Density Distribution (GOPDD) for variable parameter context independent model. Unlike the Successive State Splitting generating context dependent model, the algorithm constructs context independent model with suitable number of states and mixtures for each recognition units by automatic splitting of GOPDD in time and mixture domain. The recognition results showed that the proposed SSMS could reduce the total number of Gaussian up to 40.0% compared with the fixed parameter models at the same performance in speech recognition.

**Improved Source Modeling and Predictive Classification for Channel Robust Speech Recognition**

**Valentin Ion, Reinhold Haeb-Umbach; University of Paderborn, Germany**

Mon3BuP.12

The accuracy of distributed speech recognition has been shown to be very sensitive to errors occurring during transmission. One reason for this is that the classifier, usually trained under error free conditions, is unable to cope with the mismatch between an error free and error prone channel. In this paper we present a novel decision rule for classification which is able to account for channel errors. To achieve this, the classical Bayesian speech recognition approach has been reformulated for the server side, where the observation is known only to the extent, as is given by its a posteriori density function. We present a method to estimate the a posteriori density which is based on a Markov model of the source, which captures correlations of both static and dynamic features. A practical implementation is given, accompanied by experimental results for distributed speech recognition over an IP-network.

**Mon3CaP : Acoustic Signal Segmentation and Classification**

Poster Session, Cambria Room, 16:30 – 18:30

Chair: Jose M. Pardo, International Computer Science Institute, USA

**Automatic English Stop Consonants Classification Using Wavelet Analysis and Hidden Markov Models**

**Marco Kühne, Roberto Togneri; University of Western Australia, Australia**

Mon3CaP.1

This paper compares wavelet and STFT analysis for a speaker-independent stop classification task using the TIMIT database. In the designed experiment the HMM classifier had to assign each test token to one of the following stop classes [d,g,b,t,k,p,dx]. On 6332 stops the wavelet features obtained an overall accuracy of 86% which corresponds to a 14% relative error reduction compared to the STFT baseline system. Furthermore an analysis of the HMM misclassifications revealed that voiced stops were highly confused with their voiceless unaspirated counterparts.

**Single Frame Selection for Phoneme Classification**

**Tingyao Wu, Dirk Van Compernolle, Jacques Duchateau, Hugo Van hamme; Katholieke Universiteit Leuven, Belgium**

Mon3CaP.2

Our former study [1] has shown that maximum likelihood (ML) based frame selection, which selects reliable frames from a high resolution along the time axis, helps to improve the discrimination between phonemes. In this paper, we present our recent research on single frame selection for a phoneme classification task. A new single selection, which only selects one frame for one state in an Hidden Markov Model (HMM), is proposed. The new technique takes likelihoods of frames and their positions in a phoneme segment into account at the same time, and selects very few frames to represent the spectral evolution of the phoneme. Furthermore, we also show that for a low model complexity, a phoneme model trained by selected frames is more discriminative than a model using all frames.
Discriminating Speaking Styles

Goshu Nagino, Makoto Shozakai; Asahi Kasei Corporation, Japan

Mon3CaP.6

Discriminating speaking styles is an important issue in speech recognition, speaker recognition and speaker segmentation. This paper compares distance measures between Gaussian distributions for discriminating speaking styles. The Mahalanobis distance, the Bhattacharyya distance and the Kullback-Leibler divergence, which are in common use for a definition as a distance measure between Gaussian distributions, are evaluated in terms of an accuracy to discriminate speaking styles. In this paper, the accuracy is judged on a visualized map, where speaking style speech corpora are mapped onto two-dimensional space by utilizing a multidimensional scaling method. It is shown that speaking style clusters appear clearly grouped on the visualized map obtained by the Bhattacharyya distance and the Kullback-Leibler divergence. In addition, the visualized map corresponds to speech recognition performance, and the Kullback-Leibler shows higher sensitivity to recognition performance.

Bayesian Networks for Phonetic Classification Using Time-Scale Features

Franz Perrnkopf, Tuan Van Pham; Graz University of Technology, Austria

Mon3CaP.7

We present a phonetic classification approach based on Bayesian networks using time-scale features which are extracted from the discrete Wavelet transform. We apply Bayesian networks using discriminative and generative parameter and/or structure learning for classifying the speech frames into silence, voiced, unvoiced, mixed sounds, and two more categories, voiced closure and release of plosives. Gender dependent/independent experiments have been performed on the TIMIT database. The experiments show that (i) our time-scale features mostly outperform standard MFCC features, (ii) discriminative learning of Bayesian networks is superior to the generative approach.

Fast and Effective Retraining on Contrastive Vocal Characteristics with Bidirectional Long Short-Term Memory Nets

Nicole Beringer; 3SOFT GmbH, Germany

Mon3CaP.8

We apply Long Short-Term Memory (LSTM) recurrent neural networks to a large corpus of unprompted speech — the German part of the VERBMOBIL corpus. By training first on a fraction of the data, then retraining on another fraction, we both reduce time costs and significantly improve recognition rates. Contrastive retraining on the initial vowel cluster fraction of the data according to the Psycho-Computational Model of Sound Acquisition (PCMSA) shows higher frame by frame correctness due to more sparseness and the articulatory position of the sounds. For comparison we show recognition rates of Hidden Markov Models (HMMs) on the same corpus, and provide a promising extrapolation for HMM-LSTM hybrids.

Exploiting Dendritic Autocorrelogram Structure to Identify Spectro-Temporal Regions Dominated by a Single Sound Source

Ning Ma, Phil Green, André Coy; University of Sheffield, UK

Mon3CaP.9

Autocorrelograms exhibit tree-like structures whose spines are located at a delay of 1/F0. This paper exploits the dendritic autocorrelogram structure for the identification of spectro-temporal regions dominated by a single periodic sound source in monaural acoustic mixtures. Each frame of the mixture is first segmented into different sound sources in the autocorrelogram domain. Local pitch estimates are formed for each source and used as a cue for temporal integration. A confidence score is computed for each time-frequency pixel in the grouped regions to determine its probability of belonging to the group. The system is evaluated using simultaneous speech in a coherence measuring experiment and also employed within an ASR system where it produces improved results for the Interspeech 2006 Speech Separation Challenge.
Locating Phone Boundaries from Acoustic Discontinuities Using a Two-Staged Approach
Pairote Leelaphattarakij, Proadpran Punyabukkana, Atiwong Suchato; Chulalongkorn University, Thailand
Mon3CaP.10
Ability to automatically align phonetic transcriptions with their associated acoustic signal is crucial to the development of computer-assisted speech training system where, it is frequently needed to locate phone boundaries from a known transcription in the signal. In this paper, an attempt to locate phone boundaries in the case where only the numbers of boundaries are known is described. Phone boundaries are hypothesized based solely on acoustic discontinuities without knowing exact transcriptions. This allows speech segmentation to be performed without large number of in-domain speech data for training. The boundary identification is done in two stages. Candidates for possible boundaries are selected in the first stage via local maxima of spectral changes. Dynamic programming is used to search for the best locations of phone boundaries from the candidate list. Allowing at most 20 ms. deviation from the actual boundaries, approximately 75% accuracy is achieved on a Thai continuous speech corpus.

Investigation on Rescoring Using Minimum Verification Error (MVE) Detectors
Qiang Fu, Biing-Hwang Juang; Georgia Institute of Technology, USA
Mon3CaP.11
Discriminative training, especially Minimum Verification Error (MVE) method plays an important role in the detection-based ASR. Recently, discriminative training also has been shown to be effective in large vocabulary continuous speech recognition [1]. In this paper, we propose a rescoring framework to show the improvement by fusing MVE-trained detectors with a conventional recognizer. The recognizer performs regular Viterbi decoding, generating possible recognition candidates with corresponding likelihood in a fashion of either N-best lists or word graphs. Detectors trained under MVE criterion form and conduct hypothesis testing for all test tokens to accomplish additional scores. A number of linear or non-linear rescoring methods are then presented to combine these two groups of scores. The experiments were carried out on the TIMIT database, and the results indicates that combining based on word graphs outperforms the one on N-best lists in the final accuracy. This rescoring framework explores possible ways to combine other independent knowledge sources with a conventional recognizer. Further more, it can guide the future research of the pure detection-based ASR techniques.

Generalization of the Minimum Classification Error (MCE) Training Based on Maximizing Generalized Posterior Probability (GPP)
Qiang Fu1, Antonio Moreno-Daniel1, Biing-Hwang Juang1, Jian-Lai Zhou2, Frank K. Soong2; 1Georgia Institute of Technology, USA; 2Microsoft Research Asia, China
Mon3CaP.12
In this paper, we generalize the training error definitions for minimum classification error (MCE) training and investigate their impact on recognition performance. Starting the conventional MCE method, we discuss the three issues in regard to training error definition, which may affect the recognizer performance and need to be extensively studied. We focus our discussions on the first two aspects in this paper. We re-visit the fact that the objective function in MCE training can be formulated into an equivalent form for maximizing the “posterior probability” of the corresponding training units. Based on the framework of the generalized posterior probability (GPP) [1], we design experiments to demonstrate effects about different training units and different constraints on segmentation boundaries for the MCE training. We also provide a performance analysis to illustrate our generalization for both phone recognition and word recognition tasks based on the Wall Street Journal (WSJ0) [2, 3] database.

Unsupervised Detection of Whispered Speech in the Presence of Normal Phonation
Michael A. Carlin1, Brett Y. Smolenski2, Stanley J. Wenndt2; 1Temple University, USA; 2Research Associates for Defense Conversion, USA; 1Air Force Research Laboratory, USA
Mon3CaP.13
The results of an investigation into unsupervised detection of whispered speech segments in the presence of normally phonated speech are presented. The Whispered Speech Detection system presented here extracts features which exploit both waveform energy and periodicity. Unsupervised classification of these features was performed to identify and label long segments (approx. 2 – 2.5 seconds) of whispered speech which is typically an indication of criminal activity over telephone networks, for instance, in a correctional facility environment. Experiments indicate that it is possible to automatically detect long segments of whispering in the presence of normally phonated speech; testing of the algorithm presented in this paper yields promising results in correct identification of whispered speech segments.

Friends and Enemies: A Novel Initialization for Speaker Diarization
Xavier Anguera1, Chuck Wooters1, Javier Hernando2; 1International Computer Science Institute, USA; 2Universitat Politècnica de Catalunya, Spain
Mon3CaP.14
The task of speaker diarization consists of answering the question “Who spoke when?”. The most commonly used approach to speaker diarization is agglomerative clustering of multiple initial clusters. Even though the initial clustering is greatly modified by iterative cluster merging and possibly multiple resegmentations of the data, the initialization algorithm is a key module for system performance and robustness. In this paper we present a novel approach that obtains a desired initial number of clusters in three steps. It first computes possible speaker change points via a standard technique based on the Bayesian information criterion (BIC). It then classifies the resulting segments into “friend” and “enemy” groups to finally creates an initial set of clusters for the system. We test this algorithm with the dataset used in the RT05s evaluation, where we show a 13% Diarization error rate relative improvement and a 2.5% absolute cluster purity improvement with respect to our previous algorithm.

Mon3FoP : Linguistics, Phonology, and Phonetics I
Poster Session, Allegheny Ballroom Foyer, 16:30 – 18:30
Chair: Susan R. Hertz, NovaSpeech LLC, USA

Acoustic Cues for the Classification of Regular and Irregular Phonation
Kushan Surana, Janet Slifka; Massachusetts Institute of Technology, USA
Mon3FoP.1
Irregular phonation serves an important communicative function. It can be a cue to linguistic contrasts, and often serves as a marker for word and utterance boundaries. Automatic methods for classification and detection of regions of irregular phonation can be used to improve analyses of occurrences of irregular phonation and support technologies such as speech recognition and synthesis. This study proposes a set of acoustic cues from both the temporal and frequency domains — fundamental frequency, normalized RMS amplitude, smoothed-energy-difference amplitude and shift-difference amplitude
— for separation of regions of regular and irregular phonation. Tokens from the TIMIT database are classified using support vector machines, trained on 114 different speakers and tested with 37 different speakers. Both genders are well represented in the data set and the tokens occur in various contexts within the utterance. In the test set, 292 of 320 irregular tokens (recognition rate of 91.25%), and 4105 of 4320 regular tokens (recognition rate of 95.02%) are correctly identified.

Realizations and Representations of Thai Tones in Monomoraic Syllables
Rattima Nittisaroj; Georgetown University, USA
Mon3FoP.2
This study investigates the phonetic realizations of Thai tones in monomoraic syllables produced with three different speeds. The results do not support previous claim that tones in monomoraic syllables are neutralized to mid, and lend support to a phonological representation in which the mora is associated to either H or no tone at all. In normal speech, a /AI/ vs. /O/ opposition is found. In slow speech, a /AI/ vs. /L/ distinction is found. Past research used the consonant clusters observed in slow speech as evidence to posit the underlying high and low tones for monomoraic syllables. In the present study, it is argued that the moraic association of the phonological low tone takes place as a result of glottal stop insertion.

Measuring and Comparing Vowel Qualities in a Dutch Spontaneous Speech Corpus
Irene Jacobi, Louis C.W. Pols, Jan Stroop; University of Amsterdam, The Netherlands
Mon3FoP.3
Recent studies of spoken Standard Dutch support an ongoing change in the phonetic quality of the diphthong /ɛI/ [1, 2]. However, there is a need for broader analyses and larger data sets. Here, we took Dutch vowel variants of 44 speakers from a spoken Dutch speech corpus, the CGN [3]. The vowels were measured and compared on the basis of 15,000 vowel segments, consisting of productions of /ɛI/, /aI/, /ɛI/, /o:/, and /e:/, as well as the anchor vowels /a/, /æ/, /u/. It was our aim to analyze changes in vowel quality dependent on the speakers’ sociological backgrounds and ages, and to deal with the variable recording qualities of the corpus. All vowels were taken from spontaneously uttered sentences and were analyzed automatically by means of a principal component analysis (PCA) on the vowels’ bark-filtered spectra, as well as by formant analysis.

Recalculating spectral positions in the principal components (pc’s) plane displayed the spectral interaction of the first formants in the pc1-pc2 plane, and explained the better separability of the vowels compared to the F1-F2 plane, as well as the high correlation of the first three formants with pc1 and pc2. The first pc’s turned out to be rather insensitive to sex-differences, but they were sensitive to the signal-to-noise ratio of the speech data. Variable recording qualities manifested themselves in speaker-specific location and size of the vowel spaces. Good signal-to-noise ratios could be transformed to poorer signals by increasing the lowest possible dB values per filter. Having analyzed the influence of noise on our data, we could normalize the data by taking each speaker’s /a-I-u/ positions and the focal point as references for better inter-speaker comparison.

The results clearly show different vowel quality patterns dependent on the speakers’ education and age, and indicate a progress of quality changes with as parameters the lowering and the degree of diphthongization of the long vowels and diphthongs.

Phonetic Research on Accented Chinese in Three Dialectal Regions: Shanghai, Wuhan and Xiamen
Aijun Li, Qiang Fang, Ziyu Xiong; Chinese Academy of Social Sciences, China
Mon3FoP.4
There are 10 major dialects in China. Most people in dialectal regions are bilingual speakers, i.e. native dialect and Mandarin. Although lots of people can speak Mandarin, they speak it with different accents (called regional accented Chinese in this paper) depending on how well they grasp the language. In this study, we categorize the regional accented Chinese into 3 levels of accents according to phonetic annotation and subjective evaluation on a regional accented speech corpus of three regions: Shanghai, Wuhan and Xiamen. Three accent evaluation methods, namely segmental annotation, clustering on phonetic annotation and subjective evaluation, are compared based on phonetic error rates. The results show that objective evaluation score based on segmental pronunciation is higher than subjective evaluation for the same speaker. This implies that supra-segmental features play an important role in rating accent degree and segmental features alone are not enough for objective evaluation. In accent level criterion, the errors from prosodic and segmental aspects are not equal and the percentage of these two parts are various for different regional speakers. The result is helpful for machine evaluation, L2 teaching and acquisition.

Pronunciation Variation Modeling for Mandarin with Accent
Chi Zhang, Ji Wu, Xi Xiao, Zuoying Wang; Tsinghua University, China
Mon3FoP.5
In order to solve the problem of the performance decrease when state-of-art automatic speech recognition (ASR) system facing accent speech, we propose the Pronunciation Variation Model (PVM). Two approaches are proposed to construct the PVM in this paper. 6.38% and 7.78% relative error rate reduction is achieved for Shanghai and Wuhan accent mandarin, respectively. The experiment on these two typical accent mandarin shows it is a possible way to deal with accent speech.

Specificity and Generalizability of Spontaneous Phonetic Imitation
Kuniko Y. Nielsen; University of California at Los Angeles, USA
Mon3FoP.6
The imitation paradigm [1, 2], in which subjects’ speech is compared before and after they are exposed to target speech, has shown that subjects shift their production in the direction of the target, indicating the use of episodic traces in speech perception as well as the close tie between speech perception and production. By using this paradigm, the current study aims to investigate the psychological reality of three levels of linguistic unit (i.e., word, phoneme, and feature). An experiment was designed to test whether spontaneous phonetic imitation can be generalized from words across (a) new words which share the same initial phoneme, and (b) new words with a new phoneme falling in the same natural class (sharing a feature); and also whether word-level specificity can be obtained through physical measurements instead of perceptual assessments. The feature manipulated in the experiment was aspiration, or [+spread glottis], on the phonemes /p/ and /k/.

The results showed that subjects produced significantly longer VOTs after they were exposed to target speech with longer VOTs, replicating [2] in a non-shadowing paradigm. Furthermore, the modeled feature (increased aspiration) was generalized to new instances of /p/ (i.e., in new words) as well as to the new segment /k/. At the same time, subjects’ post-exposure VOT was significantly longer for those items that were in the target speech than items which they had not previously heard. These results, taken together, indicate that speakers possess both sub-phonemic and word-level representations.
On the Sufficiency of Automatic Phonetic Transcriptions for Pronunciation Variation Research
Christophe Van Bael, Hans van Halteren; Radboud Universiteit Nijmegen, The Netherlands
Mon3FoP.7
We investigated whether automatic phonetic transcriptions (APTs) can replace manually verified phonetic transcriptions (MPTs) in a large corpus-based study on pronunciation variation. To this end, we compared the performance of both transcription types in a classification experiment aimed at establishing the direct influence of a particular situational setting on pronunciation variation. We trained classifiers on the speech processes extracted from the alignments of an APT and an MPT with a canonical transcription. We tested whether the classifiers were equally good at verifying whether unknown transcriptions represent read speech or telephone dialogues, and whether the same speech processes were identified to distinguish between transcriptions of the two situational settings. Our results not only show that similar distinguishing speech processes were identified; our APT-based classifier yielded better classification accuracy than the MPT-based classifier whilst using fewer classification features.

Automatic Detection of Voice Onset Time Contrasts for Use in Pronunciation Assessment
Abe Kazemzadeh1, Joseph Tepperman1, Jorge Silva1, Hong You2, Sungbok Lee1, Abeer Alwan2, Shrikanth Narayanan1; 1University of Southern California, USA; 2University of California at Los Angeles, USA
Mon3FoP.8
This study examines methods for recognizing different classes of phones from accented speech based on voice onset time (VOT). These methods are tested on data from the Tball corpus of Los Angeles-area elementary school children [1]. The methods proposed and tested are: 1) to train models based on standard English VOT contrasts and then extract the VOT characteristics of the phones by measuring the duration of phone-level and sub-phone-level alignments, 2) to train phone models with explicit aspiration, and 3) to train different models for different phoneme classes of VOT times. Error rates of 23–53% for different phone classes are reported for the first method, 5–57% for the second method, and 0–36% for the third. The results show that different methods work better on different phone classes. We interpret these results in relation to past research on VOT, explain possible uses for these findings, and propose directions for future research.

Unfilled Pauses in Japanese Sentences Read Aloud by Non-Native Learners
Hiroko Hirano1, Goh Kawai2, Keikichi Hirose1, Nobuaki Minematsu1; 1University of Tokyo, Japan; 2Hokkaido University, Japan
Mon3FoP.9
Perception experiments suggest that natives judge non-native unfilled pauses as indiscriminate and indecisive. Multiple regression analyses of unfilled pauses indicate a connection between syntactic structure and pause location and duration. Native speakers uniformly pause at large syntactic breaks with marked duration, whereas non-natives’ unfilled pauses are spread over various locations, possibly reflecting limited syntactic planning. Our method might be used to synthesize appropriate unfilled pauses in text-to-speech systems, and to train pausing behavior in automated pronunciation learning systems for non-native learners.

Detection of Quotations and Inserted Clauses and Its Application to Dependency Structure Analysis in Spontaneous Japanese
Ryoji Hamabe1, Kiyotaka Uchimoto2, Tatsuya Kawahara1, Hitoshi Ishahara2; 1Kyoto University, Japan; 2NICT, Japan
Mon3FoP.10
Japanese dependency structure is usually represented by relationships between phrasal units called bunsetsus. One of the biggest problems with dependency structure analysis in spontaneous speech is that clause boundaries are ambiguous. This paper describes a method for detecting the boundaries of quotations and inserted clauses and that for improving the dependency accuracy by applying the detected boundaries to dependency analysis. The clauses and inserted clauses are determined by using an SVM-based text chunking method that considers information on morphemes, pauses, fillers, etc. The information on automatically analyzed dependency structure is also used to detect the beginning of the clauses. Our evaluation experiment using Corpus of Spontaneous Japanese (CSJ) showed that the automatically estimated boundaries of quotations and inserted clauses helped to improve the accuracy of dependency structure analysis.

Chinese Input Method Based on Reduced Mandarin Phonetic Alphabet
Chun-Han Tseng, Chia-Ping Chen; National Sun Yat-Sen University, Taiwan
Mon3FoP.11
In this paper we study the problem of simplifying Chinese input method and making it suitable for use with mobile devices. To see the feasibility of aggressively reducing the number of keystrokes per Chinese character, we compare three input modes: character-based, syllable-based and first-symbol-based. Specifically, we use these linguistic units as token types and compare the perplexities. With the language model trained by data based on the ASBC corpus, the perplexity of the data set we collect from on-line chat and instant messages is 102.6 for character-based model, 67.7 for syllable-based model and 16.3 for first-symbol-based model. Arguing from the relation between the perplexity and the number of “typical” sentences of a language model, our conclusion is that on average there are 6 to 7 characters per first-symbol in natural Chinese language.

Thesaurus Expansion Using Similar Word Pairs from Patent Documents
Yoshimi Suzuki, Fumiyo Fukumoto; University of Yamanashi, Japan
Mon3FoP.12
In both written and spoken languages, we sometimes use different words in order to describe the same meaning. For instance, we use “constraint” (seigen) and “restriction” (seiyaku) as the same meaning. This makes text classification and text summarization difficult. In order to deal with this problem, dictionaries especially thesauri are used. However, in technical paper and patent documents, a lot of new words which are not given in the dictionary. In this paper, we propose a method to accurately extract words which are semantically similar to each other. Using this method, we extracted similar word pairs from patent documents. We also expand a thesaurus using the extracted similar words.

Low-Resource Autodiacritization of Abjads for Speech Keyword Search
Patrick Schone; U.S. Department of Defense, USA
Mon3FoP.13
Keyword search in speech requires retrieval systems to know the pronunciation of keywords. Many languages of the world are either largely alphabetic or have pronouncing dictionaries so that deducing pronun-
cations at run-time is manageable. There are many under-resourced languages, though, with writing systems where only some of the vowels are represented in the orthography (i.e., “abjads”). The absence of vowels makes direct mapping of abjads to pronunciation non-trivial. We describe an automatic system for inferring pronunciations from abjadic languages which seamlessly integrates into an existing context-sensitive pronunciation generator that serves a language-universal keyword search system. We also identify Web resources and system performance for each of five abjadic languages: Arabic, Farsi, Hebrew, Pashto, and Urdu. We show that almost effortlessly, the system can learn new rules which increase pronunciation accuracies by as much as 31.2% relative.
Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources

Jason Riesa ¹, Behrang Mohit ², Kevin Knight ³, Daniel Marcu ³; ¹ Johns Hopkins University, USA; ² University of Pittsburgh, USA; ³ University of Southern California, USA

This paper presents an English-Iraqi Arabic speech-to-speech statistical machine translation system using limited resources. In it, we explore the constraints involved, how we endeavored to mitigate such problems as a non-standard orthography and a highly inflected grammar, and discuss leveraging existing plentiful resources for Modern Standard Arabic to assist in this task. These combined techniques yield a reduction in unknown words at translation time by over 40% and a +3.65 increase in BLEU score over a previous state-of-the-art system using the same parallel training corpus of spoken utterances.

A Phrase-Level Machine Translation Approach for Disfluency Detection Using Weighted Finite State Transducers

Sameer Maskey ¹, Bowen Zhou ², Yuqing Gao ²; ¹ Columbia University, USA; ² IBM T.J. Watson Research Center, USA

We propose a novel algorithm to detect disfluency in speech by formulating the problem as phrase-level statistical machine translation using weighted finite state transducers. We approach the task as translation of noisy speech to clean speech. We simplify our translation framework such that it does not require fertility and alignment models. We tested our model on the Switchboard disfluency-annotated corpus. Using an optimized decoder that is developed for phrase-based translation at IBM, we are able to detect repeats, repairs and filled pauses for more than a thousand sentences in less than a second with encouraging results.

Improving Phrase-Based Korean-English Statistical Machine Translation

Jonghoon Lee, Donghyeon Lee, Gary Geunbae Lee; POSTECH, Korea

In this paper, we describe several techniques to improve Korean-English statistical machine translation. We have built a phrase-based statistical machine translation system in a travel domain. On the baseline phrase-based system, several techniques are applied to improve the translation quality. Each technique can be applied or removed easily since the techniques are part of the preprocessing method or corpus processing method. Our experiments show that most of the techniques were successful except reordering the word sequence. The combination of the successful techniques has significantly improved the translation quality.

A Hybrid Phrase-Based/Statistical Speech Translation System

David Stallard, Fred Choi, Krste Krstovski, Prem Natarajan, Rohit Prasad, Shirin Saleem; BBN Technologies, USA

Spoken communication across a language barrier is of increasing importance in both civilian and military applications. In this paper, we present a system for task-directed 2-way communication between speakers of English and Iraqi colloquial Arabic. The application domain of the system is force protection. The system supports translingual dialogue in areas that include municipal services surveys, detainee screening, and descriptions of people, houses, vehicles, etc. N-gram speech recognition is used to recognize both English and Arabic speech. The system uses a combination of a pre-recorded questions and statistical machine translation with speech synthesis to translate the recognition output.

High-Quality Speech Translation in the Flight Domain

Chao Wang, Stephanie Seneff; Massachusetts Institute of Technology, USA

Portability is an important issue to the viability of a domain-specific translation approach. This paper describes an English to Chinese translation system for flight-domain queries, utilizing an interlingua translation framework that has been successfully applied in the weather domain. Portability of various components is tested, and new technologies to handle parse ambiguities and ill-formed inputs are developed to enhance the translation framework. Evaluation of translation quality is conducted manually on a set of 432 unseen flight-domain utterances, which are translated into Chinese using a formal method and a new robust back-off method in tandem. We achieved 96.7% sentence accuracy with a rejection rate of 7.6% on manual transcripts, and 89.1% accuracy with an 8.6% rejection rate on speech input. A game for language learning using the translation capability is currently under development.

Optimizing Components for Handheld Two-Way Speech Translation for an English-Iraqi Arabic System

Roger Hsiao ¹, Ashish Venugopal ¹, Thilo Köhler ², Ying Zhang ¹, Paisarn Charoenpornsawat ¹, Andreas Zollmann ¹, Stephan Vogel ¹, Alan W. Black ¹, Tanja Schultz ¹, Alex Waibel ¹; ¹ Carnegie Mellon University, USA; ² Mobile Technologies LLC, USA

This paper described our handheld two-way speech translation system for English and Iraqí. The focus is on developing a field usable handheld device for speech-to-speech translation. The computation and memory limitations on the handheld impose critical constraints on the ASR, SMT, and TTS components. In this paper we discuss our approaches to optimize these components for the handheld device and present performance numbers from the evaluations that were an integral part of the project. Since one major aspect of the TransTac program is to build fieldable systems, we spent significant effort on developing an intuitive interface that minimizes the training time for users but also provides useful information such as back translations for translation quality feedback.
Tue1A20 : Acoustic Modeling II — Adaptation

Oral Session, Allegheny Ballroom, center, 10:30 - 12:30
Chair: Jean-Luc Gauvain, LIMSI, France

Distant-Talking Continuous Speech Recognition Based on a Novel Reverberation Model in the Feature Domain
Armin Sehr, Marcus Zeller, Walter Kellermann; Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
Tue1A20.1
A novel approach for automatic speech recognition in highly reverberant environments, proposed in [1] for isolated word recognition, is extended to continuous speech recognition (CSR) in this paper. The approach is based on a combined acoustic model consisting of a network of clean speech HMMs and a reverberation model. Because the grammatical information and the information about the acoustic environment are strictly separated in the combined model, a high degree of flexibility for adapting the system to new tasks and new environments is attained. We show that virtually all known CSR search algorithms can be used for decoding the proposed combined model if a few extensions are added. In a simulation of a connected digit recognition task, the proposed method achieves more than 40% reduction of the word error rate compared to a conventional HMM-based system trained on reverberant speech, at the cost of an increased decoding complexity.

Robust Feature Space Adaptation for Telephony Speech Recognition
Xin Lei1, Jon Hamaker2, Xiaodong He2;1 University of Washington, USA;2 Microsoft Speech & Natural Language Group, USA
Tue1A20.2
Speaker adaptation is critical for modern speech recognition systems. Due to the computational and multi-channel model sharing considerations, the use of model adaptation techniques is limited in telephony speech recognition systems. On the other hand, feature space adaptation methods such as feature space maximum likelihood linear regression (fMLLR) are efficient approaches suitable for telephony systems. In this work, we first describe techniques for efficient implementation of online fMLLR adaptation. Then feature space maximum a posteriori linear regression (fMAPLR) is proposed to incorporate prior knowledge for the feature transform estimation and improve the robustness of the conventional fMLLR approach. Experiments on telephony data indicate that fMAPLR is significantly more robust than fMLLR, and outperforms fMLLR especially when the adaptation data is very limited.

A Simulated-Data Adaptation Technique for Robust Speech Recognition
Nattanun Thatphithakkul1, Boontee Krutarchue1, Chai Wuttiiwitchai2, Sanparith Marukatat2, Vataya Boonpiam2;1 King Mongkut's Institute of Technology Ladkrabang, Thailand;2 NECTEC, Thailand
Tue1A20.3
This paper proposes an efficient acoustic model adaptation method based on the use of simulated-data in maximum likelihood linear regression (MLLR) adaptation for robust speech recognition. Online MLLR adaptation is an unsupervised process which requires an input speech with phone labels transcribed automatically. Instead of using only the input signal in adaptation, our proposed simulated data method increases the size of adaptation data by adding noise portions extracted from the input speech to a set of pre-recorded clean speech, whose correct transcriptions are known. Various configurations of the proposed method are explored. Evaluations are performed with both additive and real noisy speech. The experimental results show that the proposed system achieves higher recognition rate than the system using only the input speech in adaptation and the system using a multi-conditioned acoustic model.

A New HMM Adaptation Approach for the Case of a Hands-Free Speech Input in Reverberant Rooms
Hans-Günter Hirsch, Harald Finster; Niederrhein University of Applied Sciences, Germany
Tue1A20.4
A new method is presented for adapting the HMMs of a speech recognition system to the condition of a hands-free speech input in a room environment. The reverberation in a room usually has a bad effect on the performance of a recognition system. Reverberation causes an artificial extension of acoustic excitations what gets visible as so called reverberation tail when looking at the envelope of the short-term energy over the whole frequency range or in subbands. The approach is based on the assumption that the acoustic excitation of a speech segment, as modeled by an HMM state, will be seen as attenuated versions at successive HMM states. Adding this attenuated excitations in the spectral domain at each HMM state leads to a considerable improvement of the recognition performance. Furthermore a new approach is presented to adapt the Delta parameters that are usually taken as additional acoustic features. The efficiency of both new techniques has been proved by some experiments on isolated and connected word recognition with the TIDigits speech data base.

A Vector Space Approach to Environment Modeling for Robust Speech Recognition
Yu Tsao, Chin-Hui Lee; Georgia Institute of Technology, USA
Tue1A20.5
We propose a vector space approach to characterizing environments for robust speech recognition. We represent a given environment by a super-vector formed by concatenating all the mean vectors of the Gaussian mixture components of the state observation densities of all hidden Markov models trained in the particular environment. New environment super-vectors can now be obtained either by an interpolation method with a collection of super-vectors trained from many real or simulated environments or by a transformation performed on an anchor super-vector for a specific environment, such as a clean condition. At a 5dB signal-to-noise (SNR) level, both interpolation- and transformation-based approaches achieve a significant error rate reduction of close to 47% from a baseline system with cepstral mean subtraction (CMS) with only two adaptation utterances. When incorporating N-best information to perform unsupervised adaptation at 5dB SNR with the same two utterances, we achieve a relative error reduction of about 40%, close to that achieved in the supervised mode.

Subspace Modeling and Selection for Noisy Speech Recognition
Jen-Tzung Chien, Chuan-Wei Ting; National Cheng Kung University, Taiwan
Tue1A20.6
This paper presents a new subspace modeling and selection approach for noisy speech recognition. In subspace modeling, we develop factor analysis (FA) for representing noisy speech. FA is a data generation model where the common factors are extracted with factor loading matrix and specific factors. We bridge the connection of FA to signal subspace (SS) approach. Interestingly, FA partitions noisy speech space into a principal subspace containing speech and noise and a minor subspace containing residual speech and residual noise. To estimate clean speech, we minimize the energies of speech distortion in principal subspace as well as minor subspace. More importantly, in
subspace selection, we explore optimal subspace partition via solving hypothesis test problems. We test the equivalence of eigenvalues in minor subspace so as to determine subspace dimension. To fulfill FA spirit, we further examine the hypothesis of uncorrelated residual speech. Optimal solutions are realized through likelihood ratio test with the approximated chi-square distributions as test statistics. Subspace partition is performed according to the confidence towards rejecting null hypotheses. In the experiments on Aurora2 database, FA outperforms S5 in subspace modeling. New selection algorithms effectively determine subspace dimension for noisy speech recognition.

Tue1A3O : Emotional Speech and Speaker State
Oral Session, Allegheny Ballroom, right, 10:30 - 12:30
Chair: Helen M. Meng, Chinese University of Hong Kong, China

Recognition of Interest in Human Conversational Speech
Björn Schuller, Niels Köhler, Ronald Müller, Gerhard Rigoll; Technische Universität München, Germany

tue1A3O.1 Recognition of interest of a speaker within a human dialog bears great potential in many commercial applications. Within this work we therefore introduce an approach that analyses acoustic and linguistic cues of a spoken utterance. A systematic generation of more than 3k hi-level features basing on prosodic and spectral feature contours by means of descriptive statistical analysis and subsequent feature space optimization is used to find relevant acoustic attributes. For linguistic information integration a bag-of-words representation is used relying on a speech recognizer's output. One main aspect is the database of more than 2k spontaneous sub-speaker turns recorded and annotated for this analysis. Several influence factors as microphone distance and ASR versus annotation of spoken content are discussed. Overall remarkable performance of a running prototype can be reported discriminating between three levels of interest.

Using System and User Performance Features to Improve Emotion Detection in Spoken Tutoring Dialogs
Hua Ai, Diane J. Litman, Kate Forbes-Riley, Mihai Rotaru, Joel Tetreault, Amruta Purandare; University of Pittsburgh, USA

tue1A3O.2 In this study, we incorporate automatically obtained system/user performance features into machine learning experiments to detect student emotion in computer tutoring dialogs. Our results show a relative improvement of 2.7% on classification accuracy and 8.08% on Kappa over using standard lexical, prosodic, sequential, and identification features. This level of improvement is comparable to the performance improvement shown in previous studies by applying dialog acts or lexical/prosodic/discourse-level contextual features.

Real-Life Emotions Detection with Lexical and Paralinguistic Cues on Human-Human Call Center Dialogs
Laurence Devillers, Laurence Vidrascu; LIMSI, France

tue1A3O.3 The emotion detection work reported here is part of a larger study aiming to model user behavior in real interactions. We already studied emotions in a real-life corpus with human-human dialogs on a financial task. We now make use of another corpus of real agent-caller spoken dialogs from a medical emergency call center in which emotion manifestations are much more complex, and extreme emotions are common. Our global aims are to define appropriate verbal labels for annotating real-life emotions, to annotate the dialogs, to validate the presence of emotions via perceptual tests and to find robust cues for emotion detection. Annotations have been done by two experts with twenty verbal classes organized in eight macro-classes. We retained for experiments in this paper four macro classes: Relief, Anger, Fear and Sadness. The relevant cues for detecting natural emotions are paralinguistic and linguistic. Two studies are reported in this paper: the first investigates automatic emotion detection using linguistic information, whereas the second investigates emotion detection with paralinguistic cues. On the medical corpus, preliminary experiments using lexical cues detect about 78% of the four labels showing very good detection for Relief (about 90%) and Fear (about 86%) emotions. Experiments using paralinguistic cues show about 60% of good detection, Fear being best detected.

Real vs. Acted Emotional Speech
Janneke Wilting, Emiel Krahmer, Marc Swerts; Tilburg University, The Netherlands

tue1A3O.4 Even though the use of actors is a popular method for researching the expression of emotion, little is known about the relation between acted and real emotions. To shed some light on this, we set up a novel experiment, based on the Velten mood induction procedure, during which participants have to utter pre-defined sentences with a strong emotional content. In one group of participants, real positive or negative emotions were induced, while another group was instructed to act positive or negative while uttering Velten sentences. Results of a mood questionnaire revealed that participants in the real emotion condition, indeed felt positive or negative, depending on whether they read positive or negative sentences, while participants in the acted emotion condition felt neutral afterwards. In a second, perception experiment, it was found that acted emotions (especially negative ones) were perceived more strongly than the real emotions. This suggests that actors do not feel the acted emotion, and may engage in overacting, which casts doubt on the usefulness of actors as a way to study real emotions.

Emotion Recognition in Spontaneous Speech Using GMMs
Daniel Nelberg, Kjell Elenius; Kornel Laskowski; 1 KTH, Sweden; 2 Carnegie Mellon University, USA

tue1A3O.5 Automatic detection of emotions has been evaluated using standard Mel-frequency Cepstral Coefficients, MFCCs, and a variant, MFCC-low, calculated between 20 and 300 Hz, in order to model pitch. Also plain pitch features have been used. These acoustic features have all been modeled by Gaussian mixture models, GMMs, on the frame level. The method has been tested on two different corpora and languages; Swedish voice controlled telephone services and English meetings. The results indicate that using GMMs on the frame level is a feasible technique for emotion classification. The two MFCC methods have similar performance, and MFCC-low outperforms the pitch features. Combining the three classifiers significantly improves performance.

Personality Factors in Human Deception Detection: Comparing Human to Machine Performance
Frank Enos, Stefan Benus, Robin L. Cautin; Martin Graciareno, Julia Hirschberg, Elizabeth Shriberg; 1 Columbia University, USA; 2 Manhattanville College, USA; 3 SRI International, USA

tue1A3O.6 Previous studies of human performance in deception detection have found that humans generally are quite poor at this task, comparing unfavorably even to the performance of automated procedures. However, different scenarios and speakers may be harder or easier to judge. In this paper we compare human to machine performance
detecting deception on a single corpus, the Columbia-SRI-Colorado
Corpus of deceptive speech. On average, our human judges scored
worse than chance — and worse than current best machine learning
performance on this corpus. However, not all judges scored poorly.
Based on personality tests given before the task, we find that several
personality factors appear to correlate with the ability of a judge to
detect deception in speech.

Tue1WeS : Speech and Language in Education
Special Session, Westmoreland Room, 10:30 - 12:30
Chair: Patti Price, PPRICE Speech and Language Technology
Consulting, USA

Developing an Automatic Assessment Tool for
Children’s Oral Reading
Leen Cleuren, Jacques Duchateau, Alain Sips, Pol
Ghesquière, Hugo Van hamme; Katholieke Universiteit
Leuven, Belgium
Tue1WeS.1
Automation of oral reading assessment and of feedback in a reading
tutor is a very challenging task. This paper describes our research aim-
ing at developing such automated systems. First topic is the recording
and annotation of CHOREC, the Flemish database of children’s oral
reading we develop in order to characterize oral reading processes
statistically. Next, we propose a classification of both oral reading
strategies and errors, which provides the basis of the envisaged
assessment and feedback. Finally, experimental results show that our
two-layered recognition system is able to provide high reading miscue
detection rates, while only few correctly read words are erroneously
tagged as miscue.

Prototyping a Call System for Students of Japanese
Using Dynamic Diagram Generation and Interactive
Hints
Christopher Waple, Yasushi Tsubota, Masatake Dantsuji,
Tatsuya Kawahara; Kyoto University, Japan
Tue1WeS.2
In this paper we discuss the concept and design of a new CALL (Computer Assisted Language Learning) system being developed to aid
students learning Japanese as a second language. The system is being
designed to allow students to create and speak their own sentences based on visual prompts, before receiving feedback on their mistakes.
The students may choose to receive guidance in order to complete
each task, selecting the level of help that best suits their needs. Having
described the concept of the system and its design, we discuss some
tests recently carried out using a prototype of the system, summarize
the results obtained, and give some thought as to the significance of
these results on the future development of the system.

A Multilingual Embodied Conversational Agent for
Tutoring Speech and Language Learning
Dominic W. Massaro1, Ying Liu2, Trevor H. Chen1,
Charles Perfetti2;1 University of California at Santa Cruz,
USA; 2University of Pittsburgh, USA
Tue1WeS.3
Speech and language science and technology evolved under the
assumption that speech was a solely auditory event. However, a bur-
gening record of research findings reveals that our perception and
understanding are influenced by a speaker’s face and accompanying
gestures, as well as the actual sound of the speech. Perceivers expertly
use these multiple sources of information to identify and interpret
the language input. Given the value of face-to-face interaction, our
persistent goal has been to develop, evaluate, and apply animated
agents to produce realistic and accurate speech. Baldi is an accurate
three-dimensional animated talking head appropriately aligned with
either synthesized or natural speech. Baldi has a realistic tongue and
palate, which can be shown by making his skin transparent.

Based on this research and technology, we have implemented
computer-assisted speech and language tutors for children with
language challenges and for all persons learning a second language.
Our language-training program utilizes Baldi (or his likeness) as the
conversational agent, who guides students through a variety of exer-
cises designed to teach vocabulary and grammar, to improve speech
articulation, and to develop linguistic and phonological awareness.
We have also implemented multilingual agents, using a client/server
architecture system. This system has been used to develop Bao, a
Mandarin talker, which has been used in an initial training study
for college students learning Mandarin as a new language. The
results address the potential for using visible speech technology and
pedagogy in language learning of both similar segments in the two
languages and new speech segments in the new language. Although
visible speech did not facilitate pronunciation learning relative to just
auditory speech, we expect that a more prolonged training period
would show an advantage of visible speech.

Some of the advantages of the Baldi pedagogy and technology include
the popularity and proven effectiveness of computers and embodied
conversational agents, the perpetual availability of the program, and
individualized instruction. The science and technology of Baldi holds
great promise in language learning, dialog, human-machine interaction,
education, and edutainment.

Classroom Success of an Intelligent Tutoring System
for Lexical Practice and Reading Comprehension
Michael Heilman, Kevyn Collins-Thompson, Jamie Callan,
Maxine Eskenazi; Carnegie Mellon University, USA
Tue1WeS.4
We present an intelligent tutoring system called REAP that provides
reader-specific lexical practice for improved reading comprehension.
REAP offers individualized practice to students by presenting authen-
tic and appropriate reading materials selected automatically from the
web. We encountered a number of challenges that must be met in
order for the system to be effective in a classroom setting. These in-
clude general challenges for a system that uses authentic materials, as
well as more specific challenges that arise from integrating the system
with pre-existing classroom curricula. We discuss how these challenges
were met, and present evidence that REAP has gained acceptance into
the classroom at the English Language Institute at the University of
Pittsburgh.

Assessing the Reading Level of Web Pages
Sarah E. Petersen, Mari Ostendorf; University of
Washington, USA
Tue1WeS.5
Reading is an important part of educational development. However,
finding appropriate reading material for all students can be difficult
and time consuming for teachers. Our goal is to automate the task
of assessing the reading level of text to enable teachers to more effec-
tively take advantage of the large amounts of text available today on the
World Wide Web. Reading level assessment tools already exist for clean
corpora such as books and magazine articles. This paper presents ex-
tensions of a particular set of tools to handle web pages returned by
a standard search engine, including a step that pre-filters web pages
to eliminate “junk” pages with little or no text. Results of applying the
reading level detectors to web pages are manually evaluated by elemen-
tary school teachers, the intended audience for these tools. The tools
work well for grades 4 and 5, with room for improvement in grades 2
and 3.
Is ASR Accurate Enough for Automated Reading Tutors, and How Can We Tell?
Jack Mostow; Carnegie Mellon University, USA
Tue1WeS.6
We discuss pros and cons of several ways to evaluate ASR accuracy in automated tutors that listen to students read aloud. Whether ASR is accurate enough for a particular reading tutor function depends on what ASR-based judgment it requires, the visibility of that judgment to students and teachers, and the amount of input speech on which it is based. How to tell depends on the purpose, criterion, and space of the evaluation.

Development of a Program for Self Assessment of Japanese Pronunciation by English Learners
Chiharu Tsurutani1, Yutaka Yamauchi2, Nobuaki Minematsu3, Dean Luo3, Kazutaka Maruyama3, Keikichi Hirose3; 1Griffith University, Australia; 2Tokyo International University, Japan; 3University of Tokyo, Japan
Tue1Wes.7
A program for self assessment of Japanese pronunciation by English-speaking learners was developed using a language model built with input from a language teacher in collaboration with speech engineers. This collaboration enhanced the program’s capacity for accurate assessment and provides practical support to users by linking evaluation with feedback, and an editorial function of error patterns. The program drew positive responses from participants in a trial run. This paper discusses the development of our language model, the function and evaluation of this self assessment program.

Pronunciation Verification of Children’s Speech for Automatic Literacy Assessment
Joseph Tepperman1, Jorge Silva1, Abe Kazemzadeh1, Hong You2, SungBok Lee1, Abeer Alwan2, Shrikanth Narayanan1; 1University of Southern California, USA; 2University of California at Los Angeles, USA
Tue1Wes.8
Arguably the most important part of automatically assessing a new reader’s literacy is in verifying his pronunciation of read-aloud target words. But the pronunciation evaluation task is especially difficult in children, non-native speakers, and pre-literate. Traditional likelihood ratio thresholding methods do not generalize easily, and even expert human evaluators do not always agree on what constitutes an acceptable pronunciation. We propose new recognition- and alignment-based features in a decision tree classification framework, along with the use of prior linguistic information and human perceptual evaluations. Our classification methods demonstrate a 91% agreement with the voted results of 20 human evaluators who agree among themselves 85% of the time.

Computer Aided Pronunciation Learning System Using Speech Recognition Techniques
Sherif Mahdy Abdou1, Salah Eldeen Hamid2, Mohsen Rashwan1, Abdurrahman Samir1, Ossama Abdel-Hamid1, Mostafa Shahin1, Waleed Nazih1; 1Cairo University, Egypt; 2Umm Al-Qura University, Saudi Arabia
Tue1Wes.9
This paper describes a speech-enabled Computer Aided Pronunciation Learning (CAPL) system HAFSS. This system was developed for teaching Arabic pronunciations to non-native speakers. A challenging application of HAFSS is teaching the correct recitation of the holy Qur’an. HAFSS uses a state of the art speech recognizer to detect errors in user recitation. To increase accuracy of the speech recognizer, only probable pronunciation variants, that cover all common types of recitation errors, are examined by the speech decoder. A module for the automatic generation of pronunciation hypotheses is built as a component of the system. A phoneme duration classification algorithm is implemented to detect recitation errors related to phoneme durations. The decision reached by the recognizer is accompanied by a confidence score to reduce effect of misleading system feedbacks to unpredictable speech inputs. Performance evaluation using a data set that includes 6.6% wrong speech segments showed that the system correctly identified the error in 62.4% of pronunciation errors, reported “Repeat Request” for 22.4% of the errors and made false acceptance of 14.9% of total errors.

An Information Theoretic Tool for Investigating Speech Perception
Bryce Lobdell, Jont B. Allen; University of Illinois at Urbana-Champaign, USA
Tue1BuP.1
A method for investigating human speech perception that combines information about human perception of speech, auditory modeling, signal detection theory, and information theory is described. For that purpose, a model for detection of signals in the auditory nerve is developed and used to analyze speech sounds. Examples are given that combine detectability information with P(heard = h | spoken = s) to access information about “cues” to some some speech sounds. These example agree qualitatively with previous results about perception of these sounds. Refinements and prospects for this approach are discussed in light of the examples.

An Adaptive Sampling Procedure for Speech Perception Experiments
Geoffrey Stewart Morrison; University of Alberta, Canada
Tue1BuP.2
Synthetic speech perception experiments may make use of several acoustic dimensions in order to adequately model listeners’ perception; however, the number of stimuli increases exponentially as dimensions are added. A relatively large number of identification responses per stimulus are needed in the vicinity of category boundaries in order to model the boundaries with reasonable accuracy. Fewer responses per stimulus are needed to model portions of the stimulus space where a single response category predominates. Rather than collecting the same number of responses for each stimulus, an experiment can therefore be shortened via adaptive sampling. An adaptive sampling procedure is described. After an initial pass through the stimuli, the procedure uses a logistic regression model to select stimuli to resample in subsequent rounds. Results of simulations indicated that the number of trials in the experiment could be reduced by a third without substantially affecting the results.

Disentangling Gestural and Auditory Contrast Accounts of Compensation for Coarticulation
Navin Viswanathan, James S. Magnuson, Carol A. Fowler; University of Connecticut, USA
Tue1BuP.3
Compensation for coarticulation (CFC), a context effect in which the articulatory characteristics of one segment influence the perception of a neighboring segment [1], has been a matter of considerable debate between proponents of gestural [2] and auditory theories of speech perception[3]. We set out to distinguish the two accounts by using non-native liquids (Tamil with American English listeners) that have distinct articulatory and acoustic characteristics from the native phoneme categories to which they are assimilated. We report three experiments that show that the auditory contrast account of CFC cannot explain compensa-
satory effects with our non-native stimuli. We argue that these context effects reflect perceptual compensation for coarticulation, as predicted on a gestural account, but discuss problems for both theories.

The Role of Positional Probability in the Segmentation of Cantonese Speech

Michael C.W. Yip; Open University of Hong Kong, China

This study investigated whether the steady-state suppression method was beneficial for native Cantonese listeners making use of probabilistic phonotactics information of words in the segmentation process of Cantonese continuous speech. Because some sounds appear more frequently at the beginning or ending of Cantonese syllables than the others, these kinds of probabilistic information of syllables may be likely to cue the locations of possible syllable boundaries in Cantonese continuous speech. A syllable-spotting experiment was conducted and the results indicated that native Cantonese listeners indeed made use of the positional probabilities of a syllable's onset but not for the case of a syllable's final in the segmentation process. Along with my previous study [1], I argue that probabilistic phonotactics is one useful source of information in Cantonese speech segmentation.

Nasality Perception of Vowels in Different Language Background

Shahina Haque, Tomio Takara; University of the Ryukyus, Japan

Nasality is a distinctive feature of Bangla vowels. In this paper, we describe our study on nasality perception of Bangla vowels by Bangla and Japanese listeners. We discuss an interesting result that Japanese listeners perceive most Bangla nasal /i/ as Japanese non-nasal /uː/. As the amount of nasalization of /i/ was increased synthetically, perception of this vowel change was found to be more categorical for Japanese listeners than that of Bangla listeners' vowel perception.

Steady-State Suppression in Reverberation: A Comparison of Native and Nonnative Speech Perception

Nao Hodoshima1, Dawn Behne2, Takayuki Arai1; 1Sophia University, Japan; 2Norwegian University of Science & Technology, Norway

This study investigated whether the steady-state suppression method proposed by Arai et al. (2001, 2002) improved consonant identification for nonnative listeners in reverberation. It also compared the effect of steady-state suppression on consonant identification by native and nonnative listeners in reverberation. We used steady-state suppression as a preprocessing technique which processes speech signals before they are radiated from loudspeakers in order to reduce the amount of overlap-masking. Participants were 24 native English (native listeners) and 24 Japanese speakers (nonnative listeners), both with normal hearing. A diotic Modified Rhyme Test was conducted with and without steady-state suppression for reverberation times of 0.4, 0.7 and 1.1 s and a non-reverberant condition. The results showed that native listeners performed better than nonnative listeners, and that the mean percentage of correct answers in initial consonants was higher than in final consonants. The results also showed that processed and unprocessed speech was comparable for word initial and final consonants. These findings indicate that parameters of steady-state suppression would need adjustment to accommodate speech materials and reverberant conditions. They also suggest that the difficulties that nonnative listeners have might not be due to the actual acoustic-phonetic information from the signal.

Effect of Dynamic Information of Formants on Discrimination of English Vowels in Consonantal Contexts by Japanese Listeners

Akiyo Joto; Prefectural University of Hiroshima, Japan

This study examined how differently native speakers of Japanese discriminated between the American English vowels /e/ and /æ/ in /CVP/ syllables with 20 different initial consonants, and how the differing discrimination was related to the formant changes throughout the vowels in comparison with the Japanese vowels /e/ and /aː/. A perceptual test and formant analyses of the English and Japanese vowels were conducted. The results showed that there were significant differences in discrimination across the consonantal contexts: the discrimination of /e/ was significantly poorer when the initial consonant was /dɡ/ or /g/, /ʃ/, /m/ or /θ/ (voiceless), and that of /æ/, when it was /ʃ/, /θ/ or /t/. It was found that the poorer discrimination was more related to the smaller formant ratios (F2/F1) and the higher F1 frequency in the latter part of /e/, and to the larger formant ratios and the lower F1 frequency in the latter part of /æ/. The changing of formant patterns throughout the two English vowels could be attributed to the poorer discrimination in the particular consonantal contexts.

Native and Nonnative Audio-Visual Perception of English Fricatives in Quiet and Cafe-Noise Backgrounds

Yue Wang1,2, Haisheng Jiang1, Chad Danyluck1,3, Simon Fraser University, Canada; 2Norwegian University of Science & Technology, Norway

This study examines audio-visual perception of second-language (L2) speech, with the goal of investigating the extent to which the auditory and visual input modalities are integrated in processing unfamiliar L2 speech. Native (Canadian English) and nonnative (Mandarin) perceivers responses were collected for a set of fricative-initial syllables presented with a quiet and a cafe-noise background, and presented in four ways: congruent audio-visual (AVc), incongruent audio-visual (AVi), audio-only (A) and visual-only (V). Results show that for both native groups, performance was better in the AVc condition than A or V condition; and better in quiet than in cafe-noise background. A comparison of the native and nonnative performance revealed that Mandarin participants showed (1) poorer identification of the L2 interdental fricatives, (2) a greater degree of reliance on visual information, even when auditory information was available, and (3) a higher percentage of McGurk responses with the incongruent AV speech. These findings indicate that although nonnatives were able to use visual information, they failed to adopt the visual cues that are linguistically characteristic of the L2 sounds, suggesting a language-specific AV processing pattern. However, similarities between the two native groups are also indicative of possible perceptual universals involved. Together they point to an integrated network in speech processing across modalities.

Perceptive and Acoustic Measurement of Average Speaking Pitch of Female and Male Speakers in German Radio News

Sven Grawunder1, Ines Bose2, Birgit Hertha2, Franziska Trauselt2, Lutz Christian Anders2; 1Max Planck Institute for Evolutionary Anthropology, Germany; 2Martin-Luther-Universität Halle-Wittenberg, Germany

The average pitch of 68 news broadcasters (34 female / 34 male speakers) was evaluated by 6 expert listeners. Additionally, the average fundamental frequency for all samples was analyzed by means of a series of standard pitch detection algorithms. The results show a strong correlation of acoustic mean and auditory median values for male voices, whereas the auditory mean values female voices are slightly higher than their acoustic counterparts. Vice versa interlistener reliability is higher for the evaluation of female speakers, whereas listeners report more difficulties in rating for male speaking voices.
Effects of Frequency Shifts on Perceived Naturalness and Gender Information in Speech

Peter F. Assmann\textsuperscript{1}, Sophia Dembling\textsuperscript{1}, Terrance M. Nearey\textsuperscript{2};\textsuperscript{1}University of Texas at Dallas, USA; \textsuperscript{2}University of Alberta, Canada

Tue1BuP.10

In natural speech, there is a moderate correlation between the fundamental frequency and formant frequencies across talkers. The present study used a high-quality vocoder to manipulate these properties and determine their contribution to perceived naturalness and voice gender. The stimuli were re-synthesized sentences spoken by two adult males and two adult females. Scale factors were chosen for each sentence and for each talker to produce frequency-shifted versions with a specified mean fundamental frequency ($F_0$) ranging from 60 Hz to 450 Hz in 10 steps, paired with 10 steps in geometric mean formant frequencies ranging from 850 Hz to 2500 Hz. Listeners judged frequency-shifted sentences as more natural when $F_0$ and formant frequencies followed the co-variation of $F_0$ and formant frequencies in natural voices. Sentences with low $F_0$s and low formant frequencies were perceived as masculine, while sentences with high $F_0$s and high formant frequencies were assigned high ratings of femininity. Sentences with “mismatched” $F_0$ and formant frequencies were assigned ratings near the midpoint of the range, indicating gender ambiguity. Frequency-shifted sentences derived from male talkers received consistently higher ratings of masculinity than those derived from females and vice versa, even when assigned scale factors appropriate for the opposite gender, indicating that factors other than $F_0$ and mean formant frequencies contribute to perceived gender.

Influence of Pause Length on Listeners’ Impressions in Simultaneous Interpretation

Hitomi Tohyama, Shigeki Matsubara; Nagoya University, Japan

Tue1BuP.11

We have been attempting to realize simultaneous machine interpretation. However, determining the interpreting utterance timing is as difficult as determining translation units. This remains a major concern for the development of such a speech translation system. It is also crucial for the system’s users that the speech generated by the system is clear and easy to listen to. In this paper, we focus attention on the pauses that partly characterize simultaneous interpreters’ utterances. We attempt to analyze the results of an experiment conducted using 31 subjects on the relationship between listener-friendliness and the number of pauses in speech, using the CIAIR simultaneous interpretation database as the data source. The results generated some knowledge about listener impressions of simultaneous interpretation, which will be helpful for realizing simultaneous machine interpretation.

New Measures to Chart Toddlers’ Speech Perception and Language Development: A Test of the Lexical Restructuring Hypothesis

Iris-Corinna Schwarz, Denis Burnham; University of Western Sydney, Australia

Tue1BuP.12

Language acquisition factors at work in toddlers between 2\frac{1}{2} and 3 years of age were investigated in the first longitudinal study of this kind. New age-appropriate tasks were devised to measure the development of vocabulary size; articulation accuracy; sensitivity to the phonemic features of, in this case, Australian English; and the degree of specialization towards the native tongue, as measured by language-specific speech perception; LSSP, with 45 Australian English learning toddlers (18 male, 27 female) at 30, 33, and 36 months of age. Results indicated (i) that nearly all measures improved linearly over age; (ii) that there were significant correlations between articulation ability and vocabulary size; and (iii) that, in confirmation of the lexical restructuring hypothesis, vocabulary size is significantly predicted by the broad range of native language abilities under the rubric of Phoneme Sensitivity, but not by the more specific measure of LSSP.

Perception of Fundamental Frequency in Cochlear Implant Patients

Ángel de la Torre\textsuperscript{1}, Cristina Roldán\textsuperscript{2}, Manuel Sainz\textsuperscript{2};\textsuperscript{1}Universidad de Granada, Spain; \textsuperscript{2}Hospital Universitario San Cecilio, Spain

Tue1BuP.13

Fundamental frequency is important in speech perception, since it provides prosodic information to emphasize key words and contributes to speaker identification. Additionally, fundamental frequency contains phonetic value in tonal languages. In the case of hearing impaired patients who have received a cochlear implant, the perception of the fundamental frequency is affected by the technical limitations associated to the procedure of auditory nerve stimulation. In this paper we analyze the possibilities and limitations of cochlear implant systems with respect to perception of the fundamental frequency.

Significance of Formants from Difference Spectrum for Speaker Identification

Kishore Prahallad\textsuperscript{1}, Sudhakar Varanasi\textsuperscript{2}, Ranganatham Veluru\textsuperscript{2}, Bharat Krishna M.\textsuperscript{2}, Debashish S. Roy\textsuperscript{2};\textsuperscript{1}Carnegie Mellon University, USA; \textsuperscript{2}Satyam Computer Services Ltd., India

Tue1CaP.1

In this paper, we describe a prototype speaker identification system using auto-associative neural network (AANN) and formant features. Our experiments demonstrate that formants extracted from difference spectrum perform significantly better than formants extracted from normal spectrum for the task of speaker identification. We also demonstrate that formants from difference spectrum provide comparable speaker identification performance with that of features such as weighted linear predictive Cepstral coefficients and Mel-Frequency Cepstral coefficients. Finally, we combine the results of formant based systems and linear predictive Cepstral coefficients based system to achieve 100% identification performance.

Using Genetic Algorithms to Weight Acoustic Features for Speaker Recognition

Maider Zamalloa\textsuperscript{1}, Germán Bordel\textsuperscript{1}, Luis Javier Rodríguez\textsuperscript{1}, Mikel Penagarikano\textsuperscript{1}, Juan Pedro Uribe\textsuperscript{2};\textsuperscript{1}Universidad del País Vasco, Spain; \textsuperscript{2}IKERLAN, Spain

Tue1CaP.2

The Mel-Frequency Cepstral Coefficients (MFCC) are widely accepted as a suitable representation for speaker recognition applications. MFCC are usually augmented with dynamic features, leading to high dimensional representations. The issue arises of whether some of those features are redundant or dependent on other features. Probably, not all of them are equally relevant for speaker recognition. In this work, we explore the potential benefit of weighting acoustic features to improve speaker recognition accuracy. Genetic algorithms (GAs) are used to find the optimal set of weights for a 38-dimensional feature set. To evaluate each set of weights, recognition error is measured over a validation dataset. Naive speaker models are used, based on empirical distributions of vector quantizer labels. Weighting acoustic features yields 24.58% and 14.68% relative error reductions in two series of speaker recognition tests. These results provide evidence that further improve-
ments in speaker recognition performance can be attained by weighting acoustic features. They also validate the use of GAs to search for an optimal set of feature weights.

Missing Feature Theory with Soft Spectral Subtraction for Speaker Verification

Michael T. Padilla\textsuperscript{1}, Thomas F. Quatieri\textsuperscript{2}, Douglas A. Reynolds\textsuperscript{2} ;\textsuperscript{1}Stanford University, USA; \textsuperscript{2}Massachusetts Institute of Technology, USA

Tue1CaP.3

This paper considers the problem of training/testing mismatch in the context of speaker verification and, in particular, explores the application of missing feature theory in the case of additive white Gaussian noise corruption in testing. Missing feature theory allows for corrupted features to be removed from scoring, the initial step of which is the detection of these features. One method of detection, employing spectral subtraction, is studied in a controlled manner and it is shown that with missing feature compensation the resulting verification performance is improved as long as a minimum number of features remain. Finally, a blending of “soft” spectral subtraction for noise mitigation and missing feature compensation is presented. The resulting performance improves on the constituent techniques alone, reducing the equal error rate by about 15% over an SNR range of 5 – 25 dB.

Prosodic Features for Speaker Verification

Leena Mary, Yegnanarayana B.; Indian Institute of Technology Madras, India

Tue1CaP.4

In this paper we study the effectiveness of prosodic features for speaker verification. We hypothesize that prosody is linked to linguistic units such as syllables and prosodic features can be better represented with reference to the syllabic sequence. For extracting prosodic features, speech is segmented into syllable-like regions using the knowledge of vowel onset points (VOP). We use a technique based on excitation source information to detect VOPs automatically. The location of VOPs serve as reference for extracting prosodic features directly from speech signal. Various parameters are used to represent the pitch and energy dynamics of the region between two consecutive VOPs. The effectiveness of the derived prosodic features for speaker verification is demonstrated on NIST SRE 2003 extended data. The complementary nature of prosodic features and spectral features help to improve the accuracy of the combined speaker verification system.

Unsupervised Learning of HMM Topology for Text-Dependent Speaker Verification

Ming Liu, Thomas S. Huang; University of Illinois at Urbana-Champaign, USA

Tue1CaP.5

Usually, text-dependent speaker verification can achieve better performance than text-independent system because of the constraint that the enrollment and testing utterance share the same phonetic content. However, the enrollment data for text-dependent system usually is very limited. Expectation Maximization (EM) training of HMM will suffer from noisy estimation because of limited enrollment. Adaptation is a popular solution in this scenario. The target model is formed by adapting the generic model based on limited speaker specific training data. Although the adaptation scheme can tolerate much less training data than direct EM method, the traditional method does not account the topology of HMM might be different for different speaker. The topology information further distinguish the target speaker from impostors. In this paper, we propose a unsupervised learning method to learn the topology of HMM for each speaker. The experimental results indicate that with learning the topology, the framework is more effective than traditional adaptation methods. In the pure acoustic matching experiments, the proposed method is the best system under extremely small amount enrollment data (1 training utterance) and moderate training data. That mainly due to explicitly including the label information in background modeling and discriminant capability of unsupervised learning of HMM topology.

On the Use of Jacobian Adaptation in Real Speaker Verification Applications

Jan Anguita, Javier Hernando; Universitat Politècnica de Catalunya, Spain

Tue1CaP.6

Jacobian Adaptation (JA) of the acoustic models is a fast adaptation technique that has been successfully used in both speech and speaker recognition systems. This technique adapts the acoustic models on the basis of the difference between the testing and the training noise conditions. For this reason, a noise reference of both the training and the testing phase is needed. In previous works, the noise conditions have been commonly supposed to be known, or estimated from the first part of the signals or using manually obtained speech and non-speech labels. However, in a real application the noise conditions are not generally known, it is not sure that the first part of the signal does not contain speech and manual labels are not usually available. In this work we propose to obtain the noise references by using continuous noise estimation methods, which are more appropriate for real applications. Several noise estimation methods are compared and the obtained results show that these techniques are effective for JA.

A Novel Framework of Text-Independent Speaker Verification Based on Utterance Transform and Iterative Cohort Modeling

Ming Liu\textsuperscript{1}, Huazhong Ning\textsuperscript{1}, Thomas S. Huang\textsuperscript{1}, Zhengyou Zhang\textsuperscript{2};\textsuperscript{1}University of Illinois at Urbana-Champaign, USA; \textsuperscript{2}Microsoft Research, USA

Tue1CaP.7

A novel framework for text-independent speaker verification is proposed. The framework is based on a new interpretation of Universal Background Model. The UBM in our framework actually defines a transform which maps the variable length observation into a fixed dimensional supervector (supervisor space). Each speech utterance is then mapped into a point in this supervector space. The similarity measure in this vector space is progressively refined via an iterative cohort modeling scheme. The experiments on NIST 2002 corpus show the effectiveness of this new framework. Overall the EER drops from the baseline system (with T-Norm) 9.21% to final improved system (without T-Norm) 8.07%. The new framework can effectively reduce the data dependence in the final output score which is clearly indicated in the second sets of experiments. The EER after T-Norm of final system marginally increases by relatively 1.73% compared to the EER of baseline system drops 16.12% relatively after T-Norm. Also, the relative improvement of DCF after T-Norm is marginal for the final improved system (2.47%) compared to 33.68% in baseline system. It clear shows that the iterative cohort modeling effectively reduce the data dependence of the final scores, so that T-Norm will not further improve the system performance. Also, the performance of novel frame clearly increases as the iteration grows which suggest that the framework progressively refine the similarity measure on the supervector space with the iterative cohort modeling.

A Cohort - UBM Approach to Mitigate Data Sparseness for In-Set/Out-of-Set Speaker Recognition

Vinoth Prakash, John H.L. Hansen; University of Texas at Dallas, USA

Tue1CaP.8

In this study, the problem of identifying in-set versus out-of-set speakers is addressed. Here the emphasis is on low enrolment and test data durations, in a text-independent mode. In order to compensate for the limited enrolment data (5 sec), a method is proposed that utilizes data from speakers that are acoustically close to a particular in-set speaker.
A speaker specific model is obtained by adaptation of a base model that is built using data from such speakers. The performance of the proposed algorithm is evaluated using the TIMIT database with an adapted GMM classifier (GMM-UBM) employed as the baseline system. Experimental results show a consistent increase in system performance, with a relative improvement ranging from 10.57–58.33% depending on in-set speaker size and test data duration.

**Analysis of Lombard Effect Under Different Types and Levels of Noise with Application to In-Set Speaker ID Systems**

Vaishnevi S. Varadarajan, John H.L. Hansen; University of Texas at Dallas, USA

**Tue1Cap.9**

This paper presents an analysis of Lombard speech produced under different types and levels of noise. The speech used for the analysis forms a part of the UT-SCOPE database and consists of sentences from the well-known TIMIT corpus, spoken in the presence of highway, large crowd and pink noise. Differences are shown to exist in the speech characteristics under these varying noise types. The deterioration of the EER of an in-set speaker identification system trained on neutral and tested with Lombard speech is also illustrated. A clear demarcation between the effect of noise and Lombard effect on noise is also given by testing with noisy Lombard speech. The test-token duration on system performance under the Lombard condition is addressed. It is seen that test duration has no effect on the EER under Lombard effect. The average EER for 3s test duration is 14.7, 28.3, 48.2, 51.3 for neutral clean, clean Lombard, noisy neutral and noisy Lombard respectively, and 7.2, 26.4, 45.8, 50.8 respectively for 12s test duration.

**Reducing Speech Coding Distortion for Speaker Identification**

Alan McCree; Massachusetts Institute of Technology, USA

**Tue1Cap.10**

In this paper, we investigate the degradation of speaker identification performance due to speech coding algorithms used in digital telephone networks, cellular telephony, and voice over IP. By analyzing the difference between front-end feature vectors derived from coded and uncoded speech in terms of spectral distortion, we are able to quantify this coding degradation. This leads to two novel methods for distortion compensation: codebook and LPC compensation. Both are shown to significantly reduce front-end mismatch, with the second approach providing the most encouraging results. Full experiments using a GMM-UBM speaker ID system confirm the usefulness of both the front-end distortion analysis and the LPC compensation technique.

**A Text-Prompted Distributed Speaker Verification System Implemented on a Cellular Phone and a Mobile Terminal**

Tsuneo Kato, Hisashi Kawai; KDDI R&D Laboratories Inc., Japan

**Tue1Cap.11**

For a practical application of biometrics authentication on cellular phones, a speaker verification system was implemented on a cellular phone and a PDA type mobile terminal. The system has the following three features: 1) a distributed system configuration for connectivity to internet services, 2) text-prompted speaker verification using connected digit patterns for robustness to imposture, and 3) incremental model update for preventing deterioration of accuracy. The system was examined in a two-month test, and effects of the incremental model update were evaluated with those data. Experimental results showed that the average equal error rate over this period was reduced from 8.6% to 4.4% by monthly update, to 2.9% by weekly update, and to 0.9% by daily update.

**Automatic Detection of Irregular Phonation in Continuous Speech**

Srikanth Vishnubhotla, Carol Y. Espy-Wilson; University of Maryland, USA

**Tue1Cap.12**

Voice quality is one of the most important source characteristics of a speaker’s speech production process, and creakiness is one of the variations of voice quality. This paper describes the development of an algorithm to automatically detect irregular phonation, including creakiness and other variations, in continuous running speech. The algorithm is an extension of the Aperiodicity, Periodicity and Pitch (APP) Detector. The algorithm has been run on 485 files of the TIMIT database, which contained 677 instances of irregular phonation. The test set comprised of 97 speakers, of which 57 were male and 40 were female. The algorithm has been found to give an accuracy of 86.7% on average, with performance being almost the same for both male and female speakers. Automatic detection of irregular phonation should help characterize speakers for speaker identification applications.

**Effects of Word Frequency on the Acoustic Durations of Affixes**

Mark Pluymaekers, Mirjam Ernestus, R. Harald Baayen; Radboud Universiteit Nijmegen, The Netherlands

**Tue1FoP.1**

This study investigates whether the acoustic durations of derivational affixes in Dutch are affected by the frequency of the word they occur in. In a word naming experiment, subjects were presented with a large number of words containing one of the affixes ge-, ver-, ont, or -lijk. Their responses were recorded on DAT tapes, and the durations of the affixes were measured using Automatic Speech Recognition technology. To investigate whether frequency also affected durations when speech rate was high, the presentation rate of the stimuli was varied. The results show that a higher frequency of the word as a whole led to shorter acoustic realizations for all affixes. Furthermore, affixes became shorter as the presentation rate of the stimuli increased. There was no interaction between word frequency and presentation rate, suggesting that the frequency effect also applies in situations in which the speed of articulation is very high.

**A Noninvasive, Low-Cost Device to Study the Velopharyngeal Port During Speech and Some Preliminary Results**

Xiaochuan Niu, Alexander B. Kain, Jan P.H. van Santen; Oregon Health & Science University, USA

**Tue1FoP.2**

It is desirable to monitor the status of the velopharyngeal port during speech. This paper reports on the design and usage of a noninvasive device that measures the static nasal airflow from the nose during speech. The signal can be recorded directly from a generic sound card. Neither does the usage of this device interfere with the articulatory process of speech, nor does it introduce distortions to the simultaneously recorded acoustic signal. We successfully tested the device in an experiment to analyze the velopharyngeal status during normal speech.

**Characterization of Cued Speech Vowels from the Inner Lip Contour**

Noureddine Aboutabit1, Denis Beaufremp1, Laurent Besacier2; 1 ICP, France; 2 CLIPS, France

**Tue1FoP.3**
Cued Speech (CS) is a manual code that complements lip-reading to enhance speech perception from visual input. The phonetic translation of CS gestures needs to combine the manual CS information with information from the lips, taking into account the desynchronization delay (Attina et al. [1], Aboutabib et al. [2]) between these two flows of information. This paper focuses on the analysis of the lip flow for vowels in French Cued Speech. The vocalic lip targets are defined automatically at the instant of minimum velocity of the inner lip contour area parameter, constrained by the corresponding acoustic labeling. We discuss in particular the possibility of discriminating the vowels with geometric lip parameters using the values at the instant of vocalic targets when associated to a Cued Speech hand position.

Modelling Aspiration Noise During Phonation Using the LF Voice Source Model

Christer Gobl; Trinity College Dublin, Ireland

This paper presents a technique for modelling the aspiration noise produced during phonation. The method employs the widely used LF voice source model, which is here extended to include a turbulence noise source for the generation of aspiration. Drawing on speech production theory and on empirical data, the overall amplitude level as well as the within-pulse modulation of the noise are determined by the specific shape of the LF model waveform. The main advantage of this approach is that it simplifies the control of the glottal source, as the temporal dynamics of the variation in noise level need not be explicitly set, but is generated automatically. Informal listening tests also suggest that the proposed modelling technique may contribute to the naturalness of synthesised speech, particularly with regard to breathy voicing.

A Simulation Based Parameter Optimization for a Coarticulation Model

Jianguo Wei, Xugang Lu, Jianwu Dang; JAIST, Japan

A coarticulation model, namely 'carrier model', has been proposed previously by Dang et al. to improve the performance of a physiological articulatory model based speech synthesizer. The carrier model offers a good framework to account for coarticulation in the planning stage, while its parameters need to be refined for improving the performance of the model. This study is to refine the parameters of the carrier model and estimate typical phonetic targets by minimizing the differences between model simulations and observations. A simulation based optimization framework is proposed for this purpose. The framework consists of two layers: obtaining planned targets in a low layer; estimating phonetic targets and optimizing the parameters in a high layer. A discrete search method was applied to the low layer due to the non-analytic nature of the articulation model, while the high layer adopts bilevel optimization strategy to decompose the complicated problem into a set of subproblems. A general evaluation was conducted by combining the refined carrier model and the learned phonetic targets together using the physiological articulatory model and the average error between observations and simulations was 0.15 cm over 103 VCV combinations on the jaw, tongue tip and tongue dorsum.

Multivariate Analysis of Frame-Based Acoustic Cues of Dysperiodicities in Connected Speech

A. Kacha, Francis Grenez, Jean Schoentgen; Université Libre de Bruxelles, Belgium

Generalized variogram is used to extract vocal dysperiodicities in disordered speech produced by dysphonic speakers. Both signal and dysperiodicity are passed through a filter bank and a segmental signal-to-dysperiodicity ratio is defined in each frequency band. Multivariate analysis is carried out to summarize the degree of perceived hoarseness. The predictor variables are the segmental signal-to-dysperiodicity ratios in the different bands. It is shown that high correlations are achieved by linear regression analysis of the segmental signal-to-dysperiodicity ratios in different non-overlapping frequency bands.

Effects of Midline Tongue Piercing on Spectral Centroid Frequencies of Sibilants

Tom Kovacs, Donald S. Finan; University of Colorado at Boulder, USA

Compensatory speech motor patterns occur in response to sensory changes in the vocal tract to facilitate natural sounding speech. The purpose of this project was to investigate speech compensation to midline lingual piercing. In the first experiment, two groups of female speakers (one with tongue piercing) produced two sets of speech samples. Tongue piercing barbells were kept in place as baseline for the first set, and removed for the second. No differences between subject groups were found for /s/ spectral centroid frequencies for the baseline set. A progressive change of /s/ centroid frequencies after barbell removal was observed, indicating rapid compensation to the lingual perturbation. In a second experiment speech samples were recorded from three subjects immediately prior to and following tongue piercing. Changes in centroid frequencies of /s/ and /ʃ/ were dependent on the speaker. It is likely that physiological factors contributed to altered speech observed following piercing.

Assessment of Articulatory Sub-Systems of Dysarthric Speech Using an Isolated-Style Phoneme Recognition System

P. Vijayalakshmi, M.R. Reddy, Douglas O'Shaughnessy; 1Indian Institute of Technology Madras, India; 2Université du Québec, Canada

In this work, the variation in the acoustic realizations of phonemes between normal and dysarthric speech is utilized for the assessment of articulatory sub-systems of dysarthric speech. Using a speech recognition system for the assessment of dysarthric speech is well established either by considering the continuous speech as a sequence of phonemes, or by considering isolated words. These systems will provide information about the intelligibility alone. However, the problems associated with the subsystems of speech can be well captured when a dysarthric speaker is asked to speak continuously and the resultant speech is analyzed at phoneme-level in isolation. Considering this aspect, in our work, an isolated-style triphone-based phoneme recognition system is developed for analyzing continuous speech. Acoustic variations of the phonemes based on place of articulation alone provides information associated with the malfunctioning of articulatory sub-system and are correlated with the Frenchay dysarthric assessment (FDA) scores. The correlation error between our system and FDA scores is found to be only 9%.

Respiratory/Laryngeal Interactions During Sustained Vowel Production in Children

Donald S. Finan, Carol A. Boliek; 1University of Colorado at Boulder, USA; 2University of Alberta, Canada

Like many other rhythmic movements, respiratory behaviors arise from a central pattern generator (CPG). Modulation of the respiratory CPG function is requisite to generate the air pressures and flows essential for speech production. The emergence of voluntary respiratory system control and the coordination of respiratory and laryngeal structures likely parallels and contributes to the development of speech. Perturbation experiments have proven to be fruitful in exploring neural control of many CPG-modulated behaviors. Little is known about the development of respiratory control and the coordination between respiratory and laryngeal systems, especially for complex behaviors such as speech production. The purpose of this study was to investigate the
interaction between laryngeal and respiratory responses to mechanical respiratory perturbation in children.

**Acoustic Characterization of Children with Speech Delay**

*H. Timothy Bunnell, James B. Polikoff; Alfred I. duPont Hospital for Children, USA*

*Tue1FoP.10*

Over the past two decades, significant advances have been made in speech analysis and speech pattern recognition techniques, however, the penetration of these advances (notably in pattern recognition techniques) into the speech disorders research arena has lagged, and penetration into the clinic is virtually non-existent. Here we examine one approach to adapting and extending speech recognition technology based on Hidden Markov Modeling (HMM) to an analysis of speech from children with speech disorders of unknown origin. Specifically, we examine the use of normal-speech trained HMMs to identify acoustically defined categories of segmental distortions, and use those categories to characterize differences among a group of children with developmental speech delays.

**Study of Time and Frequency Variability in Pathological Speech and Error Reduction Methods for Automatic Speech Recognition**

*Oscar Saz, Antonio Miguel, Eduardo Lleida, Alfonso Ortega, Luis Buera; Universidad de Zaragoza, Spain*

*Tue1FoP.11*

In this work, we study the variations in the time and frequency domains inside a Spanish language corpus of speakers with non-pathological and pathological speech. We show how pathological speech has a greater variability in the duration of the words than non-pathological speech, while in the frequency domain we show that the vowels confusability increases by a 18%. The baseline experiments in Automatic Speech Recognition (ASR) with this corpus demonstrate that this variability causes a loss in the performance of ASR systems. To reduce the impact of time and frequency variability we use a recent Vocal Tract Length Normalization (VTLN) system: MATE (augMented stAte space acousTic modEl), as a way of improving the performance of ASR systems when dealing with speakers who suffer any kind of speech pathology. Experiments with MATE show a 17.04% and 11.19% WER reduction by using frequency and time MATE respectively.
Tue2A10 : Formant Estimation
Oral Session, Allegheny Ballroom, left, 14:00 – 16:00
Chair: Thomas F. Quatieri, Massachusetts Institute of Technology, USA

Tracking of Involuntary Formant Frequency Variations and Application to Parkinsonian Speech
Laurence Cnockaert¹, Jean Schoentgen¹, Pascal Auzou², Canan Ozsançak³, Francis Grenez¹; ¹Université Libre de Bruxelles, Belgium; ²Fondation Hopale, France; ³CHRU Lille, France

The objective of this paper is to present a formant frequency estimation method, developed with a view to track small variations due to involuntary vocal tract movement. The formant frequency estimation is based on the instantaneous frequencies obtained by means of a complex wavelet transform and is synchronised with the glottal cycle. Results for synthetic speech signals show the precision of the formant frequency estimation method. Results are presented for speech signals from both normophonic speakers and speakers with Parkinson’s disease.

All-Pole Model Estimation of Vocal Tract on the Frequency Domain
Luis Weruaga¹, Amar Al-Khayat¹; ¹Austrian Academy of Sciences, Austria; ²Philips Austria GmbH, Austria

The most powerful method for speech analysis is the linear prediction analysis, or LPC analysis, one of its main characteristics being the estimation of time-domain related parameters from time-domain samples. This paper proposes a novel speech analysis framework for estimating the spectral poles directly from spectral samples in voiced speech utterances. The method can be described in plain words as the task of fitting the spectral envelope of an all-pole model directly on the log energy of the harmonics. This problem is addressed with an analysis-by-synthesis mechanism supported on a Newton-Raphson algorithm of fast convergence. The proposed method differs clearly from previous approaches commonly used in Harmonic or Sinusoidal Coding. Comparative results on synthetic signals show the excellent performance of the novel analysis technique.

HMM-Based MAP Prediction of Voiced and Unvoiced Formant Frequencies from Noisy MFCC Vectors
Jonathan Darch, Ben Milner; University of East Anglia, UK

This paper describes how formant frequencies of voiced and unvoiced speech can be predicted from mel-frequency cepstral coefficients (MFCC) vectors using maximum a posteriori (MAP) estimation within a hidden Markov model (HMM) framework. Gaussian mixture models (GMMs) are used to model the local joint density of MFCCs and formant frequencies. More localised prediction is achieved by modelling speech using voiced, unvoiced and non-speech GMMs for every state of each model of a set of HMMs. To predict formant frequencies from a MFCC vector, first a prediction of the speech class (voiced, unvoiced or non-speech) is made. Formant frequencies are predicted from voiced and unvoiced speech using a MAP estimation made using the state-specific GMMs. This ‘HMM-GMM’ prediction of speech class and formant frequencies was evaluated on a male 5000-word unconstrained large vocabulary speaker-independent database.

Extracting Formants from Short Segments of Speech Using Group Delay Functions
Anand Joseph M., Guruprasad S., Yegnanarayana B.; Indian Institute of Technology Madras, India

Speech is a non-stationary signal, with the shape of the vocal tract changing over several pitch periods, and also within the open and closed glottis phases. The effect of these changes is reflected in the locations of the formants which correspond to the resonant frequencies of the vocal tract. To observe these changes, the analysis window should be small enough (relative to a pitch period), and appropriately anchored. A non-model based method is proposed in this paper to accurately determine formants from short segments (less than a pitch period) of speech signals. It makes use of high resolution properties of group delay function to estimate formants from segments of duration less than a pitch period. The main advantage of this method is its lack of dependence on the parameters of a model. Analysis segments are synchronised with instants of glottal closure, to increase the robustness of formant extraction. Since continuity or additional acoustic-phonetic knowledge are not used, this method is fairly reliable and robust.

Tracking of Visible Vocal Tract Resonances (VVTR) Based on Kalman Filtering
İ Yücel Özbek, Mübeccel Demirekler; Middle East Technical University, Turkey

This paper analyzes vocal tract resonance (VTR) frequency trajectories and their relationship to formants from a new point of view. Considering abrupt/continuous changes in the physical geometry of vocal tract, VTR may change in number, suddenly change their positions or may leak to some regions where they usually do not exist. We define the visible VTR (VVTR) as VTR that can be seen from the spectrogram. So we propose an algorithm, based on Kalman filtering, that can handle all these changes in VVTR. The suggested properties of VVTR trajectories and the performance of the algorithm are demonstrated on several examples.

Wavelet Ridge Track Interpretation in Terms of Formants
Salma Chaïari, Kais Ouni, Noureddine Ellouze; ENIT, Tunisia

This paper proposes two new approaches for formant tracking using Fourier and wavelet ridges. The speech signal is decomposed into Time-Frequency representations issued from windowed Fourier transform and wavelet transform. Formant tracking is achieved by exploring ridges from time-frequency representation and imposing continuity constraints on formant trajectories. These approaches are validated by their application on synthesized vowels. The formant tracking shows a reliable estimation in comparison with the values given in synthesis. In the same way, the two analyses are applied on natural voiced speech. All the results are compared to two traditional methods using some type of LPC analysis. Fourier ridge detection analysis is shown to be a useful tool for formant tracking compared to traditional methods.
Tue2A2O : Language Processing Beyond and Below the Word-Level
Oral Session, Allegheny Ballroom, center, 14:00 – 16:00
Chair: Renato De Mori, LIA, France

Unsupervised Segmentation of Words into Morphemes — Morpho Challenge 2005 Application to Automatic Speech Recognition
Mikko Kurimo1, Mathias Creutz1, Matti Varjokallio1, Ebru Arisoy2, Murat Saraclar2, 1Helsinki University of Technology, Finland; 2Boğaziçi University, Turkey

Tue2A2O.1
Within the EU Network of Excellence PASCAL, a challenge was organized to design a statistical machine learning algorithm that segments words into the smallest meaning-bearing units of language, morphemes. Ideally, these are basic vocabulary units suitable for different tasks, such as speech and text understanding, machine translation, information retrieval, and statistical language modeling. Twelve research groups participated in the challenge and had submitted segmentation results obtained by their algorithms. In this paper, we evaluate the application of these segmentation algorithms to large vocabulary speech recognition using statistical n-gram language models based on the proposed word-segments instead of entire words. Experiments were done for two agglutinative and morphologically rich languages: Finnish and Turkish. We also investigate combining various segmentations to improve the performance of the recognizer.

Lattice Extension and Rescoring Based Approaches for LVCSR of Turkish
Ebru Arisoy, Murat Saraclar; Boğaziçi University, Turkey

Tue2A2O.2
In this paper, we present some techniques to solve the problems of Turkish Large Vocabulary Continuous Speech Recognition (LVCSR). Its agglutinative nature makes Turkish a challenging language in terms of speech recognition since it is impossible to include all possible words in the recognition lexicon. Therefore, data-driven sub-word recognition units, in addition to words, are used in a newspaper content transcription task. We obtain Word Error Rates (WER) of 38.8% for the baseline word model and 33.9% for the baseline sub-word model. In addition, some new methods are investigated. Baseline lattice outputs of each model are rescored with the root and root-class language models instead of words and first-sub-word language model for sub-words. The word-root interpolation achieves 0.5% decrease in the WER. Other two approaches fail due to the non-robust estimates over the baseline models. Moreover, we have tried dynamic vocabulary extension techniques to handle the Out-of-Vocabulary (OOV) problem in the word model and to remove non-word items in the sub-word model. Applying this method to the 50K baseline word model, in the best situation, we obtain an error rate of 36.2%. In average, the lexicon size of this method is around 188K. However, the error rate is approximately same as the 120K lexicon recognizer. For sub-words, 1.1% absolute improvement is achieved with the vocabulary extension technique giving us our best result.

Exploiting Semantic Relations for a Spoken Language Understanding Application
Catherine Kobus1, Géraldine Damnati1, Lionel Delphin-Poulat1, Renato De Mori2; 1France Télécom, France; 2LIA, France

Tue2A2O.3
This article proposes a new confidence measure estimated for concept hypotheses provided by a semantic language model used in the context of a dialog application. This confidence measure is based upon the ontology and more precisely, upon the semantic relations between concepts. It aims at measuring how high a concept hypothesis is related to the other hypotheses of an utterance. The semantic relation confidence measure is evaluated alone, and in combination with a classical acoustic confidence measure. The two measures are also used as parameters of a decision tree. It is shown that the two confidence measures are complementary and yield good performance in terms of cross entropy relative reduction.

Sentence Boundary Detection of Spontaneous Japanese Using Statistical Language Model and Support Vector Machines
Yuya Akita1, Masahiro Saikou1, Hiroaki Nanjo2, Tatsuya Kawahara1, 1Kyoto University, Japan; 2Ryukoku University, Japan

Tue2A2O.4
This paper presents two different approaches utilizing statistical language model (SLM) and support vector machines (SVM) for sentence boundary detection of spontaneous Japanese. In the SLM-based approach, linguistic likelihoods and occurrence of pause are used to determine sentence boundaries. To suppress false alarms, heuristic patterns of end-of-sentence expressions are also incorporated. On the other hand, SVM is adopted to realize robust classification against a wide variety of expressions and speech recognition errors. Detection is performed by an SVM-based text chunker using lexical and pause information as features. We evaluated these approaches on manual and automatic transcription of spontaneous lectures and speeches, and achieved F-measures of 0.85 and 0.78, respectively.

Compact N-Gram Models by Incremental Growing and Clustering of Histories
Sami Virpioja, Mikko Kurimo; Helsinki University of Technology, Finland

Tue2A2O.5
This work concerns building n-gram language models that are suitable for large vocabulary speech recognition in devices that have a restricted amount of memory and space available. Our target language is Finnish, and in order to evade the problems of its rich morphology, we use sub-word units, morphs, as model units instead of the words. In the proposed model we apply incremental growing and clustering of the morph n-gram histories. By selecting the histories using maxima a posteriori estimation, and clustering them with information radius measure, we obtain a clustered varigram model. We show that for restricted model sizes this model gives better cross-entropy and speech recognition results than the conventional n-gram models, and also better recognition results than non-clustered varigram models built with another recently introduced method.

Opinion Mining in a Telephone Survey Corpus
Nathalie Camelin1, Géraldine Damnati2, Frédéric Béchet1, Renato De Mori1; 1LIA, France; 2France Télécom, France

Tue2A2O.6
Telephone surveys are often used by Customer Services to evaluate their clients’ satisfaction and to improve their services. Large amounts of data are collected to observe the evolution of customers’ opinions. Within this context, the automatization of the process of these databases becomes a crucial issue. This paper addresses the automatic analysis of audio messages where customers are asked to give their opinion over several dimensions about a Customer Service. Interpretation methods that integrate automatically and manually acquired knowledge are proposed. Experimental results, done on a database collected from a deployed Customer Service in real conditions with real customers are given.
A User Simulator Based on VoiceXML for Evaluation of Spoken Dialog Systems
Akinori Ito, Keisuke Shimada, Motoyuki Suzuki, Shozo Makino; Tohoku University, Japan
Tue2A3O.1
This paper describes a user simulator based on analysis of VoiceXML description. A user simulator is a method to evaluate a spoken dialog system without the use of human evaluators. The new feature of our simulator is that it uses a VoiceXML description that describes the dialog system’s behavior. By using the VoiceXML description, the proposed simulator can be used for any dialog system that works with VoiceXML. We constructed a prototype of the user simulator and carried out an evaluation experiment. The experimental result showed that the dialog between the simulator and the dialog system had similar properties to that between human subjects and the dialog system.

User Expectations and Real Experience on a Multimodal Interactive System
Kristiina Jokinen, Topi Hurtig; University of Tampere, Finland
Tue2A3O.2
We present evaluation results of a multimodal route navigation system that allows interaction using speech and tactile/visual modes. Various functional aspects of the system were studied, related especially to the IO-modalities and their use as means of communication. We compared the users’ expectations before the evaluation with their actual experience of the system, and found significant differences among various user groups.

Detecting Anger in Automated Voice Portal Dialogs
F. Burkhardt1, J. Ajmera2, Roman Englert2, J. Stegmann1, W. Burleson2; 1T-Systems Enterprise Services GmbH, Germany; 2Technische Universität Berlin, Germany
Tue2A3O.3
Anger detection is a topic that is gaining more and more attention with voice portal carriers, as it can be useful for quality measurement and emotion-aware dialog strategies. In the context of a prototype voice portal we describe methods to search for training data, report on the performance of the prosodic classifier under real world conditions, and explore the use of dialog information for anger prediction. The results show that, although significantly worse than under laboratory conditions, anger detection still works well above chance level and can be used to enhance real world voice-portal usability.

Evaluation of a Spoken Dialogue System with Usability Tests and Long-Term Pilot Studies: Similarities and Differences
Markku Turunen, Jaakko Hakulinen, Anssi Kainulainen; University of Tampere, Finland
Tue2A3O.4
We present findings from the long-term study of a speech-based bus timetable system. After the deployment of the prototype system we have collected data from real usage for 30 months. In addition, we have conducted usability tests to get subjective ratings of the pilot system. The comparison of these evaluations shows that the results obtained with usability tests differ significantly from those gained from the real usage, and the data of the initial use differs significantly from the data collected after that. For example, the differences in help requests, interruptions, speech recognition rejections, silence timeouts, and repeat requests are highly significant, and in some cases, such as explicit quit requests, enormous (65% versus 3%).

CHAT: A Conversational Helper for Automotive Tasks
Fuliang Weng1, Sebastian Varges2, Badri Raghunathan1, Florin Ratiu2, Heather Pon-Baury1, Brian Lathrop3, Qi Zhang1, Harry Bratt4, Tobias Scheideck1, Kui Xu1, Matthew Purver2, Rohit Mishra3, Annie Lien1, M. Rayla1, S. Peters2, Y. Meng1, J. Russell1, Lawrence Cavedon5, Elizabeth Shriberg4, H. Schmidt1, R. Prieto2; 1Robert Bosch Corp., USA; 2Stanford University, USA; 3Volkswagen of America, USA; 4SRI International, USA; 5National ICT Australia, Australia
Tue2A3O.5
Spoken dialogue interfaces, mostly command-and-control, become more visible in applications where attention needs to be shared with other tasks, such as driving a car. The deployment of the simple dialog systems, instead of more sophisticated ones, is partly because the computing platforms used for such tasks have been less powerful and partly because certain issues from these cognitively challenging tasks have not been well addressed even in the most advanced dialog systems. This paper reports the progress of our research effort in developing a robust, wide-coverage, and cognitive load-sensitive spoken dialog interface called CHAT: Conversational Helper for Automotive Tasks. Our research in the past few years has led to promising results, including high task completion rate, dialog efficiency, and improved user experience.

User Simulation for Spoken Dialogue Systems: Learning and Evaluation
Kalirroi Georgila, James Henderson, Oliver Lemon; University of Edinburgh, UK
Tue2A3O.6
We propose the “advanced” n-grams as a new technique for simulating user behaviour in spoken dialogue systems, and we compare it with two methods used in our prior work, i.e. linear feature combination and “normal” n-grams. All methods operate on the intention level and can incorporate speech recognition and understanding errors. In the linear feature combination model user actions (lists of ⟨speech act, task⟩ pairs) are selected, based on features of the current dialogue state which encodes the whole history of the dialogue. The user simulation based on “normal” n-grams treats a dialogue as a sequence of lists of ⟨speech act, task⟩ pairs. Here the length of the history considered is restricted by the order of the n-gram. The “advanced” n-grams are a variation of the normal n-grams, where user actions are conditioned not only on speech acts and tasks but also on the current status of the tasks, i.e. whether the information needed by the application (in our case flight booking) has been provided and confirmed by the user. This captures elements of goal-directed user behaviour. All models were trained and evaluated on the Communicator corpus, to which we added annotations for user actions and dialogue context. We then evaluate how closely the synthetic responses resemble the real user responses by comparing the user response generated by each user simulation model in a given dialogue context (taken from the annotated corpus) with the actual user response. We propose the expected accuracy, expected precision, and expected recall evaluation metrics as opposed to standard precision and recall used in prior work. We also discuss why they are more appropriate metrics for evaluating user simulation models compared to their standard counterparts. The advanced n-grams produce higher scores than the normal n-grams for small values of n, which proves their strength when little amount of data is available to train larger n-grams. The linear model produces the best expected accuracy but with respect to expected precision and
expected recall it is outperformed by the large n-grams even though it is trained using more information. As a task-based evaluation, we also run each of the user simulation models against a system policy trained on the same corpus. Here the linear feature combination model outperforms the other methods and the advanced n-grams outperform the normal n-grams for all values of n, which again shows their potential. We also calculate the perplexity of the different user models.

Tue2WeO : Corpora, Annotation, and Assessment Metrics II
Oral Session, Westmoreland Room, 14:00 – 16:00
Chair: Briony Williams, University of Wales, UK

Conversational Quality Estimation Model for Wideband IP-Telephony Services
Hitoshi Aoki, Atsuko Kurashima, Akira Takahashi; NTT Corporation, Japan
Tue2WeO.1
As broadband and high-speed IP networks spread, IP-telephony services have become a popular speech communication application over IP networks. Recently, the speech quality of IP-telephony services has become close to that of conventional PSTN services. To provide better speech quality to users, speech communication with wider bandwidth (e.g., 7 kHz) is one of the most promising applications. To ensure desirable quality, we should design the quality before services start and manage it while they are being provided. To do this, an effective means for estimating users' perceptions of speech quality is indispensable. This paper describes a model for estimating the conversational quality of wideband IP-telephony services from physical characteristics of terminals and networks. The proposed model takes into account the quality enhancement effect achieved by widening speech bandwidth and has the advantage that it can evaluate the quality of both wideband and telephone-band speech on the same scale. Based on subjective conversational quality evaluation experiments, we show that the proposed model can accurately estimate the subjective quality for wideband speech as well as for telephone-band speech.

The Vocal Joystick Data Collection Effort and Vowel Corpus
Kelley Kilanski, Jonathan Malkin, Xiao Li, Richard Wright, Jeff A. Bilmes; University of Washington, USA
Tue2WeO.2
Vocal Joystick is a mechanism that enables individuals with motor impairments to make use of vocal parameters to control objects on a computer screen (buttons, sliders, etc.) and ultimately will be used to control electro-mechanical instruments (e.g., robotic arms, wireless home automation devices). In an effort to train the VJ-system, speech data from the TIMIT speech corpus was initially used. However, due to problematic issues with co-articulation, we began a large data collection effort in a controlled environment that would not only address the problematic issues, but also yield a new vowel corpus that was representative of the utterances a user of the VJ-system would use. The data collection process evolved over the course of the effort as new parameters were added and as factors relating to the quality of the collected data in terms of the specified parameters were considered. The result of the data collection effort is a vowel corpus of approximately 11 hours of recorded data comprised of approximately 23500 sound files of the monophthongs and vowel combinations (e.g. diphthongs) chosen for the Vocal Joystick project varying along the parameters of duration, intensity and amplitude. This paper discusses how the data collection has evolved since its initiation and provides a brief summary of the resulting corpus.

Comparison of the ITU-T P.85 Standard to Other Methods for the Evaluation of Text-to-Speech Systems
Dmitry Sitryaev, Katherine Knill, Tina Burrows; Toshiba Research Europe Ltd., UK
Tue2WeO.3
Evaluation of TTS systems is essential to assess performance. The ITU-T P.85 standard was introduced in 1994 to assess the overall quality of speech synthesis systems. However it has not been widely accepted or used. This paper compares the ITU test to more commonly used tests for intelligibility (semantically unpredictable sentences (SUS)) and naturalness (mean opinion score based). The aim of this research was to determine if the ITU test can provide a better performance measure and/or supplementary information to help evaluate TTS systems.

An Annotation Scheme for Complex Disfluencies
Peter A. Heeman1, Andy McMillin2, J. Scott Yaruss3; 1Oregon Health & Science University, USA; 2Hearing & Speech Institute, USA; 3University of Pittsburgh, USA
Tue2WeO.4
In this paper, we present an annotation scheme for disfluencies. Unlike previous schemes, this scheme allows complex disfluencies with multiple backtracking points to be annotated, which are common in stuttered speech. The scheme specifies each disfluency in terms of word-level annotations, thus making the scheme useful for building sophisticated language models of disfluencies. As determining the annotation codes is quite difficult, we have developed a pen and paper procedure in which the annotator lines up the words into rows and columns, from which it is straight-forward for the annotator to determine the annotation tags.

Automatic Phonetic Transcription of Large Speech Corpora: A Comparative Study
Christophe Van Bael, Lou Boves, Henk van den Heuvel, Helmer Strik; Radboud Universiteit Nijmegen, The Netherlands
Tue2WeO.5
This study investigates whether automatic transcription procedures can approximate manual phonetic transcriptions typically delivered with contemporary large speech corpora. We used ten automatic procedures to generate a broad phonetic transcription of well-prepared speech (read-aloud texts) and spontaneous speech (telephone dialogues). The resulting transcriptions were compared to manually verified phonetic transcriptions. We found that the quality of this type of transcription can be approximated by a fairly simple and cost-effective procedure.

Examining Knowledge Sources for Human Error Correction
Yongmei Shi, Lina Zhou; University of Maryland, USA
Tue2WeO.6
A variety of knowledge sources have been employed by error correction mechanisms to improve the usability of speech recognition (SR) technology. However, little is known about the effect of knowledge sources on human error correction. Advancing our understanding of the role of knowledge sources in human error correction could improve the state of automatic error correction. We selected three knowledge sources, including alternative list, imperfect context, and perfect context, and compared their usefulness to human error correction via an empirical user study. The results showed that knowledge sources had significant impact on the performance of human error correction. In particular, perfect context was the best that could significantly reduce word error rate without increasing the processing time.
An Integrated Solution for Error Concealment in DSR Systems Over Wireless Channels
Antonio M. Peinado, Angel M. Gómez, Victoria Sánchez, José L. Pérez-Córdoba, Antonio J. Rubio; Universidad de Granada, Spain

Distributed Speech Recognition involves the development of techniques to conceal the degradations that the transmission channel introduces in the speech features. This work proposes a low-complexity high-accuracy error concealment technique compatible with the DSR ETSI standards. This is achieved by combining three different techniques: fast MMSE estimation, Viterbi decoding with soft-data and subvector-based error detection. We also propose a method to extend this Viterbi decoding to dynamic features. The experimental results show the effectiveness of our proposal.

Interleaving and MMSE Estimation with VQ Replicas for Distributed Speech Recognition Over Lossy Packet Networks
Angel M. Gómez, Antonio M. Peinado, Victoria Sánchez, José L. Carmona, Antonio J. Rubio; Universidad de Granada, Spain

In this work we evaluate the performance of MMSE estimation with a media-specific FEC based on VQ replicas in comparison with MAP estimation and interleaving, both operating in a DSR system over a loss-prone packet switched network. Both schemes combine a sender-driven with a receiver-based technique and, as we show, clearly outperform the standard Aurora mitigation. However, as independent techniques, interleaving and FEC codes could be jointly applied. Although this would provide better results, a direct combination of FECs and interleaving involves a sum of the delays of both operations. In this work, we introduce a double stream-based strategy that avoids this sum of delays.

Noise-Robust Speech Recognition of Conversational Telephone Speech
Gang Chen, Hesham Tolba, Douglas O'Shaughnessy; Université du Québec, Canada

Over the past several years, the primary focus of investigation for speech recognition has been over the telephone or IP network. Recently more and more IP telephony has been extensively used. This paper describes the performance of a speech recognizer on noisy speech transmitted over an H.323 IP telephony network, where the minimum mean-square error log spectra amplitude (MMSE-LSA) method [1,2] is used to reduce the mismatch between training and deployment condition in order to achieve robust speech recognition. In the H.323 network environment, the sources of distortion to the speech are packet loss and additive noise. In this work, we evaluate the impact of packet losses on speech recognition performance first, and then explore the effects of uncorrelated additive noise on the performance. To explore how additive acoustic noise affects the speech recognition performance, seven types of noise sources are selected for use in our experiments. Finally, the experimental results indicate that the MMSE-LSA enhancement method apparently increased robustness for some types of additive noise under certain packet loss rates over the H.323 telephone network.

Lost Speech Reconstruction Method Using Speech Recognition Based on Missing Feature Theory and HMM-Based Speech Synthesis
Shingo Kuroiwa, Satoru Tsuge, Fuji Ren; University of Tokushima, Japan

In recent years, IP telephone service has spread rapidly. However, an unavoidable problem of IP telephone service is deterioration of speech due to packet loss, which often occurs on wireless networks. To overcome this problem, we propose a novel lost speech reconstruction method using speech recognition based on Missing Feature Theory and HMM-based speech synthesis. The proposed method uses linguistic information and can deal with the lack of syllable units which conventional methods are unable to handle. We conducted subjective and objective evaluation experiments under speaker independent conditions. These results showed the effectiveness of the proposed method. Although there is a processing delay in the proposed method, we believe that this method will open up new applications for speech recognition and speech synthesis technology.

Speaker Adaptation Using Evolutionary-Based Linear Transform
Sid-Ahmed Selouani1, Douglas O'Shaughnessy2; 1Université de Moncton, Canada; 2Université du Québec, Canada

This paper presents a technique to adapt HMMs to new speakers using Genetic Algorithms (GAs) in unsupervised mode. The implementation requirements of GAs, such as genetic operators and objective function, have been chosen in order to give more reliability to a global linear transformation matrix. By implementing a ‘survival of the fittest’ strategy, the proposed GA-MLLR approach allows to maintain and manipulate a population of a wide range of solutions. Experiments have been performed on ARPA-RM and TIMIT databases using a triphones HMM-based system. Results show that from a new speaker, significant decrease of word error rate which can reach 6% for a particular speaker, has been achieved by the evolutionary approach, compared to the conventional MLLR-based adaptation method.

A Speaker Adaptation Algorithm Using Principal Curves in Noisy Environments
Jingying Wang, Zuoying Wang; Tsinghua University, China

A new speaker adaptation method of speech recognition is proposed in this paper utilizing principal curves algorithm. The key feature of this method is the construction of a transformation function based on the correlation information between observations of different acoustic states. This is an important a priori information crucial to improving system's recognition performance. Herein the relationships between the statistics information required choosing the best reconstruction of an audio speech pattern and the codebook state parameters of the new algorithm are described, and then the method is applied to a large database of continuous speech. Experiment results on large vocabulary continuous speech recognition database showed that this new method is superior to MLLR adaptation approach in noisy cases, demonstrating that the principal curves speaker adaptation algorithm successfully exploits the correlation information and improve robustness.

Limitations of MLLR Adaptation with Spanish-Accented English: An Error Analysis
Constance Clarke1, Daniel Jurafsky2; 1University at Buffalo SUNY, USA; 2Stanford University, USA

This paper presents a novel lost speech reconstruction method using speech recognition based on Missing Feature Theory and HMM-based speech synthesis. The proposed method uses linguistic information and can deal with the lack of syllable units which conventional methods are unable to handle. We conducted subjective and objective evaluation experiments under speaker independent conditions. These results showed the effectiveness of the proposed method. Although there is a processing delay in the proposed method, we believe that this method will open up new applications for speech recognition and speech synthesis technology.
We studied the effect of MLLR adaptation with Spanish-accented English to understand the strengths and weaknesses of MLLR with unseen foreign accents. We trained a global MLLR transform on 10 adaptation sentences per speaker, giving a 3.4% absolute decrease in phone error rate. We then studied the pattern of improvements across phones and phone classes. Phones that improved the least tended to be those that do not exist in Spanish. Results suggest the poorer performance is related to increased insertion and substituter rates during the adaptation phase, as well as greater acoustic variability.

Issues with Uncertainty Decoding for Noise Robust Speech Recognition

H. Liao, M.J.F. Gales; University of Cambridge, UK

Tue2BuP.8

Recently there has been interest in uncertainty decoding for robust speech recognition. Here the uncertainty associated with the observation in noise is propagated to the recogniser. By using appropriate approximations for this uncertainty, it is possible to obtain efficient implementations during decoding. The aim of these schemes is to obtain performance which is close to that of a model-based compensated system, without the computational cost. Unfortunately, in low SNR there is a fundamental issue with front-end uncertainty decoding where the model means and variances are updated according to the features. This is described in detail using the Joint and SPLICE uncertainty schemes, but is not limited to these two techniques. A solution for the Joint scheme is presented along with the implicit approach used in SPLICE with uncertainty. In addition, a model-based Joint uncertainty scheme is described, which is more efficient and powerful than the front-end schemes, and being model-based not affected by this problem. This issue is illustrated using the AURORA 2.0 database with these various systems.

Vector Taylor Series Based Joint Uncertainty Decoding

Haitian Xu, Luca Rigazio, David Kryzezi; Aalborg University, Denmark; Panasonic San Jose Laboratory, USA

Tue2BuP.9

Joint uncertainty decoding has recently achieved promising results by using front-end uncertainty in the back-end in a mathematically consistent framework. One drawback of the method is that it relies on stereo-data or numerical algorithms, such as DPMC, which have high computational complexity and are difficult to deploy in real applications. We propose a Vector Taylor Series (VTS) approach to joint uncertainty decoding which provides a closed-form solution to the key problem of estimating the clean/noisy speech cross-covariance matrix. Our solution does not require stereo-data or numerical integration. We also propose a new strategy to deal with the cross-covariance matrix singularity. Experiments on Aurora2 show that VTS-based joint uncertainty decoding has similar accuracy compared to DPMC-based joint uncertainty decoding while being at least three times faster. Finally, VTS-based joint uncertainty decoding provided more than 2% absolute improvement when combined with our new strategy for cross-covariance singularity.

A Maximum Likelihood Training Approach to Irrelevant Variability Compensation Based on Piecewise Linear Transformations

Qiang Huo, Donglai Zhu; University of Hong Kong, China

Tue2BuP.10

In our previous works, a maximum likelihood training approach was developed based on the concept of stochastic vector mapping (SVM) that performs a frame-dependent bias removal to compensate for environmental variabilities in both training and recognition stages. Its effectiveness was confirmed by evaluation experiments on Aurora2 and Aurora3 databases. In this paper, we present an extended ML formulation to entertain some new SVM functions that are piecewise linear transformations and are more flexible than the frame-dependent bias removal. Evaluation results on Finnish Aurora3 database show that in comparison with the performance of a baseline system based on ML-trained CDHMMs without feature compensation, the previous and the new SVM-based feature compensation approaches achieve a relative word error rate reduction of 15.7% and 26.1% respectively for well-matched condition.

Speaker Clustered Regression-Class Trees for MLLR Adaptation

Arindam Mandal1, Mari Ostendorf1, Andreas Stolcke2; 1University of Washington, USA; 2SRI International, USA

Tue2BuP.11

A speaker clustering algorithm is presented that is based on an eigenspace representation of Maximum Likelihood Linear Regression (MLLR) transformations and is used for training cluster-dependent regression-class trees for MLLR adaptation. It is shown that significant automatic speech recognition (ASR) system performance gains are possible by choosing the best regression-class tree structure for individual speakers. To take advantage of the potential gains, an algorithm for combining the MLLR mean transformations from cluster-specific trees is described that effectively results in a soft regression-class tree. In conversational speech recognition, only small overall improvements are obtained, but the number of speakers that have performance degradation due to adaptation is reduced by over 70%.

Robust Speech Recognition Over Mobile Networks Using Combined Weighted Viterbi Decoding and Subvector Based Error Concealment

Zheng-Hua Tan, Paul Dalsgaard, Børge Lindberg; Aalborg University, Denmark

Tue2BuP.12

Robustness against transmission errors is one of the primary barriers to the widespread application of automatic speech recognition (ASR) in mobile communications. We have previously proposed a subvector based error concealment (EC) method that conducts error detection and mitigation in the feature-domain at the subvector level. This paper presents a weighted Viterbi decoding (WVD) algorithm that works in the model domain for counteracting unreliable features generated by the subvector based EC. The reliability of each feature is estimated during the process of subvector based EC and is used by the WVD for modifying the observation probability of the feature. Recognition experiments are conducted on the Aurora 2 database corrupted by GSM error pattern EP3. Combining the WVD and the subvector EC achieves 70% and 24% performance improvement as compared to the ETSI-DSR standard and the subvector based EC, respectively.

Speaker Adaptation of Trajectory HMMs Using Feature-Space MLLR

Heiga Zen, Yoshihiko Nankaku, Keiichi Kitamura; Nagoya Institute of Technology, Japan

Tue2BuP.13

Recently, a trajectory model, derived from the hidden Markov model (HMM) by imposing explicit relationships between static and dynamic features, has been proposed. The derived model, named trajectory HMM, can alleviate two limitations of the HMM: constant statistics within a state and conditional independence assumption of state output probabilities. In the present paper, a speaker adaptation algorithm for the trajectory HMM based on feature-space Maximum Likelihood Linear Regression (fMLLR) is derived and evaluated. Results of a simple continuous speech recognition experiment shows that adapting trajectory HMMs using the derived adaptation algorithm improves the speech recognition performance.
Feature and Model Space Speaker Adaptation with Full Covariance Gaussians

Daniel Povey, George Saon; IBM T.J. Watson Research Center, USA
Tue2BuP.14

Full covariance models can give better results for speech recognition than diagonal models, yet they introduce complications for standard speaker adaptation techniques such as MLR and iMLR. Here we introduce efficient update methods to train adaptation matrices for the full covariance case. We also experiment with a simplified technique in which we pretend that the full covariance Gaussians are diagonal and obtain adaptation matrices under that assumption. We show that this approximate method works almost as well as the exact method.

Tue2CaP : Multimodal, Translation and Information Retrieval

Poster Session, Cambria Room, 14:00 – 16:00
Chair: Pedro J. Moreno, Google Inc., USA

Linguistic Tuple Segmentation in Ngram-Based Statistical Machine Translation

Adrià de Gispert, José B. Marín; Universitat Politècnica de Catalunya, Spain
Tue2CaP.1

Ngram-based Statistical Machine Translation relies on a standard Ngram language model of tuples to estimate the translation process. In training, this translation model requires a segmentation of each parallel sentence, which involves taking a hard decision on tuple segmentation when a word is not linked during word alignment. This is especially critical when this word appears in the target language, as this hard decision is compulsory.

In this paper we present a thorough study of this situation, comparing for the first time each of the proposed techniques in two independent tasks, namely English-Spanish European Parliament Proceedings large-vocabulary task and Arabic-English Basic Travel Expressions small-data task. In the face of this comparison, we present a novel segmentation technique which incorporates linguistic information. Results obtained in both tasks outperform all previous techniques.

Sentence Boundary Detection Using Sequential Dependency Analysis Combined with CRF-Based Chunking

Takanobu Oba, Takaaki Hori, Atsushi Nakamura; NTT Corporation, Japan
Tue2CaP.2

In spoken language, sentence boundaries are much less explicit than in written language. Since conventional natural language processing (NLP) techniques are generally designed assuming the sentence boundaries are already given, it is crucial to detect the boundaries accurately for applying such NLP techniques to spoken language. Classification frameworks, such as Support Vector Machines (SVMs) and Conditional Random Fields (CRFs), can be used to detect the boundaries. With these methods, the sentence boundaries are determined based on local sentence-end-like word sequences around the boundaries. However, the methods do not evaluate whether or not each block determined by the boundaries is appropriate as a sentence. We have proposed sequential dependency analysis (SDA), which extracts the dependency structure of unsegmented word sequences with a subsidiary mechanism of sentence boundary detection. In this paper, we extend SDA by combining it with CRFs to reflect both the properties of local word sequences and the appropriateness as a sentence. In this way we achieve more accurate sentence boundary detection. The experimental result shows that our proposed method provides better detection accuracy than that obtained with SVMs or CRFs alone. Our method can also work sequentially because it is based on the SDA framework and can be used for on-line spoken applications.

Sequence Classification for Machine Translation

Srinivas Bangalore, Patrick Haffner, Stephan Kanthak; AT&T Labs Research, USA
Tue2CaP.3

Discriminatively trained classification techniques have been shown to out-perform generative techniques on many speech and natural language processing problems. However, most of the research in machine translation has been based on generative modeling techniques. The application of classification techniques to machine translation requires scaling classifiers to deal with very large label sets (the vocabulary of the target language). In this paper, we present a method to scale classifiers to very large label sets and apply it to train classifiers for machine translation. We contrast this approach to a generatively trained machine translation model represented as a weighted finite-state transducer. We show translation accuracy results on spoken language corpora in English to Spanish and English to Japanese translation tasks.

Two-Stage Vocabulary-Free Spoken Document Retrieval — Subword Identification and Re-Recognition of the Identified Sections

Yoshiaki Itoh1, Takayuki Otake1, Kohei Iwata1, Kazunori Koijima1, Masaaki Ishigame1, Kazuyo Tanaka2, Shi-wook Lee3, 1Iwate Prefectural University, Japan; 2University of Tsukuba, Japan; 3AIST, Japan
Tue2CaP.4

A query word for retrieval systems is liable to be a special term not included in a speech recognizer dictionary. Spoken document retrieval (SDR) systems must therefore be vocabulary-free to deal with arbitrary query words. This paper proposes a new method for vocabulary-free spoken document retrieval. The method exploits two-stage tactics. First, when a query word is submitted, the query word is transformed to a subword sequence according to conversion rules. The subword sequence is searched for spoken documents previously transcribed to a subword sequence by subword recognition. The identified sections are extracted according to the distance between the subword sequences of the query and the identified sections. Second, each identified section is re-recognized using a grammar that includes the query subword sequence. Retrieval experiments were conducted with an actual TV program and the results demonstrated that the proposed method improved SDR performance without long delays in retrieval.

Design and Performance Analysis of a Factoid Question Answering System for Spontaneous Speech Transcriptions

Mihai Surdeanu, David Domínguez-Sal, Pere R. Comas; Universitat Politècnica de Catalunya, Spain
Tue2CaP.5

This paper introduces a QA designed from scratch to handle speech transcriptions. The system’s strength is achieved by analyzing the speech transcriptions with a mix of IR-oriented methodologies and a small number of robust NLP components. We evaluate the system on transcriptions of spontaneous speech from several 1-hour-long seminars and presentations and show that the system obtains encouraging performance.

Performance Improvement of Dialog Speech Translation by Rejecting Unreliable Utterances

Toshiyuki Takezawa, Tohru Shimizu; NICT, Japan
Tue2CaP.6

We discuss how to measure the reliability of recognized utterances...
Recent Advances of IBM’s Handheld Speech Translation System
Weizhong Zhu, Bowen Zhou, Charles Prosser, Pavel Krbec, Yuqing Gao; IBM T.J. Watson Research Center, USA

Recent Advances in the processing capabilities of Personal Digital Assistants (PDAs) have enabled the implementation of an end-to-end speech translation system on these devices. We have presented a bi-directional speech-to-speech (English and Chinese) translation system, which is hosted on a PDA running embedded Linux. Our Multilingual Automatic Speech-to-Speech Translator (MASTOR) system includes an HMM-based large vocabulary continuous speech recognizer using statistical n-grams, a translation module, and a multi-language speech synthesis system. This paper describes our recent efforts to develop a bi-directional English and colloquial Arabic speech-to-speech system on a device running the popular Windows-CE operating system. We have created completely new translation modules and introduced a fast adaptation scheme for a new speaker or environment. In addition, the componentization of system modules speeds up the development of a new speech-to-speech system.

QASR: Question Answering Using Semantic Roles for Speech Interface
Svetlana Stenchikova 1, Dilek Hakkani-Tür 2, Gokhan Tür 3; 1 Stony Brook University, USA; 2 International Computer Science Institute, USA; 3 SRI International, USA

In this paper, we evaluate a semantic role labeling approach to the extraction of answers in the open domain question answering task. We show that this technique especially improves the system performance when answers are communicated to the user by voice. Semantic role labeling identifies predicates and semantic argument phrases in a sentence. With this information we are able to analyze and extract structure from both questions and candidate sentences, which helps us identify more relevant and precise answers in a long list of candidate sentences. When searching for an answer to a question, we match the missing argument in the question to the semantic parses of the candidate answers. This technique significantly improves the accuracy of the question answering system and results in more concise and grammatical answers, which is essential for enabling voice interfaces to question answering systems. In this paper we apply our approach to factoid questions containing predicates; however, this technique can be also useful in answering more complex questions.

Towards a Multimodal Topic Tracking System for a Mobile Robot
Jan F. Maas, Britta Wrede, Gerhard Sagerer; Bielefeld University, Germany

Topics in situated and task oriented communication depend heavily on the given, often changing environment, making the detection of predetermined topics in many cases useless. Detection of non-predefined topics can enhance Human-Robot-Interaction (HRI) in a variety of ways, though.

In this paper we propose a way to dynamically determine topics during Human-Robot-Communication using well established techniques such as Latent Semantic Analysis (LSA). The procedure is based on multimodal cues, supporting the view that topics are not simply a property of spoken or written language, but of multimodal situated communication. An online version of the topic detection system has been developed and is currently being tested on our mobile robot BIRON. To demonstrate the feasibility of our approach, we present the results of an evaluation of our system on the BITT corpus.
Edge-Splitting in a Cumulative Multimodal System, for a No-Wait Temporal Threshold on Information Fusion, Combined with an Under-Specified Display
Edward C. Kaiser, Paulo Barthelmess; Natural Interaction Systems LLC, USA
Tue2CaP.12
Predicting the end of user input turns in a multimodal system can be complex. User interactions vary across a spectrum from single, unimodal inputs to multimodal combinations delivered either simultaneously or sequentially. Early multimodal systems used a fixed duration temporal threshold to determine how long to wait for the next input before processing and integration. Several recent studies have proposed using dynamic or adaptive temporal thresholds to predict turn segmentation and thus achieve faster system response times. We introduce an approach that requires no temporal threshold. First we contrast current multimodal command interfaces to a new class of cumulative-observant multimodal systems that we introduce. Within that new system class we show how our technique of edge-splitting combined with our strategy for under-specified, no-wait, visual feedback resolves parsing problems that underlie turn segmentation errors. Test results show a 46.2% significant reduction in multimodal recognition errors, compared to not using these techniques.

Joint Interpretation of Input Speech and Pen Gestures for Multimodal Human-Computer Interaction
Pui-Yu Hui, Helen M. Meng; Chinese University of Hong Kong, China
Tue2CaP.13
This paper describes our initial work in semantic interpretation of multimodal user input that consist of speech and pen gestures. We have designed and collected a multimodal corpus of over a thousand navigational inquiries around the Beijing area. We devised a processing sequence for extracting spoken references from the speech input (perfect transcripts) and interpreting each reference by generating a hypothesis list of possible semantics (i.e. locations). We also devised a processing sequence for interpreting pen gestures (pointing, circling and strokes) and generating a hypothesis list for every gesture. Partial interpretations from individual modalities are combined using Viterbi alignment, which enforces the constraints of temporal order and semantic compatibility constraints in its cost functions to generate an integrated interpretation across modalities for overall input. This approach can correctly interpret over 97% of the 322 multimodal inquiries in our test set.
Tue3A10 : Advances in Acoustic Segmentation

Oral Session, Allegheny Ballroom, left, 16:30 – 18:30
Chair: Chuck Wooters, International Computer Science Institute, USA

Voice Activity Detector Based on Enhanced Cumulant of LPC Residual and On-Line EM Algorithm

David Cournapeau, Tatsuya Kawahara, Kenji Mase, Tomoji Toriyama, Kyoto University, Japan; Nagoya University, Japan; ATR-MIS, Japan

Infinite Models for Speaker Clustering

Fabio Valente; IDIAP Research Institute, Switzerland

A Constrained Baum-Welch Algorithm for Improved Phoneme Segmentation and Efficient Training

David Huggins-Daines, Alexander I. Rudnicky; Carnegie Mellon University, USA

The Segmentation of Multi-Channel Meeting Recordings for Automatic Speech Recognition

John Dines, Jithendra Vepa, Thomas Hain; IDIAP Research Institute, Switzerland; University of Sheffield, UK

Minimum Boundary Error Training for Automatic Phonetic Segmentation

Jen-Wei Kuo, Hsin-Min Wang; Academia Sinica, Taiwan

Dynamic Evidence Models in a DBN Phone Recognizer

William Schuler, Tim Miller, Stephen Wu, Andrew Exley; University of Minnesota, USA

Instigated the development of novel audio pre-processing approaches. In this paper we present a system for the automatic segmentation of multiple-channel individual headset microphone (HM) meeting recordings for automatic speech recognition. The system relies on an MLP classifier trained from several meeting room corpora to identify speech/non-speech segments of the recordings. We give a detailed analysis of the segmentation performance for a number of system configurations, with our best system achieving ASR performance on automatically generated segments within 1.3% (3.7%) relative of a manual segmentation of the data.

One major research challenge in the domain of the analysis of meeting room data is the automatic transcription of what is spoken during meetings, a task which has gained considerable attention within the ASR research community through the NIST rich transcription evaluations conducted over the last three years. One of the major difficulties in carrying out automatic speech recognition (ASR) on this data is dealing with the challenging recording environment, which has
Tue3A2O : Acoustic Modeling III — LVCSR

The IBM 2006 Speech Transcription System for European Parliamentary Speeches

B. Ramabhadran1, Olivier Siohan1, L. Mangin1, G. Zweig1, M. Westphal2, H. Schulz2, A. Soneiro2; 1IBM T.J. Watson Research Center, USA; 2IBM Germany, Germany

Tue3A2O.1

TC-STAR is an European Union funded speech to speech translation project to transcribe, translate and synthesize European Parliamentary Plenary Speeches (EPPS). This paper describes IBM’s English and Spanish speech recognition systems submitted to the TC-STAR 2006 Evaluation. The technical advances in this submission include two different algorithms for automatic segmentation and speaker clustering of the input audio; a system architecture that is based on cross-adaptation across these two segmentation schemes and system combination through generation of an ensemble of systems using randomized decision tree state-typing; automatic punctuation of the speech recognition output; and the incorporation of an additional 35 hours of in-domain EPPS acoustic training data. These advances reduced the error rate by 30% relative over the best-performing system in the TC-STAR 2005 Evaluation on the 2006 English development test set, and produced one of the best performing systems on the 2006 evaluation in English with a word error rate of 8.3%.

Advances in Lecture Recognition: The ISL RT-06S Evaluation System

Christian Fügen1, Matthias Wölfl1, John W. McDonough1, Shajith Ikbal1, Florian Kraft1, Kornel Laskowski1, Mari Ostendorf2, Sebastian Stüker1, Kenichi Kumatani1; 1Universität Karlsruhe, Germany; 2University of Washington, USA

Tue3A2O.2

This paper describes the 2006 lecture recognition system developed at the Interactive Systems Laboratories (ISL), for individual head-microphone (IHM), single distant microphone (SDM), and multiple distant microphones (MDM) conditions. It was evaluated in RT-06S rich transcription meeting evaluation sponsored by the US National Institute of Standards and Technologies (NIST). We describe the principal differences between our current system and those submitted in previous years, namely, improved acoustic and language models, cross adaptation between systems with different front-ends and phoneme sets, and the use of various automatic speech segmentation algorithms. Our system achieved word error rates of 38.5% (53.4%) and 22.9% (32.2%), respectively, on the MDM and IHM conditions of the RT-05S (RT-06S) lecture evaluation set.

Investigation on Mandarin Broadcast News Speech Recognition

Mei-Yuh Hwang1, Xin Lei1, Wen Wang2, Takahiro Shinozaki1; 1University of Washington, USA; 2SRI International, USA

Tue3A2O.3

This paper describes our efforts in building a competitive Mandarin broadcast news speech recognizer. We successfully incorporated the most popular speech technologies into our system. More importantly, we present two novel algorithms in smoothing pitch features and segmenting Chinese characters into word units. Additionally, we propose to borrow the principle of pointwise mutual information for creating a Chinese word lexicon automatically. Our final system achieved 6.0% character error rate (CER) on dev04 and 16.0% on eval04, with simpler acoustic models, less training data, and simpler decoding architecture compared with other state-of-the-art systems, yet was equally competitive.

Improved Tone Modeling for Mandarin Broadcast News Speech Recognition

Xin Lei1, Manhong Sia2, Mei-Yuh Hwang1, Mari Ostendorf1, Tan Lee3; 1University of Washington, USA; 2Hong Kong University of Science & Technology, China; 3Chinese University of Hong Kong, China

Tue3A2O.4

Tone has a crucial role in Mandarin speech in distinguishing ambiguous words. Most state-of-the-art Mandarin automatic speech recognition systems adopt embedded tone modeling, where tonal acoustic units are used and F0 features are appended to the spectral feature vector. In this paper, we combine the embedded approach (using improved F0 smoothing) with explicit tone modeling in rescoring the output lattices. Oracle experiments indicate 32% relative improvement can be achieved by rescoring with perfect tone information. Recognition experiments on Mandarin broadcast news show that, even with an accuracy of only 70%, the explicit tone classifier offers complementary knowledge and improves performance significantly. Through the combination of tone modeling techniques, the character error rate on the CTV test set can be improved from 13.0% to 11.5%.

Prosodic Modeling in Large Vocabulary Mandarin Speech Recognition

Jui-Ting Huang, Lin-shan Lee; National Taiwan University, Taiwan

Tue3A2O.5

The issue of incorporating prosodic information into speech recognition processes has emerged in recent years. In this work we present a complete framework for Mandarin speech recognition with prosodic modeling considering two-level hierarchical prosodic information for Mandarin Chinese. We developed a GMM-based, a decision-tree-based, and a hybrid approach. The best improvements in character recognition accuracy were obtained by the decision-tree-based prosodic models. This approach does NOT require a training corpus labeled with prosodic features, and works reasonably for a large-scale multi-speaker task.

Experiments on Chinese Speech Recognition with Tonal Models and Pitch Estimation Using the Mandarin Speechon Data

Ying Sun1, Daniel Willett1, Raymond Brueckner1, Rainer Gruhn1, Dirk Bühler2; 1Harman/Becker Automotive Systems, Germany; 2University of Ulm, Germany

Tue3A2O.6

Automatic speech recognition of a tonal and syllabic language such as Chinese Mandarin poses new challenges but also offers new opportunities. We present approaches and experimental results concerning the choice of base units for acoustic modeling, pitch estimation and how to integrate pitch estimates into the modeling framework. The experimental evaluations are carried out both on rather clean headset data and on noisy and reverberant distant talking speech data. Results show that tonal base units offer a word error rate reduction of more than 30% compared to toneless base units. This holds for both phoneme models and initial-final models. The integration of pitch as an additional feature stream yields another remarkable improvement of more than 20% over the best tonal baseline system. In a two-stream modeling approach, the pitch stream distributions can be strongly tied such that the overall model size increases only very moderately.
A Model of the Regularities Underlying Speaker Variation: Evidence from Hybrid Synthesis
Susan R. Hertz; NovaSpeech LLC, USA
Tue3A3O.1
This paper presents the framework of a speech model, tentatively called the “hybrid model,” which offers an explanation of how listeners can identify phonemes in an incoming speech signal despite the vast amount of cross-speaker and contextual variation. Fundamental to the model are two basic speech units into which listeners process the incoming speech stream: acoustic consonant clusters and acoustic nuclei. Acoustic nuclei are responsible for speaker identity, but acoustic consonant clusters are more generic and can even be substituted across speakers with negligible impact on speech quality. The paper focuses on acoustic consonant clusters, showing that much of the variability in them is perceptually irrelevant, and how the hybrid model accounts for listeners’ ability to parse them into phonemes. The paper supports the model as applied to English by drawing on experiments in hybrid synthesis, a technique in which speech is produced by splicing together segments from different speakers [1].

Pauses as a Tool to Ensure Rhythmic Wellformedness
Augustin Speyer; University of Pennsylvania, USA
Tue3A3O.2
Rhythmic wellformedness on the level of syllables and words and the mechanisms which are employed to ensure it are well known and researched on. The level of sentential accent is less often in the focus of such studies. In this paper I argue that rhythmic wellformedness plays an equally important role on the sentential level and that the preferred strategy to ensure it is the insertion of a pause (or lengthening of an already present pause), rather than the classical strategies of stress shift and destressing. This has to do with the role that focal accent plays for the semantic interpretation of an utterance. A reading experiment showed clearly that in clash cases the distance between the words in clash is increased by a pause; as we get the same effect in words in which a number of unstressed syllables intervene between the clashing focal accents, it is clear that it is not simply stress clash resolution but resolution of a clash on a higher level of representation.

Factors Affecting Speakers’ Choice of Fillers in Japanese Presentations
Michiko Watanabe1, Yasuharu Den2, Keikichi Hirose1, Shusaku Miwa1, Nobuaki Minematsu1; 1University of Tokyo, Japan; 2Chiba University, Japan
Tue3A3O.3
Disfluencies are intrinsic in spontaneous speech. Although it is known that there is a wide range of frequencies and types of disfluencies among speech, little is known about factors affecting speakers choice of disfluency types. We first conducted a correspondence analysis using ratios of seven types of fillers, and other disfluencies, in 174 presentations and quantified the data. We conducted a cluster analysis using 4 dimension scores from the correspondence analysis, and extracted five filler-type groups. We then examined frequent types of presentations (formal or casual) and speaker attributes (gender and age) in each group. The results indicate that speakers choice of filler types is affected by speech levels, speakers gender and age, and that relevant factors differ depending on the type of fillers.

Developing Consistent Pronunciation Models for Phonemic Variants
Marelle Davel, Etienne Barnard; CSIR, South Africa
Tue3A3O.4
Pronunciation lexicons often contain pronunciation variants. This can create two problems: It can be difficult to define these variants in an internally consistent way and it can also be difficult to extract generalised grapheme-to-phoneme rule sets from a lexicon containing variants. In this paper we address both these issues by creating ‘pseudo-phonemes’ associated with sets of ‘generation restriction rules’ to model those pronunciations that are consistently realised as two or more variants. By pre-processing and post-processing the lexicon appropriately, grapheme-to-phoneme algorithms that were not able to deal with pronunciation variants previously can now be extended to incorporate variants easily, without requiring changes to the standard algorithms. We evaluate the effectiveness of this approach using the Default&Refine rule extraction algorithm, and apply the method to both the English Oxford Advanced Learners Dictionary (OALD) and the Flemish FONILEX pronunciation lexicon. We find that the approach generalises to different languages, is able to model phonemic variation accurately and is able to identify inconsistent variants in pre-existing lexicons.

Grapheme-to-Phoneme Conversion Using Automatically Extracted Associative Rules for Korean TTS System
Jinsik Lee, Seungwon Kim, Gary Geunbae Lee; POSTECH, Korea
Tue3A3O.5
In this paper, we describe a method for automatically extracting grapheme-to-phoneme conversion rules directly from the transcription of speech synthesis database and introduce a weighted score and jamo similarity to overcome the rule application difficulties. We make a structured rule tree by rule pruning and rule association, and can eliminate most of the rules with almost no decrease of the performance. Our system achieves over 99.5 percent of phoneme-level accuracy and this performance is easily achievable even with the small amount of training data.

Example-Based Grapheme-to-Phoneme Conversion for Thai
Paisarn Charoenpornsawat, Tanja Schultz; Carnegie Mellon University, USA
Tue3A3O.6
Several characteristics of the Thai writing system make Thai grapheme-to-phoneme (G2P) conversion very challenging. In this paper, we propose an Example-Based Grapheme-to-Phoneme conversion approach. It generates the pronunciation of a word by selecting, modifying and combining pronunciations from syllables from training corpus. The best system achieves 80.39% word accuracy and 94.19% phone accuracy which significantly outperform previous approaches for Thai.
Results of mirror-images of the original fragments show that and that the left part of the face is more important than the right stronger cue value for prominence detection than the bottom part, screen which displayed the film fragments. The task of the subjects positioned at a distance of either 50cm, 250cm or 380cm from the could only see the upper or lower half, or the right or left part of in spoken utterances. In particular, it investigates which area of The article discusses the processing of facial markers of prominence while for the visual modality local cues are sufficient. fragments are better classified in the auditory modality, while for short fragments the visual modality works better. This suggests that fragments were presented in 3 modalities: either a bimodal presenta- tion, questioning, uncertain, happy, angry and neutral. Results showed that in all expressive modes, words with focal accent are accompanied by a greater variation of the facial parameters than are words in non-focal positions. Moreover, interesting differences between the expressions in terms of different parameters were found.

How Auditory and Visual Prosody is Used in End-of-Utterance Detection
Pashiera Barkhuysen, Emiel Krahmer, Marc Swerts; Tilburg University, The Netherlands

In this paper, we present measurements of visual, facial parameters obtained from a speech corpus consisting of short, read utterances in which focal accent was systematically varied. The utterances were recorded in a variety of expressive modes including certain, confirming, questioning, uncertain, happy, angry and neutral. Results showed that in all expressive modes, words with focal accent are accompanied by a greater variation of the facial parameters than are words in non-focal positions. Moreover, interesting differences between the expressions in terms of different parameters were found.

The Importance of Different Facial Areas for Signalling Visual Prominence
Marc Swerts, Emiel Krahmer; Tilburg University, The Netherlands

This article discusses the processing of facial markers of prominence in spoken utterances. In particular, it investigates which area of a speaker’s face contains the strongest cues to prominence, using stimuli with the entire face visible or versions in which participants could only see the upper or lower half, or the right or left part of the face. To compensate for potential ceiling effects, subjects were positioned at a distance of either 50cm, 250cm or 380cm from the screen which displayed the film fragments. The task of the subjects was to indicate for each stimulus which word they perceived as the most prominent one. Results show that, while prominence detection becomes more difficult at longer distances, the upper facial area has stronger cue value for prominence detection than the bottom part, and that the left part of the face is more important than the right part. Results of mirror-images of the original fragments show that this latter result is due both to a speaker and an observer effect.

Visual Speech Segmentation and Speaker Recognition for Transcription of TV News
Josef Chaloupka; Technical University of Liberec, Czech Republic

This paper is about a method for visual segmentation of TV news. The TV news shows are segmented according to the visual stream from the video TV recordings in this method. Human faces are found in the single visual segments with the help of the fast algorithm for face detection. The found faces are compared with the visual GMMs, that have been trained from the video picture of the single broadcasters (anchors) from the TV news. The single visual segments, where the faces of the broadcasters have been found and recognized, have been compared with the acoustic segments from the acoustic segmentation. The speaker adapted HMMs have been used for speech recognition of these acoustic segments. The recognition rate is better for the use of this speaker-adapted HMMs compared to the use of the speaker independent HMMs. It is possible to use the methods for the speaker identification and verification from the acoustic signal in the acoustic segments. The results from the visual speaker identification will be better for smaller number of speakers and for the use of the video recordings of TV news with a lot of noise in the acoustic signal.

HMM-Based Continuous Sign Language Recognition Using a Fast Optical Flow Parameterization of Visual Information
G. Cortés, L. García, Carmen Benítez, José C. Segura; Universidad de Granada, Spain

This paper presents a preliminary study of an optical flow-based parameterization of visual information in a sign language recognition system using Hidden Markov Models (HMM). Current feature extraction processes need initialization, tracking and segmentation stages in order to describe signer gestures. Our aim is to develop a single and fast technique to reduce computational complexity which doesn’t require these stages and is able to work in mobile devices with limited hardware resources. The Moving Block Distance (MBD) parameterization is an interesting first approach for this purpose, proved by two signers under a static background constraint. A lexicon of 33 basic word units (signemes) was used to build the data set containing phrases with a variable number of words. Continuous recognition results achieve more than 99% accuracy in close test.

Audio-Visual Speech Recognition in the Presence of a Competing Speaker
Xu Shao, Jon Barker; University of Sheffield, UK

This paper examines the problem of estimating stream weights for a multistream audio-visual speech recogniser in the context of a simultaneous speaker task. The task is challenging because signal-to-noise ratio (SNR) cannot be readily inferred from the acoustics alone. The method proposed employs artificial neural networks (ANNs) to estimate the SNR from HMM state-likelihoods. SNR is converted to stream weight using a mapping optimised on development data. The method produces an audio-visual recognition performance better than that of both the audio-only and the video-only baselines across a wide range of SNRs. The performance using SNR estimates based on audio state-likelihoods is compared to that obtained using both audio and visual likelihoods. Although the audio-visual SNR estimator outperforms the audio-only SNR estimator, the recognition performance benefit is small. Ideas for making fuller use of the visual information are discussed.
Interspeech 2006 — ICSLP  
September 17–21 2006, Pittsburgh Pennsylvania USA

**Expressive Prosody for Unit-Selection Speech Synthesis**

**Volker Strom, Robert A.J. Clark, Simon King; University of Edinburgh, UK**

**Tue3BuP.1**

Current unit selection speech synthesis voices cannot produce emphasis or interrogative contours because of a lack of the necessary prosodic variation in the recorded speech database. A method of recording script design is proposed which addresses this shortcoming. Appropriate components were added to the target cost function of the Festival Multisyn engine, and a perceptual evaluation showed a clear preference over the baseline system.

**Cues for Hesitation in Speech Synthesis**

**Rolf Carlson ¹, Kjell Gustafson ¹, Eva Strangert ²; ¹ KTH, Sweden; ² Umeå University, Sweden**

**Tue3BuP.2**

The current study investigates acoustic correlates to perceived hesitation based on previous work showing that pause duration and final lengthening both contribute to the perception of hesitation. It is the total duration increase that is the valid cue rather than the contribution by either factor. The present experiment using speech synthesis was designed to evaluate F0 slope and presence vs. absence of creaky voice before the inserted hesitation in addition to durational cues. The manipulations occurred in two syntactic positions, within a phrase and between two phrases, respectively. The results showed that in addition to durational increase, variation of both F0 slope and creaky voice had perceptual effects, although to a much lesser degree. The results have a bearing on efforts to model spontaneous speech including disfluencies, to be explored, for example, in spoken dialogue systems.

**Multi-Domain Text-to-Speech Synthesis by Automatic Text Classification**

**Francesc Alías, Joan Claudi Socoró, Xavier Sevillano, Ignasi Iriondo, Xavier Gonzalez; Ramon Llull University, Spain**

**Tue3BuP.3**

This paper describes a multi-domain text-to-speech (MD-TTS) synthesis strategy for generating speech among different domains and so increasing the flexibility of high quality TTS systems. To that effect, the MD-TTS introduces a flexible TTS architecture that includes an automatic domain classification module, which allows MD-TTS systems to be implemented by different synthesis strategies and speech corpus typologies. In this work, the performance of a corpus-based MD-TTS system is subjectively validated by means of several perceptual tests.

**Phrase Break Prediction Using Logistic Generalized Linear Model**

**Lifu Yi, Jian Li, Xiaoyan Lou, Jie Hao; Toshiba China R&D Center, China**

**Tue3BuP.4**

In this paper we propose a novel phrase break prediction model for Mandarin speech synthesis. It is generalized linear models (GLM) with stepwise regression solution. We assume phrase break obeys Bernoulli distribution and then model phrase break probability by Logistic GLM. The attribute set is automatically selected by stepwise regression, which is a totally data-driven method. We also introduce speaking rate as a new attribute for prediction. The proposed method is applied to 2,150 utterances of the Mandarin speech corpus, and it achieves 5.4% higher performances than CART method in open test. The method can be extended to include more linguistic and prosodic attributes and it is very compact for application.

**Joint Prosodic and Segmental Unit Selection Speech Synthesis**

**Robert A.J. Clark, Simon King; University of Edinburgh, UK**

**Tue3BuP.5**

We describe a unit selection technique for text-to-speech synthesis which jointly searches the space of possible diphone sequences and the space of possible prosodic unit sequences in order to produce synthetic speech with more natural prosody. We demonstrate that this search, although currently computationally expensive, can achieve improved intonation compared to a baseline in which only the space of possible diphone sequences is searched. We discuss ways in which the search could be made sufficiently efficient for use in a real-time system.

**Phonetically Enriched Labeling in Unit Selection TTS Synthesis**

**Yeon-Jun Kim, Ann K. Syrdal, Alistair Conkie, Mark C. Beutnagel; AT&T Labs Research, USA**

**Tue3BuP.6**

Unit selection techniques have improved the quality of text-to-speech (TTS) synthesis. However, mistakes which had been less noticeable previously in poorer quality synthetic speech become very noticeable in more natural-sounding synthetic speech. Many problems appear to be caused by mismatches between phones requested by the TTS front-end and phones selected from the labeled speech inventory. Given the input text and the added information predicted by the TTS front-end, finding the optimal units from a speech inventory database still remains a challenge in unit selection TTS synthesis. Consonants in American English affect intelligibility of speech synthesis and they are realized differently depending on their position in the syllable. Pre-vocalic plosives must have a release burst before the vowel begins while post-vocalic consonants may or may not be released. When a post-vocalic consonant is chosen to synthesize a pre-vocalic consonant, it may cause problems such as missing consonants, consonant confusion or word-boundary confusion. In this paper, a new phone labeling method which differentiates pre-vocalic and post-vocalic consonants is proposed. The proposed phone labeling method leads unit selection to choose contextually accurate phone units and minimizes unit selection errors caused by lack of specification in TTS front-end transcriptions and phone labels in the speech inventory. In a listening test the TTS voices labeled with the pre-vocalic / post-vocalic distinction were rated significantly higher (+0.33) compared to reference voices that did not use this distinction.

**Further Developments in LSM-Based Boundary Training for Unit Selection TTS**

**Jerome R. Bellegarda; Apple Computer Inc., USA**

**Tue3BuP.7**

The level of quality that can be achieved in concatenative text-to-speech synthesis depends, among other things, on a judicious segmentation of all units in the underlying unit selection inventory. We have recently advocated the iterative refinement of unit boundaries based on a data-driven feature extraction framework separately optimized for each boundary region [1]. This paper presents the formal proof of convergence of the iterative algorithm, as well as a detailed analysis of its potential benefits for concatenative TTS synthesis. A formal listening test, in particular, underscores the practical viability of the approach for unit boundary optimization.
A Style Control Technique for Speech Synthesis Using Multiple Regression HSMM
Takashi Nose, Junichi Yamagishi, Takao Kobayashi; Tokyo Institute of Technology, Japan

This paper presents a technique for controlling intuitively the degree or intensity of speaking styles and emotional expressions of synthetic speech. The conventional style control technique based on multiple regression HMM (MRHMM) has a problem that it is difficult to control phone duration of synthetic speech because HMM has no explicit parameter which models phone duration appropriately. To overcome this problem, we use multiple regression hidden semi-Markov model (MRHSM) which has explicit state duration distributions to control phone duration. We show that the duration control is important for style control of synthetic speech from the results of subjective tests. We also compare the proposed technique with another control technique based on model interpolation.

Acoustic Model Training Based on Linear Transformation and MAP Modification for HSMM-Based Speech Synthesis
Katsumi Ogata, Makoto Tachibana, Junichi Yamagishi, Takao Kobayashi; Tokyo Institute of Technology, Japan

This paper describes the use of combined linear regression and ex-post MAP methods for average-voice-based speech synthesis system based on HMM. To generate more natural sounding speech using the average-voice-based speech synthesis system when a large amount of training data is available, we apply ex-post MAP estimation after the linear transformation based adaptation. We investigate how the amount of data used in the training of the average voice model and the tying topology affect the naturalness of synthetic speech. From the results of evaluation tests, we show that the adapted average voice model trained using a large amount of data can generate more natural sounding speech than the speaker dependent model.

Improving Arabic HMM Based Speech Synthesis Quality
Ossama Abdel-Hamid, Sherif Mahdy Abdou, Mohsen Rashwan; Cairo University, Egypt

This paper presents a technique for improving the quality of Arabic speech synthesis, where speech parameters are generated directly from HMM models, a new technique relating to other speech synthesis techniques. In this paper, we propose some modifications to the basic system to improve its quality. We apply a multi-band excitation model and use samples extracted from the spectral envelop as spectral parameters. In the synthesis, the voiced and unvoiced speech parts are mixed according to bands voicing parameters. The voiced part is generated based on a harmonic sinusoidal model. Experimental tests performed on Arabic dataset show that the applied modifications improved the quality.

FarsBayan: A Unit Selection Based Farsi Speech Synthesizer
M. Mehdi Homayounpour, Majid Namnabat; Amirkabir University of Technology, Iran

In recent years, the unit selection-based concatenative speech synthesis method using a large corpus has attracted great attention. This method provides more natural quality speech compared to the parameter driven methods. The Formant Synthesis, HNM method and use of MLSA filter are the prevalent methods for synthesizing Farsi speech. In this paper, we present the structure of a proposed unit selection synthesizer for Farsi language. In the proposed system, the linear regression method has been used for determination of weights of discrete sub-costs in the target cost, while the weights of other sub-costs have been considered constant. We have also presented a pre-selection algorithm using adaptive threshold for pruning the units. In addition, the efficiency of TD-PSOLA algorithm in improvement of resulting speech quality has been studied. Informal tests show the degrading effect of this algorithm on the output quality. The output speech was found to be remarkably fluent and natural. The quality of the output speech has been evaluated using MOS subjective test and we have obtained a MOS test value of 3.8 for overall quality.

Amharic Speech Synthesis Using Cepstral Method with Stress Generation Rule
Tadesse Anberbir, Tomio Takara; University of the Ryukyus, Japan

Amharic is the official language of Ethiopia. In this paper, we present our study on Amharic stress. Stress (Gemination of consonants) in Amharic language is very important for proper pronunciation of words. It is also one of the most distinctive characteristics of the rhythm of the speech. We discuss a method employed for generating stressed syllables from unstressed syllables, and its application to our speech synthesizer. First, we analyzed waveforms of minimal pair words concerned with stressed and unstressed syllables in the time patterns of pitch, power and spectrum. Then, by combining or exchanging these patterns, speech sounds were synthesized. Using the synthesized sounds, listening tests were performed to examine the acoustic correlates of stress among pitch, spectrum, power and duration. We found that consonant’s duration is the most important factor. A further listening test was performed to determine the threshold of duration of consonants between unstressed and stressed syllables, and we observed that 50ms is the average threshold duration for voiced consonants and 70ms is for unvoiced consonants.

Automatic Syllable-Pattern Induction in Statistical Thai Text-to-Phone Transcription
AUSDANG THANGTHAI, CHATCHAWARN HANSAKUNBUNTHEUNG, RUNGKARN SIRICHAROENCHAI, CHAI WUTTIWATCHAI; NECTEC, THAILAND

This paper proposes a technique of automatic syllable-pattern induction in statistical Thai text-to-phone transcription. A general process of building a statistical text-to-phone transcription is to first build a set of rules describing syllable patterns, which is used for syllabification. Given an input text, the syllabification process generates all possible syllable sequences, which are then scored and selected using a statistical model. Updating the handcrafted rule set of syllable patterns is time-consuming and requires expert linguists. Instead of the manual process, automatic induction of new syllable patterns from a large raw text if proposed. The process that can deal with raw text is particularly needed for Thai as segmenting Thai text is a very tedious task. Experimental tests show that the proposed Thai text-to-phone transcription system after applying a large raw text for syllable-pattern induction achieves approximately 2% improvement. A comparison with other Thai text-to-phone transcription models and error analyses are also given in the paper.

Development of Prototype Text-to-Speech Systems for Northern Sotho
H.J. Oosthuizen, S.T. Phihela, M.J.D. Manamela; University of Limpopo, South Africa

Two text-to-speech synthesis systems were developed for one of the eleven official languages of South Africa, viz. Northern Sotho. A diphone synthesis system, based on extraction of diphones from nonsense words, was constructed. A cluster unit selection synthesis system, based on recordings of sentences containing a selection of
most common words in Northern Sotho, was also built. The Festival speech synthesis system was used for both systems. Both of these systems performed well in a subjective evaluation.

**Identify Language Origin of Personal Names with Normalized Appearance Number of Web Pages**

Jiali You¹, Yining Chen², Min Chu², Yong Zhao², Jinhlin Wang¹; ¹Chinese Academy of Sciences, China; ²Microsoft Research Asia, China

Tue3CuP.15

Identifying the language origin of a personal name without context is interesting and useful in many areas. Morphological structure, which has long been considered as the main source of language origin information, is modeled by N-grams of letters or letter chunks. In this paper, we introduce a new information source, the appearance number of a name in web pages of different languages, for identifying its language origin. Since the distribution of web pages in various languages is not identical, and the state-of-the-art search engines can only provide the number of pages that contain the queried words, we propose a method to normalize the appearance number obtained from a search engine and use it as a new feature. When this new feature is used independently to identify language origin of names among four closely related languages (English, German, French, and Portuguese), the error rate is 26.0%, which is comparable to that of letter 4-gram features. When it is used together with the letter N-gram models, the error rate is reduced to 14.2%, which is about 43.2% error reduction, compared with the letter 4-gram based baseline model.

**Observations of the Spoken Language Acquisition Process Based on a Multimodal Infant Behavior Corpus**

Ryo Tsuji, Tomohiko Kasami, Shogo Ishikawa, Shinya Kiriyama, Yoichi Takebayashi, Shigeyoshi Kitazawa; Shizuoka University, Japan

Tue3CaP.1

We have developed a framework to record spontaneous speech of infants. Using the framework, we have accumulated infant speech data, and proved that the data is quite efficient for the explication of the spoken language acquisition process. We aim at constructing the “multimodal infant behavior corpus,” which contributes to the elucidation of human commonsense knowledge and its acquisition mechanism. We previously established the environments to record infant behaviors as multimodal data. We have newly developed a wearable speech recording system and succeeded to record infant utterances with high quality. We have held an infant school once a week for 10 months, and accumulated infant speech data more than 100 hours long. We observed infant utterances in the aspects of acoustic and prosodic features. Through an acoustic observation, we have succeeded to analyze alteration of the pronunciation skills focused on demonstratives that appear quite often in infant utterances. As the result of a prosodic observation, we have also obtained knowledge of how infants enrich and diversify the ways to explain their intentions or emotions corresponding to their growth.

**Infants’ Ability to Extract Verbs from Continuous Speech**

Ellen Marklund, Francisco Lacerda; Stockholm University, Sweden

Tue3CaP.2

Early language acquisition is a result of the infant’s general associative and memory processes in combination with its’ ecological surroundings. Extracting a part of a continuous speech signal and associate it to for instance an object, is made possible by the structure provided by characteristically repetitive Infant Directed Speech. The parents adjust the way they speak to their infant based on the response they are given, which in turn is dependent on the infant’s age and cognitive development.

It seems probable that the ability to extract lexical candidates referring to visually presented actions is developed at a later stage than the ability to extract lexical candidates referring to visually presented objects — actions are more abstract and there is a time aspect involved.

Using the Visual Preference Paradigm, the ability to extract lexical candidates referring to actions was studied in infants at the age of 4 to 8 months. The results suggest that while the ability at this age is not greatly apparent, it seems to increase slightly with age.

**Category Formation and the Role of Spectral Quality in the Perception and Production of English Front Vowels**

Ricardo A.H. Bion¹, Paola Escudero², Andréa S. Rauber³, Barbara O. Baptista³; ¹University of Joensuu, Finland; ²University of Amsterdam, The Netherlands; ³Federal University of Santa Catarina, Brazil

Tue3CaP.3

This study aimed at comparing the perception and production of English front vowels by 17 proficient Brazilian speakers of English as a second language (L2) and 6 native speakers of American English. Towards this end, three experiments were carried out: (i) a production test measuring the first two formants of the participants’ English front vowels, (ii) an oddity discrimination test investigating the formation of vowel categories, and (iii) a discrimination test with synthetic stimuli which assessed the participants’ reliance on spectral quality when perceiving English vowels. The results of these experiments suggest a strong relationship between L2 vowel perception and production, since the vowel pairs which were produced with similar formant values by the Brazilian participants were also poorly discriminated in the two perception tests. In addition, the findings suggest that vowel perception might precede vowel production, as high rates on the discrimination of vowel pairs on both perception tests were a prerequisite for differentiating the same two vowels on the production test. Lastly, some Brazilian participants obtained native-like scores on the category formation test without manifesting native-like reliance on spectral quality, indicating that other acoustic cues, such as vowel duration, might be playing a role in their perception of English vowels.

**Productions in Bilinguism, Early Foreign Language Learning and Monolinguism: A Prosodic Comparison**

Ranka Bijeljac-Babic¹, Christelle Dodane², Sabine Metta¹, Claire Gérard¹; ¹LMDC, France; ²DDL, France

Tue3CaP.4

The degree of L2 foreign accent is likely to vary, according to the age of the acquisition, the length of contact with L2 and the possible interaction between L1 and L2. This study examined how children who master French and English at different levels pronounce disyllabic words in both languages. Acoustic analysis (F0, duration and amplitude) of syllables in disyllabic words were compared between 8 bilingual children (French-English, aged between 3;6 and 6;1 years) and 16 monolingual children (8 French children and 8 English children of the same age) and confronted to the analysis of 20 (7 years aged) early French Learners of English (FLE) children. Results showed that the bilingual children acquired prosodic patterns in both languages. However, the accent of their disyllabic words differed from those of the monolingual children. French 7 years old-aged learners of English, after only two years of acquisition, produced the native-like accent. Our findings modulate the “critical age” hypothesis and bring some new elements in favour of the L1 and L2 obligatory interaction hypothesis.
Training Native English Speakers to Identify Japanese Vowel Length with Fast Rate Sentences

Yukari Hirata\textsuperscript{1}, Elizabeth Whitehurst\textsuperscript{1}, Emily Collings\textsuperscript{1}, Jacob Whiton\textsuperscript{1}, Carol Glenn\textsuperscript{2}; \textsuperscript{1}Colgate University, USA; \textsuperscript{2}Cornell University, USA

Tue3Cap.5

Native English speakers were trained to identify Japanese vowel length with sentences spoken at a fast rate, and their test scores before and after training were compared with those of a control group. Results indicated that the present training with little variability in speaking rate did not yield a robust perceptual improvement for the trained group. However, among the three rates tested (slow, normal, fast), the trained group showed significant improvement specifically for the slow-rate stimuli, distinct from the control group's.

Formant-Based English Vowel Assessment for Chinese in Taiwan

Jiang-Chun Chen\textsuperscript{1}, Wei-Tang Hsu\textsuperscript{1}, J.-S. Roger Jang\textsuperscript{1}, Ren-Yuan Lyu\textsuperscript{2}, Yang-Chin Chiang\textsuperscript{1}; \textsuperscript{1}National Tsing Hua University, Taiwan; \textsuperscript{2}Chang Gung University, Taiwan

Tue3Cap.6

This paper proposes a formant-based approach for computer-assisted English vowel assessment. Various studies in formant-based speech synthesis have suggested the importance of formant coefficients; this motivates us to investigate pronunication assessment using formant information instead of MFCC (Mel-frequency cepstral coefficients) alone. In particular, we explore the multi-stream HMM with the addition of formant information to improve the phoneme segmentation. We then propose the use of PCN (pronunciation confusion network) together with a formant-based confidence measure to improve error detection rates. Furthermore, the pros and cons of using cross-word phone model for both native speakers and L2 learners are discussed. Experimental results demonstrate the feasibility of the proposed approach for automatic vowel pronunciation assessment.

Substitute Sounds for Ventriloquism and Speech Disorders

Jörg Metzner\textsuperscript{1}, Marcel Schmittfuß\textsuperscript{2}, Karl Schnell\textsuperscript{2}; \textsuperscript{1}University of Oxford, UK; \textsuperscript{2}Goethe-University Frankfurt, Germany

Tue3Cap.7

The restriction of articulation used by ventriloquists or caused by speech disorders can be compensated using substitute sounds. For a better understanding of these sounds, in this contribution the results of investigations of the substitute sounds by analysis and synthesis are presented. For that purpose the substitute sounds and their natural counterparts uttered by a ventriloquist are analyzed. Substitute sounds are also generated by articulatory synthesis and by a plaster model and are then compared to the original sounds. The results show that the perception of substitute sounds can reach the natural uttered sounds. The degree of the perceptive similarity depends on the substitute sound. The spectrum of the substitute sounds can replicate partially the spectrum of the natural sounds. However, for successful perception of the sound it is only necessary to reproduce the relevant spectral information. The knowledge about those sounds is interesting for speech therapy.

Automatic Mandarin Pronunciation Scoring for Native Learners with Dialect Accent

Si Wei, Qing-Sheng Liu, Yu Hu, Ren-Hua Wang; University of Science & Technology of China, China

Tue3Cap.8

This paper studies pronunciation scoring algorithm in CALL system aiming at teaching native Chinese learn standard Mandarin. Most of the pronunciation scoring algorithms focus on non-native environment, which may not be suitable for native speakers. We bring up a new algorithm based on traditional posterior log-likelihood algorithm by weighting the initial part of Mandarin syllables, where final-initial's duration ratio is introduced to control the weight. Experiments show that the proposed algorithm is much more effective than traditional posterior log-likelihood algorithm in the Mandarin learning system. The correlation with human score achieves an increase of 11%.

Quick Individual Fitting Methods of Simplified Hearing Compensation for Elderly People

Kengo Fujita, Tsuneo Kato, Hisashi Kawai; KDDI R&D Laboratories Inc., Japan

Tue3Cap.9

For simplified hearing compensation, which combines a high-frequency boost, pitch shift and speech rate conversion, two quick individual fitting procedures, based on subjective assessments, are proposed. Since conventional individual fitting procedures impose enormous hearing comparisons on a user for decision of an appropriate parameter set, the proposed procedures split the fitting process into two steps, namely a parameter fitting step for each hearing compensation method and a comprehensive combination step, to reduce the number of comparisons. Fitting experiments with twenty elderly subjects over sixty-five showed that the proposed procedures had effect on hearing with an average of only around 20 comparisons, while the conventional procedure requires hundreds of comparisons.

An Online Adaptive Filtering Algorithm for the Vocal Joystick

Xiao Li, Jonathan Malkin, Susumu Harada, Jeff A. Bilmes, Richard Wright, James Landay; University of Washington, USA

Tue3Cap.10

This paper introduces a novel adaptive direction filtering algorithm in the Vocal Joystick (VJ) setting that utilizes context information and applies real-time inference in a continuous space. The VJ system using this algorithm is endowed with the ability to produce movements in arbitrary directions and the ability to draw smooth curves. This is in contrast to previous VJ settings whereby vowel quality was used to determine mouse movement in only a finite discrete set of directions [1].

Speaking Aid System for Total Laryngectomees Using Voice Conversion of Body Transmitted Artificial Speech

Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano; Nara Institute of Science & Technology, Japan

Tue3Cap.11

The aim of this paper is to improve the naturalness of speech using a medical device such as an electrolarynx. There are several problems associated with using existing electrolarynxes, such as the fact the loud volume of the electrolarynx itself might disturb smooth interpersonal communication, and that the generated speech is unnatural. We propose a novel speaking-aid system for total laryngectomees using a new sound source as an alternative to the existing electrolarynx and a statistical voice-conversion technique. The new sound-source unit outputs extremely small signals that cannot be heard by people around the speaker. Artificial speech is recorded with a NAM microphone through soft tissues of the head. From the result of voice conversion, the body-transmitted artificial speech is consistently converted to a more natural voice. We also demonstrate that the speech recognition performance of the proposed system substantially increases in terms of objective evaluation.
A Spanish Speech to Sign Language Translation System for Assisting Deaf-Mute People

R. San-Segundo, R. Barra, L.F. D’Haro, J.M. Montero, R. Córdoba, J. Ferreiros; Universidad Politécnica de Madrid, Spain
Tue3Cap.12

This paper describes the first experiments of a speech to sign language translation system in a real domain. The developed system is focused on the sentences spoken by an officer when assisting people in applying for, or renewing the National Identification Document (NID) and the Passport. This system translates officer explanations into sign language for deaf-mute people. The translation system is composed by a speech recognizer (for decoding the spoken utterance into a word sequence), a natural language translator (for converting a word sequence into a sequence of gestures belonging to the sign language), and a 3D avatar animation module (for playing the gestures). The field experiments have reported a 27.2% GER (Gesture Error Rate) and a 0.62 BLEU (BiLingual Evaluation Understudy).

Potential Relevance of Audio-Visual Integration in Mammals for Computational Modeling

Eeva Klintfors, Francisco Lacerda; Stockholm University, Sweden
Tue3Cap.13

The purpose of this study was to examine typically developing infants’ integration of audio-visual sensory information as a fundamental process involved in early word learning. One hundred sixty pre-linguistic children were randomly assigned to watch one of four counterbalanced versions of audio-visual video sequences. The infants’ eye-movements were recorded and their looking behavior was analyzed throughout three repetitions of exposure-test-phases. The results indicate that the infants were able to learn covariance between shapes and colors of arbitrary geometrical objects and to them corresponding nonsense words. Implications of audio-visual integration in infants and in non-human animals for modeling within speech recognition systems, neural networks and robotics are discussed.

Finding the Gaps: Applying a Connectionist Model of Word Segmentation to Noisy Phone-Recognized Speech Data

C. Anton Rytting; Ohio State University, USA
Tue3Cap.14

The Christiansen model of word segmentation [1] is a connectionist framework for modeling how infants combine multiple cues in learning and processing language. Most studies applying this model assume idealized input with adult-like representations of phonemes and features, with little or no degradation of the input signal. From these studies, it is difficult to tell if the model is robust to non-idealized, noisy input, which may correspond more closely to an infant language-learner’s experience.

This study tests the robustness of the Christiansen model by providing input from a minimally-trained phone recognizer on infant-directed speech. Some degradation of performance is observed, but the model still performs above chance. This finding represents a first step in developing more realistic input representations for models of child language acquisition.

Single Channel Speech Enhancement by Frequency Domain Constrained Optimization and Temporal Masking

Wen Jin, Michael Scordilis; University of Miami, USA
Tue3FoP.1

A speech enhancement algorithm is proposed that exploits the masking properties of the human auditory system. The enhancement is formulated as a frequency domain constrained optimization problem. The noise components of the noisy speech are suppressed by a gain function subject to the constraint that both the signal distortion and residual noise should fall below the masking thresholds. Temporal as well as simultaneous masking effects are incorporated into the estimation of masking thresholds. The enhancement algorithm was tested with speech corrupted by white Gaussian and multtalker babble noise, respectively. Its performance was evaluated by ITU PESQ scores and segmental SNR. Experimental results indicate that the proposed gain function performs slightly but consistently better than a former perceptually motivated enhancement algorithm. Greater improvement is achieved by incorporating the temporal masking effects.

Speech Enhancement Based on Residual Noise Shaping

Jong Won Shin, Seung Yeol Lee, Hwan Sik Yun, Nam Soo Kim; Seoul National University, Korea
Tue3FoP.2

In this paper, we propose a novel approach to speech enhancement, which incorporates a new criterion based on residual noise shaping. In the proposed approach, our goal is to make the residual noise perceptually comfortable although the power of the residual noise is relatively high. In contrast to the conventional techniques, the proposed approach regulates not only the power of the signal distortion and residual noise but also the spectral shape of the residual noise. A predetermined ‘comfort noise’ is provided as a target for the spectral shaping. Three different versions of enhancement algorithms adopting the proposed criterion are presented. Subjective listening test results show that the proposed algorithm outperforms the conventional spectral enhancement techniques which are based on soft decision and the noise suppression module implemented in IS-893 Selectable Mode Vocoder.

Quality Improvement of Telephone Speech by Artificial Bandwidth Expansion — Listening Tests in Three Languages

Hanna Pulakka1, Laura Laaksonen2, Paavo Alku1; 1Helsinki University of Technology, Finland; 2Nokia Research Center, Finland
Tue3FoP.3

Quality and intelligibility of narrowband telephone speech can be improved by artificial bandwidth expansion (ABE), which expands the speech bandwidth using only information available in the narrowband speech signal. This paper describes an ABE method that generates a high-band expansion using spectral folding and then modifies the magnitude spectrum of the expansion band with spline curves. The performance of the ABE algorithm was evaluated by formal listening tests in three languages: American English, Russian, and Mandarin Chinese. The results of the listening tests indicate that ABE-processed speech was preferred to narrowband speech in all tested languages.
Role of Phase Estimation in Speech Enhancement

Benjamin J. Shannon, Kuldip K. Paliwal; Griffith University, Australia

Tue3Fp.4

Typical speech enhancement algorithms that operate in the Fourier domain only modify the magnitude component. It is commonly understood that the phase component is perceptually unimportant, and thus, it is passed directly to the output.

In recent intelligibility experiments, it has been reported that the Short-Time Fourier Transform (STFT) phase spectrum can provide significant intelligibility when estimated using a window function lower in dynamic range than the typical Hamming window. Motivated by this, we investigate the role of the window function for STFT phase estimation in relation to speech enhancement.

Using a modified STFT Analysis-Modification-Synthesis (AMS) framework, we show that noise reduction can be achieved by modifying the window function used to estimate the STFT phase spectra. We demonstrate this through spectrogram plots and results from two objective speech quality measures.

Speech Enhancement Based on Spectral Estimation from Higher-Lag Autocorrelation

Benjamin J. Shannon, Kuldip K. Paliwal, Climent Nadeu; Griffith University, Australia

Tue3Fp.5

In this paper, we propose a unique approach to enhance speech signals that have been corrupted by non-stationary noises. This approach is not based on a spectral subtraction algorithm, but on an algorithm that separates the speech signal and noise signal contributions in the autocorrelation domain. We call this technique the AR-HASE speech enhancement algorithm.

In this initial study, we evaluate the performance of the new algorithm using the average PESQ score computed from 10 male utterances and 10 female utterances taken from the TIMIT database as a measure of speech quality. We test the algorithm using one broadband stationary noise and two non-stationary noises. We will show that the AR-HASE enhancement algorithm produces near transparent quality for clean speech, gives poor enhancement performance for broadband stationary noises, and gives significantly enhanced quality for the two non-stationary noises.

Noise Update Modeling for Speech Enhancement: When Do We Do Enough?

Nitish Krishnamurthy, John H.L. Hansen; University of Texas at Dallas, USA

Tue3Fp.6

In speech enhancement, it is generally assumed that if you can update your noise estimate on a frame-by-frame basis, you should achieve the highest level of enhancement performance. However, for many noise types and environmental conditions, it is not necessary to perform an update on a frame-by-frame basis to achieve superior performance if the noise structure does not change rapidly. For applications where compute/memory resources are limited, better overall speech performance could be achieved if a more reasonable update rate is estimated so that available compute/memory resources could be made available to the enhancement algorithm itself. In this study, we propose a framework to predict the noise update rate with the goal of determining the best update rate required to achieve a given quality for speech enhancement. Speech systems generally develop specialized solutions for noise which are unique to each application (i.e., recognition, speaker ID, enhancement etc.). Here we propose a model to predict the noise update rate required to achieve a given quality for enhancement. We evaluate the algorithm across a corpus of four noise types under different levels of degradation. The error between the mean observed and the mean predicted Itakura-Saito (IS) values of quality are typically between 0.06 to 1.78 IS for our model selected noise frame update rate of 1 frame every 5 frames using the Log-MMSE enhancement scheme. Finally we consider mobile and resource limited applications where such a framework would be useful.

Mapping Neural Networks for Bandwidth Extension of Narrowband Speech

A. Shahina, Yegnanarayana B.; Indian Institute of Technology Madras, India

Tue3Fp.7

This paper exploits the nonlinear mapping property of feed-forward neural networks for estimation of high frequency components (4-8kHz) of the speech signals from the band-limited (0-4kHz) signals. Cepstral coefficients are used to represent the feature vectors of each frame of data. This paper also proposes an approach that uses the autocorrelation method to derive the Linear Prediction (LP) coefficients from the estimated cepstral coefficients that are obtained from the mapping network. This method guarantees the stability of the LP synthesis filter. Informal listenings indicate the effectiveness of the proposed method for estimation of wideband frequency components of speech. The enhanced speech sounds similar to the original wideband speech. Also, it does not contain any distortion that may arise due to spectral discontinuities between adjacent frames.

Decision Directed Constrained Iterative Speech Enhancement

Amit Das, John H.L. Hansen; University of Texas at Dallas, USA

Tue3Fp.8

Earlier studies have shown that degradation due to environmental background noise is non-uniform across various phoneme classes of speech. In this study, we present an improved formulation of single channel constrained iterative speech enhancement (AutoLSP) that follows a rover based paradigm. The new approach overcomes some of the drawbacks observed earlier in the baseline AutoLSP system. First, it eliminates the sensitivity to proper determination of the terminating iteration. Second, it employs a phone level non-uniform enhancement approach which significantly improves perceptual quality of the overall utterance. Third, audible noise components are suppressed by incorporating an auditory masked threshold (AMT) framework. The proposed algorithm is evaluated using Itakura-Saito (IS) objective quality measure over four noise sources and two SNR levels. Comparative evaluations with other baseline systems (AutoLSP, log-MMSE) reveal that the new algorithm exhibits consistent quality improvement for each noise case over all phoneme classes in the TIMIT corpus. Reduction in IS distance over degraded speech is observed in the range of 35.09–46.88%. The Rover scheme outperforms AutoLSP and log-MMSE by 9.21% and 11.19% respectively using IS scores.

Adaptive Filtering for Attenuating Musical Noise Caused by Spectral Subtraction

Takahiro Murakami1, Yoshihisa Ishida2; 1Tokyo University of Agriculture & Technology, Japan; 2Meiji University, Japan

Tue3Fp.9

A method of alleviating processing distortion caused by spectral subtraction is presented. It is well known that the spectral subtraction introduces annoying artifacts, which are referred to as undesirable musical noise, in the enhanced speech. The enhancement quality of the spectral subtraction quite depends on the performance of reducing the musical noise. Our approach exploits an adaptive filter in order to eliminate such distortion. In the method, the enhanced speech obtained by the spectral subtraction is used as a reference signal of the adaptive filter. The proposed method utilizes the characteristic difference between the musical noise and speech, i.e., the majority of the frequency components consisting the musical noise have the duration shorter than those of speech. Therefore, when the convergence speed
of the adaptive filter is slower than the lifetime of the musical noise but faster than that of speech, only speech components can be tracked by the filter while the musical noise components are attenuated. Simulation results show that the proposed method can efficiently reduce the musical noise and the enhancement quality is improved in comparison with the conventional spectral subtraction.

Evaluation of Objective Measures for Speech Enhancement

Yi Hu, Philipos C. Loizou; University of Texas at Dallas, USA
Tue3FoP.10

In this paper, we evaluate the performance of several objective measures in terms of predicting the quality of noisy speech enhanced by noise suppression algorithms. The objective measures considered a wide range of distortions introduced by four types of real-world noise at two SNRs by four classes of speech enhancement algorithms: spectral subtractive, subspace, statistical-model based and Wiener algorithms. The subjective quality ratings were obtained using the ITU-T P.835 methodology designed to evaluate the speech quality along three dimensions: signal distortion, noise distortion and overall quality. This paper reports the correlations of five common objective measures with these three subjective measures. Improvements to the PESQ measure are reported along with new composite objective measures.

Performance Analysis of Various Single Channel Speech Enhancement Algorithms for Automatic Speech Recognition

Myung-Suk Song, Chang-Heon Lee, Hong-Goo Kang; Yonsei University, Korea
Tue3FoP.11

This paper analyzes the performance of various single channel speech enhancement systems when they are applied to automatic speech recognition (ASR) systems as a preprocessor. Until now the researches on speech enhancement algorithms have focused on improving the perceptual quality of speech signal. However, it has not been verified yet whether the improvements of the perceptual quality also increase the speech recognition rate. By investigating several enhancement modules designed for improving the perceptual quality, we analyze the relationship between a speech recognizer and speech enhancement systems. Simulation results show that the decision-directed method and speech absence probability (SAP) estimation proposed for improving perceptual quality influence adverse effects to the speech recognition performance.
Highly Noise Robust Text-Dependent Speaker Recognition Based on Hypothesized Wiener Filtering

Ramasubramanian V.1, Deepak Vijaywargiya2, Praveen Kumar V.1; 1Siemens Information Systems Ltd., India; 2International Institute of Information Technology, India

We propose a hypothesized Wiener filtering (HWF) algorithm for noise robust variable-text text-dependent speaker-recognition. The proposed algorithm exploits an important feature of the text — dependent mode of operation of speaker-recognition, namely, the availability of the ‘clean reference templates’ of the words of the ‘password’ text which is supposed to be the text of the input noisy speech. The proposed HWF algorithm is set within the one-pass DP framework proposed by us recently for text-dependent speaker-recognition, which enables use of multiple-templates for each word in the password. We evaluate the proposed HWF algorithm for both speaker — identification and speaker — verification using the TIDIGITS database and show that the proposed HWF algorithm has very high recognition accuracies for both additive white-noise conditions and non-stationary color noise conditions (factory, chopper and babble noises), which are also the typical conditions where conventional spectral subtraction techniques perform poorly.

Speaker Identification Under Noisy Environments by Using Harmonic Structure Extraction and Reliable Frame Weighting

Hiromasa Fujihara1, Tetsuro Kitahara1, Masataka Goto2, Kazunori Komatani1, Tetsuya Ogata1, Hiroshi G. Okuno1; 1Kyoto University, Japan; 2AIST, Japan

We present methods for automatic speaker identification in noisy environments. To improve noise robustness of speaker identification, we developed two methods, the harmonic structure extraction method and the reliable frame weighting method. The harmonic structure extraction method enables the speaker of input speech signals to be identified after environmental noise has been reduced. This method first extracts harmonic components of the speech from the sound mixtures and then resynthesizes a clean speech signal by using a sinusoidal model driven by harmonic components. The reliable frame weighting method then determines how each frame of the resynthesized speech is reliable (i.e. little influenced by environmental noises) by using two Gaussian mixture models for the speech and noise. The reliable frame weighting method then determines how each frame of the resynthesized speech is reliable (i.e. little influenced by environmental noises) by using two Gaussian mixture models for the speech and noise. The speech can be robustly identified by attaching importance to reliable frames. Experimental results with thirty speakers showed that our method was able to reduce the influences of environmental noise and achieved an error rate of 10.7%, while the error rate for a conventional method was 18.9%.

Enhancing the Performance of a GMM-Based Speaker Identification System in a Multi-Microphone Setup

Andreas Stergiou, Aristodemos Pnevmatikakis, Lazaros C. Polymenakos; Athens Information Technology, Greece

In this paper the speaker identification system developed at Athens Information Technology is presented. It is based on the Gaussian Mixture modeling of the Mel-Frequency Cepstral Coefficients of speech. Starting from this basic algorithm, we describe and discuss two significant modifications that have resulted in performance enhancements, in terms of both processing speed and identification accuracy. We present the performance of our system in the recent CLEAR 2006 evaluation workshop and also discuss approaches to further improve our system by fusing decisions derived from a multitude of sensors in a multi-microphone setup.

Discriminative Adaptation for Speaker Verification

C. Longworth, M.J.F. Gales; University of Cambridge, UK

Speaker verification is a binary classification task to determine whether a claimed speaker uttered a phrase. Current approaches to speaker verification tasks typically involve adapting a general speaker Universal Background Model (UBM), normally a Gaussian Mixture Model (GMM), to model a particular speaker. Verification is then performed by comparing the likelihoods from the speaker model to the UBM. Maximum A-Posteriori (MAP) is commonly used to adapt the UBM to a particular speaker. However speaker verification is a classification task. Thus, robust discriminative-based adaptation schemes should yield gains over the standard MAP approach. This paper describes and evaluates two discriminative approaches to speaker verification. The first is a discriminative version of MAP based on Maximum Mutual Information (MMI-MAP). The second is to use an augmented-GMM (A-GMM) as the speaker-specific model. The additional, augmented, parameters are discriminatively, and robustly, trained using a maximum margin estimation approach. The performance of these models is evaluated on the NIST 2002 SRE dataset. Though no gains were obtained using MMI-MAP, the A-GMM system gave an Equal Error Rate (EER) of 7.31%, a 30% relative reduction in EER compared to the best performing GMM system.

Within-Class Covariance Normalization for SVM-Based Speaker Recognition

Andrew O. Hatch 1, Sachin Kajarekar2, Andreas Stolcke2; 1International Computer Science Institute, USA; 2SRI International, USA

This paper extends the within-class covariance normalization (WCCN) technique described in [1, 2] for training generalized linear kernels. We describe a practical procedure for applying WCCN to an SVM-based speaker recognition system where the input feature vectors reside in a high-dimensional space. Our approach involves using principal component analysis (PCA) to split the original feature space into two subspaces: a low-dimensional “PCA space” and a high-dimensional “PCA-complement space.” After performing WCCN in the PCA space, we concatenate the resulting feature vectors with a weighted version of their PCA-complements. When applied to a state-of-the-art MLLR-SVM speaker recognition system, this approach achieves improvements of up to 22% in EER and 28% in minimum decision cost function (DCF) over our previous baseline. We also achieve substantial improvements over an MLLR-SVM system that performs WCCN in the PCA space but discards the PCA-complement.

A New Set of Features for Text-Independent Speaker Identification

Carol Y. Espy-Wilson, Sandeep Manocha, Srikanth Vishnubhotla; University of Maryland, USA

The success of a speaker identification system depends largely on the set of features used to characterize speaker-specific information. In this paper, we discuss a small set of low-level acoustic parameters that capture information about the speaker’s source, vocal tract size and vocal tract shape. We demonstrate that the set of eight acoustic parameters has comparable performance to the standard sets of 26 or 39 MFCCs for the speaker identification task. Gaussian Mixture Models were used for constructing speaker models.
Wed1A2O: Robust ASR
Oral Session, Allegheny Ballroom, center, 10:30 - 12:30
Chair: Michael L. Seltzer, Microsoft Research, USA

Rapid Speaker Adaptation Using Regression-Tree Based Spectral Peak Alignment
Shizhen Wang¹, Xiaodong Cui², Abeer Alwan¹;
¹University of California at Los Angeles, USA; ²IBM T.J. Watson Research Center, USA

In this paper, regression-tree based spectral peak alignment is proposed for rapid speaker adaptation using the linearization of VTLN. Two different regression classes are investigated: phonetic classes (using combined knowledge and data-driven techniques) and mixture classes. Compared to MLLR and VTLN, improved performance can be obtained for both supervised and unsupervised adaptations on both medium vocabulary and transient noise, and that the use of the synchrony mechanism provides further improvement beyond that which is provided by the mean rate mechanism. Experimental results show that the performance improvements can be achieved even with limited adaptation data.

Physiologically-Motivated Synchrony-Based Processing for Robust Automatic Speech Recognition
Chanwoo Kim, Yu-Hsiang Chiu, Richard M. Stern;
Carnegie Mellon University, USA

In this paper we address the problem of robust speech recognition. We propose a new method based on the individual variance adaptation of frequency filtered parameters to reduce the deleterious effects of additive narrow-band noise. The method can be interpreted as a spectral weighting that assigns increased importance to the most reliable spectral components, typically the spectral peaks. The experiments confirm that the suggested method results in significantly improved recognition rates for additive narrow-band noise.

Recent Progress on the Discriminative Region-Dependent Transform for Speech Feature Extraction
Bing Zhang, Spyros Matsoukas, Richard Schwartz; BBN Technologies, USA

The region-dependent transform (RDT) is a feature extraction method for speech recognition that employs the Minimum Phoneme Error (MPE) criterion to optimize a set of feature transforms, each concentrating on a region of the acoustic space. Previous results have shown that RDT gives significant recognition-error reduction in a large vocabulary speaker-independent (SI) system. As a follow-up investigation, this paper presents the recent progress of applying RDT in speaker-adaptive training (SAT). Similar to previous SI results, the integration of RDT with SAT yields 7% relative improvement in word error rate (WER). Also, theoretical comparisons are made between RDT and other discriminative feature extraction methods, including the improved version of the feature-space MPE (fMPE) that uses the “mean-offsets” as additional input features.

Improved Warping-Invariant Features for Automatic Speech Recognition
Jan Rademacher, Matthias Wächtcher, Alfred Mertins;
University of Oldenburg, Germany

In this paper, we extend a previously introduced method for the generation of vocal tract length invariant (VTLI) features. The novelty is a reduction of the number of obtained invariances to the more desired ones, which results in a significant improvement of recognition rates. In experiments on the TIMIT database, the enhanced discrimination capabilities and robustness to mismatches between training and test conditions are shown.

Sub-Ward Unit Based Non-Audible Speech Recognition Using Surface Electromyography
Matthias Walliczek¹, Florian Kraft¹, Szu-Chen Jou²,
Tanja Schultz², Alex Waibel¹; ¹Universität Karlsruhe, Germany; ²Carnegie Mellon University, USA

In this paper we present a novel approach for a surface electromyographic speech recognition system based on sub-word units. Rather than using full word models as integrated in our previous work we propose here smaller sub-word units as prerequisites for large vocabulary speech recognition. This allows the recognition of words not seen in the training set based on seen sub-word units. Therefore we report on experiments with syllables and phonemes as sub-word units. We also developed a new feature extraction method that gains significant improvement for words and sub-word units.

Individual On-Line Variance Adaptation of Frequency Filtered Parameters for Robust ASR
Jesús Vicente-Peña¹, Fernando Díaz-de-María¹, Bastiaan Kleijn²; ¹Universidad Carlos III de Madrid, Spain; ²KTH, Sweden

In this paper we address the problem of robust speech recognition. We propose a new method based on the individual variance adaptation of frequency filtered parameters to reduce the deleterious effects of additive narrow-band noise. The method can be interpreted as a spectral weighting that assigns increased importance to the most reliable spectral components, typically the spectral peaks. The experiments confirm that the suggested method results in significantly improved recognition rates for additive narrow-band noise.
Effects of Featural Similarity and Overlap Position on Lexical Confusions and Overt Similarity Judgments
Sarah C. Creel, Delphine Dahan, Daniel Swingley; University of Pennsylvania, USA

Spoken word recognition involves selecting one word out of many competing candidate word-forms. However, there is disagreement as to what candidates actually compete with each other given a particular signal, though similarity is commonly invoked. We investigate whether the word similarity literature, drawing from explicit similarity judgments of word forms, can speak to this current state of confusion. Many word similarity models emphasize the primacy of featural overlap. With a single set of stimuli, we find that featural overlap and overlap position influence a non-time-pressured lexical learning task as well as explicit similarity judgments. The implications for appropriate metrics of similarity in word recognition and models of word similarity are discussed.

Word Structure and Tone Perception in Mandarin
Hansjörg Mixdorff1, Yu Hu2; 1TFH Berlin, Germany; 2University of Science & Technology of China, China

This paper presents results concerning the relationship between word structure in terms of number of syllables and tonal realization in Mandarin. It examines whether the fact that a word (in our context a prosodic word) is more complex implies certain tonal reductions. Our hypothesis is that a monosyllabic word will be uttered more carefully than a polysyllabic word due to the potentially larger number of possibly confusable words. We also examine whether the total number of syllables in a word has an effect, creating more tonal reductions in longer than in shorter words. A database of Mandarin originally designed for concatenative speech synthesis and segmented into prosodic words was statistically analyzed regarding the occurrences of syllable/tone combinations in prosodic words of varying length. 10 sets of syllables were selected comprising all four tones of Mandarin who had to decide which tone they perceived. The results of the perception test indicate, inter alia, that perception of syllables categorized into three sets, depending on the consonant. The consonant confusions correlate closely with the high frequency spectra of the consonants while the vowel confusions correlate with the vowel durations and the second formant frequencies.

Identification of Regional Accents in French: Perception and Categorization
Cécile Woehrling, Philippe Boula de Mareuil; LIMSI, France

This article is dedicated to the perceptual identification of French varieties by listeners from the surroundings of Paris and Marseilles. It is based on the geographical localization of about forty speakers from 6 Francophone regions: Normandy, Vendee, Romand Switzerland, Languedoc and the Basque Country. Contrary to the speech type (read or spontaneous speech) and the listeners’ region of origin, the speakers’ degree of accentedness has a major effect and interacts with the speakers’ age. The origin of the oldest speakers (who have the strongest accent according to Paris listeners’ judgments) is better recognized than the origin of the youngest speakers. However confusions are frequent among the Southern varieties. On the whole, three accents can be distinguished (by clustering and multidimensional scaling techniques): Northern French, Southern French and Romand Swiss.

Consonant and Vowel Confusions in Speech-Weighted Noise
Sandep Phatak, Jont B. Allen; University of Illinois at Urbana-Champaign, USA

Confusion analysis for closed-set recognition of 64 consonant-vowel (CV) sounds, spoken by 18 talkers, is presented. In presence of speech-weighted noise the confusions patterns of the CV syllables categorize into three sets, depending on the consonant. The consonant confusions correlate closely with the high frequency spectra of the consonants while the vowel confusions correlate with the vowel durations and the second formant frequencies.

Accident – Execute: Increased Activation in Nonnative Listening
Mirjam Broersma; Max Planck Institute for Psycholinguistics, The Netherlands

Dutch and English listeners’ perception of English words with partially overlapping onsets (e.g., accident – execute) was investigated. Partially overlapping words remained active longer for nonnative listeners, causing an increase of lexical competition in nonnative compared with native listening.

Estimation of the Quality Dimension “Directness/Frequency Content” for the Instrumental Assessment of Speech Quality
Kirstin Scholz1, Marcel Wältermann2, Lu Huo1, Alexander Raake3, Sebastian Möller3, Ulrich Heute1; 1Christian-Albrechts-Universität zu Kiel, Germany; 2Ruhr-Universität Bochum, Germany; 3Technische Universität Berlin, Germany

The presented work aims at an instrumental method for the assessment of speech-quality of modern telecommunication networks that features two functions: the prediction of a system’s overall speech-quality as well as the analysis of the overall speech-quality. The introduced method is based on the estimation of quality dimensions, each describing a particular characteristic relevant for the quality of speech signals. For narrowband telephone-speech three relevant dimensions have been identified: “directness/ frequency content”, “continuity”, and “noisiness”. The estimates of the single dimensions allow for an analysis of speech quality while together forming a model for the overall quality.

This paper presents a method for the estimation of the quality dimension “directness/frequency content”. The estimator is based on two parameters considered suitable to measure “directness/ frequency content” of a speech signal y[k] that has been transmitted over a telecommunication system. The estimates DF of this dimension show correlations of ρ > 0.93 with the results of corresponding auditory tests.
Summarization Evaluation for Text and Speech: Issues and Approaches

Ani Nenkova; Stanford University, USA

This paper surveys current text and speech summarization evaluation approaches. It discusses advantages and disadvantages of these, with the goal of identifying summarization techniques most suitable to speech summarization. Precision/recall schemes, as well as summary accuracy measures which incorporate weightings based on multiple human decisions, are suggested as particularly suitable in evaluating speech summaries.

Summarization of Spontaneous Conversations

Xiaodan Zhu, Gerald Penn; University of Toronto, Canada

Most speech summarization research is conducted on broadcast news. In our viewpoint, spontaneous conversations are a more “typical” speech source that distinguishes speech summarization from text summarization, and hence a more appropriate domain for studying speech summarization. For example, spontaneous conversations contain more spoken-language characteristics, e.g. disfluencies and false starts. They are also more vulnerable to ASR errors. Previous research has studied some aspects of this type of data, but this paper addresses the problem further in several important respects. First, we summarize spontaneous conversations with features of a wide variety that have not been explored before. Second, we examine the role of disfluencies in summarization, which in all previous work was either not explicitly handled or removed as noise. Third, we breakdown and analyze the impact of WER on the individual features for summarization.

Perplexity Based Linguistic Model Adaptation for Speech Summarisation

Pierre Chattain, Edward Whittaker, Joanna Mrozinski, Sadaoki Furui; Tokyo Institute of Technology, Japan

The performance of automatic speech summarisation has been improved in previous experiments by using linguistic model adaptation. One of the problems encountered was the high computational cost and low efficiency of the development phase. In this paper we compare our original development approach of evaluating summaries produced by an exhaustive search over all parameters with a much faster development method using an expectation maximization algorithm that minimizes perplexity in order to find the optimal combination of linguistic models for the speech summarisation task. Perplexity proves to be sufficiently correlated to the objective evaluation metrics used in the summarisation literature that it can be used in this fashion. For a much reduced computational cost (approximately 500 times faster), final relative improvements are very similar to those previously obtained, ranging from 1.5% to 21.3% on all investigated metrics for summaries made from automatic speech recogniser transcriptions.

Multi-Layered Summarization of Spoken Document Archives by Information Extraction and Semantic Structuring

Lin-shan Lee, Sheng-yi Kong, Yi-cheng Pan, Yi-sheng Fu, Yu-tsun Huang; National Taiwan University, Taiwan

The spoken documents are very difficult to be shown on the screen, and very difficult to retrieve and browse. It is therefore important to develop technologies to summarize the entire archives of the huge quantities of spoken documents in the network content to help the user in browsing and retrieval. In this paper we propose a complete set of multi-layered technologies to handle at least some of the above issues: (1) Automatic Generation of Titles and Summaries for each of the spoken documents, such that the spoken documents become much more easier to browse, (2) Global Semantic Structuring of the entire spoken document archive, offering to the user a global picture of the semantic structure of the archive, and (3) Query-based Local Semantic Structuring for the subset of the spoken documents retrieved by the user’s query, providing the user the detailed semantic structure of the relevant spoken documents given the query he entered. The Probabilistic Latent Semantic Analysis (PLSA) is found to be helpful, and an initial prototype system for the functions mentioned above has been successfully developed, in which the broadcast news archive in Mandarin Chinese is taken as the example archive.

Soundbite Detection in Broadcast News Domain

Sameer Maskey, Julia Hirschberg; Columbia University, USA

In this paper, we present results of a study designed to identify soundbites in Broadcast News. We describe a Conditional Random Field-based model for the detection of these included speech segments uttered by individuals who are interviewed or who are the subject of a news story. Our goal is to identify direct quotations in spoken corpora which can be directly attributable to particular individuals, as well as to associate these soundbites with their speakers. We frame soundbite detection as a binary classification problem in which each turn is categorized either as a soundbite or not. We use lexical, acoustic/prosodic and structural features on a turn level to train a CRF. We performed a 10-fold cross validation experiment in which we obtained an accuracy of 67.4% and an F-measure of 0.566 which is 20.9% and 38.6% higher than a chance baseline.

Dialogue Act Compression via Pitch Contour Preservation

Gabriel Murray, Steve Renals; University of Edinburgh, UK

This paper explores the usefulness of prosody in automatically compressing dialogue acts from meeting speech. Specifically, this work attempts to compress utterances by preserving the pitch contour of the original whole utterance. Two methods of doing this are described in detail and are evaluated subjectively using human annotators and objectively using edit distance with a human-authored gold-standard. Both metrics show that such a prosodic approach is much better than the random baseline approach and significantly better than a simple text compression method.

Wed1WeS  : Speech Summarization
Special Session, Westmoreland Room, 10:30 – 12:30
Chair: Sadaoki Furui, Tokyo Institute of Technology, Japan
The Subspace Precision and Mean model (SPAM) is a way of representing environmental information explicitly from the useful information for discrimination. It is proposed that a feature parameter used in pattern recognition involves categorical information and also environmental information. A well-known HLDA tries to extract useful information (UI) to represent categorical information from the feature parameter. However, environmental information is still remained in the UI parameters extracted by HLDA, and it causes slight degradation in performance. This is because HLDA does not handle the environmental information explicitly. The proposed MHLDA also tries to extract UI like HLDA, but it handles environmental information explicitly. This handling makes MHLDA-based UI parameter less influenced of environment. However, as compensation, in MHLDA, the categorical information is little bit destroyed. In this paper, we try to combine HLDA-based UI and MHLDA-based UI for pattern recognition, and draw benefit of both parameters. Experimental results show the effectiveness of this combining method.

**Time-Dependent Cross-Probability Model for Multi-Environment Model Based Linear Normalization**

Luis Buera¹, Eduardo Lleida¹, Juan A. Nolasco-Flores², Antonio Miguel¹, Alfonso Ortega¹;¹Universidad de Zaragoza, Spain;²ITESM, Mexico

In a previous work, Multi-Environment Model based Linear Normalization, MELIN, was presented and it was proved to be effective to compensate environment mismatch. MELIN is an empirical feature vector normalization which models clean and noisy spaces by Gaussian Mixture Models (GMMs). In this algorithm, the probability of the clean model Gaussian, given the noisy model one and the noisy feature vector (cross-probability model) is a critical point. In the previous work the cross-model probability was approximated as time-independent. In this paper, a time-dependent estimation of the cross-probability model based on GMM is proposed. Some experiments with SpeechDat Car database were carried out in order to study the performance of the proposed estimation in a real acoustic environment. MELIN with time-independent cross-probability model reached 70.21% of mean improvement in Word Error Rate (WER), however, when time-dependent cross-probability model based on GMM was applied, the mean improvement in WER went up to 78.47%.

**SPAM and Full Covariance for Speech Recognition**

Daniel Povey; IBM T.J. Watson Research Center, USA

The Subspace Precision and Mean model (SPAM) is a way of representing Gaussian precision and mean values in a reduced dimension. This paper presents some large vocabulary experiments with SPAM and introduces an efficient way to optimize the SPAM basis. We present experiments comparing SPAM, diagonal covariance and full covariance models on a large vocabulary task. We also give explicit formulae for an implementation of SPAM.
Pronunciation Variant-Based Multi-Path HMMs for Syllables

Annika Hämaläinen, Louis ten Bosch, Lou Boves; Radboud Universiteit Nijmegen, The Netherlands
Wed1BuP.8

Recent research suggests that it is more appropriate to model pronunciation variation with syllable-length acoustic models than with context-dependent phones. Due to the large number of factors contributing to pronunciation variation at the syllable level, the creation of multi-path model topologies appears necessary. In this paper, we propose a novel approach for constructing multi-path models for frequent syllables. The suggested approach uses phonetic knowledge for the initialisation of the parallel paths, and a data-driven solution for their re-estimation. When applied to 94 frequent syllables in a 37-hour corpus of Dutch read speech, it leads to improved recognition performance when compared with a triphone recogniser of similar complexity.

A New State-Dependent Phonetic Tied-Mixture Model with Head-Body-Tail Structured HMM for Real-Time Continuous Phoneme Recognition System

Junho Park, Hanseok Ko; Korea University, Korea
Wed1BuP.9

An acoustic model for a real-time continuous phoneme recognition system must exhibit the following desirable features: an ability to minimize the recognition performance degradation while solving the model complexity problem to confine the delay to a minimum in recognition process. To cope with the challenges, we introduce the state-dependent Phonetic Tied-Mixture (PTM) model with Head-Body-Tail (HBT) structured HMM as an acoustic model optimization. The proposed acoustic modeling method shows a significant improvement in recognition performance and becomes a solution to the sparse training data problem and the model complexity problem. Moreover, defining the exceptional Gaussian mixtures in tying process achieves a drastic reduction in phoneme error rate compared to traditional state-dependent PTM method. In this paper, we describe the new acoustic model optimization procedure and show the outstanding performance evaluation results for real-time continuous phoneme recognition system.

Conversion from Phoneme Based to Grapheme Based Acoustic Models for Speech Recognition

Andrä Žgank, Zdravko Kačič; University of Maribor, Slovenia
Wed1BuP.10

This paper focuses on acoustic modeling in speech recognition. A novel approach how to build grapheme based acoustic models with conversion from existing phoneme based acoustic models is proposed. The grapheme based acoustic models are created as weighted sum from monophone acoustic models. The influence of particular monophone is determined with the phoneme to grapheme confusion matrix. Further, the context-dependent acoustic models are being trained within the grapheme training procedure. The decision tree based clustering approach is used to tie similar states. A modified data-driven method for generation of grapheme broad classes needed during the initialization of decision tree is being applied. The data-driven broad classes are created using the grapheme based confusion matrix. All experiments were performed with the Slovenian language (1000 FDB SpeechDat(R) database), which is a highly inflectional language with no fixed set of rules for grapheme to phoneme conversion. The achieved results showed improvements of speech recognition results with the proposed methods.

Phone Vector DHMM to Decode a Phone Recognizer’s Output

Bong-Wan Kim, Dae-Lim Choi, Yongnam Um, Yong-Ju Lee; Wonkwang University, Korea
Wed1BuP.11

In this paper we introduce a Phone Vector Discrete HMM (PVDHMM) that decodes a phone recognizer’s output. The proposed PVDHMM treats a phone recognizer as a vector quantizer whose codebook size is equal to the size of its phone set. To examine the proposed method we perform two experiments. First, the output of a phone recognizer is recognized by the PVDHMM, and its results are compared with those of a continuous speech recognizer (CSR). Second, to investigate its potential application in the field of open-vocabulary spoken document retrieval, a retrieval experiment through word spotting is carried out on the output of a phone recognizer, and its results are compared with those of retrieval through the phone-based vector space model.

Combining Multiple-Sized Sub-Word Units in a Speech Recognition System Using Baseform Selection

T. Nagarajan, P. Vijayalakshmi, Douglas O’Shaughnessy; Université du Québec, Canada
Wed1BuP.12

A longer-sized sub-word unit is known to be a better candidate in the development of a continuous speech recognition system. However, the basic problem with such units is the data sparsity. To overcome this problem, researchers have tried to combine longer-sized sub-word unit models with phoneme models. In this paper, we have considered only frequently occurring syllables and VC (Vowel + Consonant) units, and phone-sized units (monophones and triphones) for the development of a continuous speech recognition system. In such a case, even for a single pronunciation of a word, there can be multiple representational baseforms in the lexicon, each with different-sized units. We show that a considerable improvement in recognition performance can be achieved if the baseforms are selected properly. Out of all possible baseforms for a given word in the lexicon, the baseform that maximizes the acoustic likelihood, for possible sub-word unit concatenations to make a word, alone is considered. In the baseline systems’ word-lexicon, like pure monophone or triphone-based systems, since only the acoustically weaker baseforms are replaced by baseforms with longer-sized units, the resultant performance is guaranteed to be better than that of baseline systems. The preliminary experiments carried out on the TIMIT speech corpus show a considerable improvement in the recognition performance over a pure monophone/triphone-based systems when the larger-sized units are combined using proper selection of baseforms.

Local Transformation Models for Speech Recognition

Antonio Miguel 1, Eduardo Lleida 1, Alfons Juan 2, Luis Buera 1, Alfonso Ortega 1, Oscar Saz 1; 1 Universidad de Zaragoza, Spain; 2 Universitat Politècnica de València, Spain
Wed1BuP.13

This paper presents a novel acoustic modeling framework that naturally extends the Hidden Markov Model (HMM) approach. The novel models reduce the errors caused by speaker variability by means of a local spectral mismatch reduction. A more complex and flexible speech production scheme can be assumed, in which the local temporal and frequency elastic deformations of the speech are captured by the model. In the new framework the states of a standard HMM, which are usually associated with temporal transitions, are expanded so that a new degree of freedom for the model is provided and it is then possible to estimate an optimum frequency warping factor at the same time as the decoder finds the best state sequence. In the local spectral warping based models the states become time-frequency related states and the number of parameters of the model is comparable to the standard HMM since they share a certain amount of parameters as it will be shown. The novel models are evaluated in the noise-free TIDIGITS
corpus, which includes connected digits uttered by male, female and children. It has been found that, under speaker group (age-gender) mismatch conditions, the local frequency warping reduced Word Error Rate (WER) in mean by a 70%, using the initial models. When matched speaker group conditions were tested the error was reduced in mean in a 9.7% after reestimating the models.

Wed1CaP : Large Vocabulary Speech Recognition
Poster Session, Cambria Room, 10:30 – 12:30
Chair: Alex Acero, Microsoft Research, USA

Online Speech Detection and Dual-Gender Speech Recognition for Captioning Broadcast News
Toru Imai, Shohei Sato, Akio Kobayashi, Kazuo Onoe, Shinichi Homma; NHK, Japan
Wed1CaP.1
This paper describes two new methods, online speech detection and dual-gender speech recognition, for captioning broadcast news. The proposed online speech detection performs dual-gender phoneme recognition and detects a start-point and an end-point based on the ratio between the cumulative phoneme likelihood and the cumulative non-speech likelihood with a very small delay from the audio input. As soon as the start-point is detected, the subsequent continuous speech recognizer with paralleled gender-dependent acoustic models starts a search using gender change information from the preceding phoneme recognizer to reduce computational cost. Speech recognition experiments on conversational commentaries and field reporting from Japanese broadcast news showed that the proposed speech detection method was effective in reducing false segmentations and also recognition errors in comparison with a conventional method using adaptive energy thresholds. The proposed dual-gender speech recognition with the new speech detection significantly reduced the word error rate by 11.2% relative to a conventional gender-independent system, while keeping the computational cost in real-time.

Automatic Alignment and Error Correction of Human Generated Transcripts for Long Speech Recordings
Timothy J. Hazen; Massachusetts Institute of Technology, USA
Wed1CaP.2
In this paper we examine the issues of aligning and correcting approximate human generated transcripts for long audio files. Accurate time-aligned transcriptions help provide easier access to audio materials by aiding downstream applications such as the indexing, summarizing and retrieving of audio segments. Accurate time alignments are also necessary when incorporating audio data into the training data for a speech recognizer’s acoustic model. We provide some initial analysis of manual transcriptions which show that there can be significant differences between the “approximate” manual transcriptions generated by typical commercial transcription services and what was actually spoken in the recording. We then present a new alignment approach for approximate transcriptions of long audio files which is designed to discover and correct errors in the manual transcription during the alignment process.

Improving Speech Recognition Accuracy with Multi-Confidence Thresholding
Shuangyu Chang; Tellme Networks Inc., USA
Wed1CaP.3
Confidence-based thresholding plays an important role in practical speech recognition applications. Most previous works have focused on directly improving confidence estimation within the recognition engine. A complementary approach that does not require access to recognizer internal is to optimize confidence threshold settings. This paper describes a general multi-confidence thresholding algorithm that automatically learns different confidence thresholds for different utterances, based on discreet or continuous features associated with a speech utterance. The algorithm can be applied to any speech recognition engine with a confidence output. A learned multi-threshold setting is guaranteed to perform at least as well as a baseline single-threshold system on training data. A significant improvement on overall accuracy can often be obtained on test data, as demonstrated with experiments on two real-world applications.

Conceptual Decoding from Word Lattices: Application to the Spoken Dialogue Corpus MEDIA
Christophe Servan, Christian Raymond, Frédéric Béchet, Pascal Nocéra; LIA, France
Wed1CaP.4
Within the framework of the French evaluation program MEDIA on spoken dialogue systems, this paper presents the methods proposed at the LIA for the robust extraction of basic conceptual constituents (or concepts) from an audio message. The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the Automatic Speech Recognition (ASR) process. This approach allows us to keep the probabilistic search space on sequences of words produced by the ASR module and to project it to a probabilistic search space of sequences of concepts. This paper presents the first ASR results on the French spoken dialogue corpus MEDIA, available through ELSA. The experiments made on this corpus show that the performance reached by our approach is better than the traditional sequential approach that looks first for the best sequence of words before looking for the best sequence of concepts.

Improving the Performance of Out-of-Vocabulary Word Rejection by Using Support Vector Machines
Shilei Huang, Xiang Xie, Jingming Kuang; Beijing Institute of Technology, China
Wed1CaP.5
Support Vector Machines (SVM) represents a new approach to pattern classification developed from the theory of structural risk minimization [1]. In this paper, we propose an approach to improve the performance of confidence measurements for out-of-vocabulary word rejection by using SVM. Confidence measures are computed from the information of n-best candidates and anti-word by a Hidden Markov Model (HMM) based speech recognizer. The acceptance/rejection decision for a word is based on the confidence score which is provided by SVM classifier. And the decision is performed for each word in vocabulary separately. The performance of the proposed SVM classifier is compared with method based on posterior probability and anti-word probability. Experiments of Mandarin command recognition have shown that better performance can be obtained when using the proposed method.

Robust Phone Lattice Decoding
Kris Demuynck, Dirk Van Compernolle, Hugo Van hamme; Katholieke Universiteit Leuven, Belgium
Wed1CaP.6
Most ASR systems adopt an all-in-one approach: acoustic model, lexicon and language model are all applied simultaneously, thus forming a single large search space. This way, both lexicon and language model help in constraining the search at an early stage which greatly improves its efficiency. However, such close integration comes at a cost: all resources must be kept simple. Achieving higher accuracy in unconstrained LVCSR tasks will require more complex resources while at the same time the ‘unconstrainedness’ of the task reduces the effectiveness of the all-in-one approach. Therefore, we propose a modular two-layered architecture. First, a pure acoustic-phonemic search generates a dense phone network. Next a robust decoder finds those words from the lexicon that match well with the phone sequences encoded in the phone network. In this paper we investigate
the properties the robust word decoder must have and we propose an efficient search algorithm.

**Imperfect Transcript Driven Speech Recognition**

*Benjamin Lecouteux, Georges Linares, Pascal Nocera, Jean-Francois Bonastre; LIA, France*

Wed1CaP.7

In many cases, textual information can be associated with speech signals such as movie subtitles, theater scenarios, broadcast news summaries etc. This information could be considered as approximated transcripts and corresponds rarely to the exact word utterances. The goal of this work is to use this kind of information to improve the performance of an automatic speech recognition (ASR) system. Multiple applications are possible: to follow a play with closed caption aligned to the voice signal (while respecting to performer variations) to help deaf people, to watch a movie in another language using aligned and corrected closed captions, etc. We propose in this paper a method combining a linguistic analysis of the imperfect transcripts and a dynamic synchronization of these transcripts inside the search algorithm.

The proposed technique is based on language model adaptation and on-line synchronization of the search algorithm. Experiments are carried out on an extract of the ESTER evaluation campaign [4] database, using the LIA Broadcast News system. The results show that the transcript-driven system outperforms significantly both the original recognizer and the imperfect transcript itself.

**New Improvements in Decoding Speed and Latency for Automatic Captioning**

*Jian Xue, Rusheng Hu, Yunxin Zhao; University of Missouri-Columbia, USA*

Wed1CaP.8

In this paper, we present new improvements in decoding speed and latency for automatic captioning in telehealth. Complementary local word confidence scores are used to prune uncompetitive search paths. Subspace distribution clustering hidden Markov modeling (SDCHMM) is used for fast generation of acoustic and local confidence scores, where overlap accumulative probability (OAP) is used to measure the similarity of Gaussian pdf's in SDCHMM. We propose to use pre-backtrace based on detection of prosodic boundaries defined by unfilled pauses, filled pauses, as well as pitch contour to decrease latency. Experiments were conducted on a telehealth captioning task with vocabulary sizes of 21K and 46K. The proposed methods led to 33% improvement in decoding speed without loss of word accuracy, and to 3 folds of decrease in maximum latency with about 1.6% loss of word accuracy.

**Colloquial Iraqi ASR for Speech Translation**

*Shirin Saleem, Rohit Prasad, Prem Natarajan; BBN Technologies, USA*

Wed1CaP.9

In this paper we describe a real-time speech recognition system developed for colloquial Iraqi Arabic. This system is currently used in our speech-to-speech translation system configured for bi-directional communication in English and Iraqi on a laptop. We present experimental results on Iraqi utterances from different speech-to-speech translation domains, and analyze the usefulness of acoustic and language modeling data from different domains. We highlight the improvements obtained by modeling techniques that are language-independent, such as lattice-based discriminative training and domain-biased language model interpolation. In addition, we report on initial experiments we have performed to address specific challenges posed by Iraqi for speech recognition such as absence of short vowels and multiple forms of glottal stop, or the hamza, in the written form.

**Reducing Computation on Parallel Decoding Using Frame-Wise Confidence Scores**

*Tomohiro Hakamata, Akinobu Lee, Yoshikiko Nankaku, Keiichi Tokuda; Nagoya Institute of Technology, Japan*

Wed1CaP.10

Parallel decoding based on multiple models has been studied to cover various conditions and speakers at a time on a speech recognition system. However, running many recognizers in parallel applying all models causes the total computational cost to grow in proportion to the number of models. In this paper, an efficient way of finding and pruning unpromising decoding processes during search is proposed. By comparing temporal search statistics at each frame among all decoders, decoders with relatively unmatched model can be pruned in the middle of recognition process to save computational cost. This method allows the model structures to be mutually independent. Two frame-wise pruning measures based on maximum hypothesis likelihoods and top confidence scores respectively, and their combinations are investigated. Experimental results on parallel recognition with an acoustic models showed that by using the both criteria, the total computational cost was reduced to 36.53% compared to full computation without degrading the recognition accuracy.

**Posterior Based Keyword Spotting with a priori Thresholds**

*Hamed Ketabdar, Jithendra Vepa, Samy Bengio, Herve Bourlard; IDIAP Research Institute, Switzerland*

Wed1CaP.11

In this paper, we propose a new posterior based scoring approach for keyword and non keyword (garbage) elements. The estimation of these scores is based on HMM state posterior probability definition, taking into account long contextual information and the prior knowledge (e.g. keyword model topology). The state posteriors are then integrated into keyword and garbage posteriors for every frame. These posteriors are used to make a decision on detection of the keyword at each frame. The frame level decisions are then accumulated (in this case, by counting) to make a global decision on having the keyword in the utterance. In this way, the contribution of possible outliers are minimized, as opposed to the conventional Viterbi decoding approach which accumulates likelihoods and the prior knowledge. Experimental results on keywords from the Conversational Telephone Speech (CTS) and Numbers'95 databases are reported. Results show that the new scoring approach leads to better trade off between true and false alarms compared to the Viterbi decoding approach, while also providing the possibility to pre-calculate keyword specific spotting thresholds related to the length of the keywords.

**A Multi-Pass Error Detection and Correction Framework for Mandarin LVCSR**

*Zhengyu Zhou, Helen M. Meng, Wai Kit Lo; Chinese University of Hong Kong, China*

Wed1CaP.12

We previously proposed a multi-pass framework for Large Vocabular Continuous Speech Recognition (LVCSR). The objective of this framework is to apply sophisticated linguistic models for recognition, while maintaining a balance between complexity and efficiency. The framework is composed of three passes: initial recognition, error detection and error correction. This paper presents and evaluates a prototype of the multi-pass framework based on Mandarin dictation. In this prototype, the first pass recognizes speech with a well-trained state-of-the-art recognizer incorporating an efficient language model; the second pass detects recognition errors by a new three-step error detection procedure; and the third pass corrects errors detected in those lightly erroneous utterances by a novel error correction approach. The error correction algorithm corrects recognition errors by first creating candidate lists for errors, and then re-ranking the candidates with a combined model of mutual information and trigram. Mandarin dictation experiments show a relative reduction of 4% in
character error rate (CER) over the initial recognition performance based on those light erroneous utterances detected.

**Continual On-Line Monitoring of Czech Spoken Broadcast Programs**

**Jan Nouza, Jindrich Zdansky, Petr Cerva, Jan Kolorenč; Technical University of Liberec, Czech Republic**

Wed1CaP.13

In the paper we describe the development of the first practical system that performs automatic on-line monitoring of Czech broadcast stations. It is based on our own speech recognition server that operates with 300K word lexicon and 2.3 RT factor. For true on-line service, several servers are connected to the platform that controls acoustic stream segmentation, distribution of data to the servers, collection of results and production of the final transcription. We show practical results achieved on different types of broadcast programs, such as news (21% WER), parliament debates (21% WER) and talk-shows (34%).

**Wed1FoP : Speech/Noise/Music Segmentation**

**Poster Session, Allegheny Ballroom Foyer, 10:30 – 12:30**

**Chair: Carol Y. Espy-Wilson, University of Maryland, USA**

**Fast SVM Training Based on the Choice of Effective Samples for Audio Classification**

**Shilei Zhang, Hongchen Jiang, Shuwu Zhang, Bo Xu; Chinese Academy of Sciences, China**

Wed1FoP.1

In this paper, we propose a new method to choose the effective samples for support vector machines (SVM) training based on regression tree in audio classification task. The objective is to reduce the training time of SVM by choosing effective examples from the training set and to balance the number of training points of binary classes. One obvious advantage of such method is that it provides a flexible framework to implement the choice procedure based on the training data for a given classification task. We test the performances of our new method on a dataset composed of about 6-hour audio data which illustrate that the computation time can be significantly reduced without a significant decrease in the prediction accuracy.

**Online Speaker Change Detection by Combining BIC with Microphone Array Beamforming**

**Joerg Schmalenstroeer, Reinhold Haeb-Umbach; University of Paderborn, Germany**

Wed1FoP.2

In this paper we consider the problem of detecting speaker changes in audio signals recorded by distant microphones. It is shown that the possibility to exploit the spatial separation of speakers more than makes up the degradation in detection accuracy due to the increased source-to-sensor distance compared to close-talking microphones. Speaker direction information is derived from the filter coefficients of an adaptive Filter-and-Sum Beamformer and is combined with BIC analysis. The experimental results reveal significant improvements compared to BIC-only change detection, be it with the distant or close-talking microphone.

**Speech/Non-Speech Discrimination Combining Advanced Feature Extraction and SVM Learning**

**Javier Ramirez, Pablo Yélamos, J.M. Górriz, José C. Segura, L. Garcia; Universidad de Granada, Spain**

Wed1FoP.3

This paper shows an effective speech/non-speech discrimination method for improving the performance of speech processing systems working in noisy environment. The proposed method uses a trained support vector machine (SVM) that defines an optimized non-linear decision rule over different sets of speech features. Two alternative feature extraction processes based on: i) subband SNR estimation after denoising, and ii) long-term SNR estimation were compared. Both methods show the ability of the SVM-based classifier to learn how the signal is masked by the acoustic noise and to define an effective non-linear decision rule. However, it is shown that a feature vector incorporating contextual information yielded better speech/non-speech discrimination even when no denoising is applied. The experimental analysis carried out on the Spanish SpeechDat-Car database shows clear improvements over standard VADs including ITU G.729, ETSI AMR and ETSI AFE for distributed speech recognition (DSR), and other recently reported VADs.

**Cooperation Between Global and Local Methods for the Automatic Segmentation of Speech Synthesis Corpora**

**Safaa Jarifi1, Dominique Pastor1, Olivier Rosec2; 1ENST Bretagne, France; 2France Télécom, France**

Wed1FoP.4

This paper introduces a new approach for the automatic segmentation of corpora dedicated to speech synthesis. The main idea behind this approach is to merge the outputs of three segmentation algorithms. The first one is the standard HMM-based (Hidden Markov Model) approach. The second algorithm uses a phone boundary model, namely a GMM (Gaussian Mixture Model). The third method is based on Brandt’s GLR (Generalized Likelihood Ratio) and aims to detect signal discontinuities in the vicinity of the HMM boundaries. Different combination strategies are considered for each phonetic class. The experiments presented in this paper show that the proposed approach yields better accuracy than existing methods.

**Speaker Independent Voiced-Unvoiced Detection Evaluated in Different Speaking Styles**

**Martin Heckmann1, Marco Moebus2, Frank Joublin1, Christian Goerick1; 1Honda Research Institute Europe GmbH, Germany; 2Darmstadt University of Technology, Germany**

Wed1FoP.5

We propose a new algorithm for voiced/unvoiced classification of speech on a phoneme or sample level. The algorithm is inspired by auditory based approaches and combines two cues. One cue is based on the energy distribution of the signal and the other on the harmonicity. In order to extract the harmonicity of the signal we calculate a histogram of the zero crossings of the filter channels after applying a Gammatone filterbank to the signal. A measure similar to the variance of the zero crossings yields the harmonicity cue. The performance of the algorithm was measured on several minutes of read and spontaneous speech with various speakers. An algorithm proposed by Mustafa et al. [1] served as benchmark. The results show that our algorithm performs significantly better as well on read as on spontaneous speech and seems in particular be better able to cope with different speaking styles.

**Robust Speaker Diarization for Meetings: ICSI RT06s Evaluation System**

**Xavier Anguera, Chuck Wooters, Jose M. Pardo; International Computer Science Institute, USA**

Wed1FoP.6

In this paper we present the ICSI speaker diarization system submitted for the NIST Rich Transcription evaluation (RT06s) [1] conducted on the meetings environment. This is a set of yearly evaluations which in the last two years have included speaker diarization of two kinds of distinct meetings: conference room and lecture room. The system presented focuses on being robust to changes in the meeting conditions by not using any training data. In this paper we introduce four of the main improvements to the system from last years’ submission:
The first is a new training-free speech/non-speech detection algorithm. The second is the introduction of a new algorithm for system initialization. The third is the use of a frame purification algorithm to increase clusters differentiability. The last improvement is the use of inter-channel delays as features, greatly improving performance. We show the diarization error rate (DER) score of this system on all available meeting datasets to date for the multiple distant microphone (MDM) and single distant microphone (SDM) conditions.

A Multipitch Tracker for Monaural Speech Segmentation

André Coy, Jon Barker; University of Sheffield, UK Wed1Fp.7

This paper presents a novel algorithm for forming coherent harmonic fragments from a mixture of speech sources. A multiple pitch detection algorithm is used to produce pitch candidates which are tracked using a pair of parallel HMMs. One novel aspect of the technique is that it systematically models pitch doubling and halving errors, thereby facilitating the identification of smooth pitch segments even in the absence of the fundamental frequency. The system does not face the problem of incorrect source assignment that can occur when sources have similar fundamental frequency or are harmonically related. An evaluation of the technique shows that the algorithm’s emphasis on tracking coherent segments leads to the formation of speech fragments with high coherence, indicating a more reliable segmentation of the harmonic speech regions.

Novel Entropy Based Moving Average Refiners for HMM Landmarks

Rahul Chitturi1, Mark Hasegawa Johnson2; 1University of Texas at Dallas, USA; 2University of Illinois at Urbana Champaign, USA Wed1Fp.8

The training of precise speech recognition models depends on accurate segmentation of the phonemes in a training corpus. Segmentation is typically performed using HMMs, but recent speech recognition work suggests that the transient acoustic features characteristic of manner-class phoneme boundaries (landmarks) may be more precisely localized using acoustic classifiers specifically designed for the task of landmark detection. This paper makes an empirical exploration of entropy based moving average techniques that are capable of improving the time alignment of phoneme boundaries proposed by an HMM-based speech recognizer. On a standard benchmark data set (A database of Hindi — National Language of India), we achieve new state-of-the-art performance, reducing RMS phone boundary alignment error from 28ms to 15ms.

Two-Microphone Voice Activity Detection in the Presence of Coherent Interference

Gibak Kim, Nam Ik Cho; Seoul National University, Korea Wed1Fp.9

In this paper, we propose a two-microphone Voice Activity Detection (VAD) method in the presence of coherent interference. The proposed method is based on the Cross Power Spectrum Phase (CPSP) which is an implementation of the Phase Transform (PHAT) weighted cross correlation between two microphones. The PHAT weighting whitens the spectrum of input signals and makes the cross correlation dependent entirely on the phase of the cross spectrum. If we assume that the direction of desired speech signal is known and the time delay between microphones is compensated, the Averaged CPSP (A-CPS) can be utilized as a VAD measure. In order to enhance the VAD performance in the presence of strong coherent interference from other direction, we propose a Maximum Partially Averaged Real CPSP (MPA-RCPSP) method which detects the cophased frequency region with high Signal-to-Interference Ratio (SIR). Simulation results demonstrate that the proposed MPA-RCPSP is a more reliable measure to the conventional A-CPS in the presence of strong coherent interference.

On a Greedy Learning Algorithm for dPLRM with Applications to Phonetic Feature Detection

Tor André Myrvoll1, Tomoko Matsui2; 1NTNU, Norway; 2Institute of Statistical Mathematics, Japan Wed1Fp.10

In this work we investigate the use of a greedy training algorithm for the dual Penalized Logistic Regression Machine (dPLRM), and our target application is detection of broad class phonetic features. The use of a greedy training algorithm is meant to alleviate the infeasible memory and computational demands that arises during the learning phase when the amount of training data increases. We show that using only a subset of the training data, chosen in a greedy manner, we can achieve as good as or better performance as when using the full training set. We can also train dPLRMs using data sets that are significantly larger than what our current computational resources can accommodate when using non-greedy approaches.
**An Automatic Singing Skill Evaluation Method for Unknown Melodies Using Pitch Interval Accuracy and Vibrato Features**

Tomoyasu Nakano¹, Masataka Goto², Yuzuru Hiraga¹; ¹University of Tsukuba, Japan; ²AIST, Japan

**Pitch Determination Using Aligned AMDF**

M. Shahidur Rahman, Hirobumi Tanaka, Tetsuya Shimamura; Saitama University, Japan

---

**Improving Glottal Waveform Estimation Through Rank-Based Glottal Quality Assessment**

Elliot Moore II, Juan Torres; Georgia Institute of Technology, USA

**A Pitch Marks Filtering Algorithm Based on Restricted Dynamic Programming**

Francesc Alías, Carlos Monzo, Joan Claudi Socoró; Ramon Llull University, Spain

**Analysis of Nonmodal Phonation Using Minimum Entropy Deconvolution**

Nicolas Malyska, Thomas F. Quatieri; Massachusetts Institute of Technology, USA

---

**Pitch Estimation**

Oral Session, Allegheny Ballroom, left, 14:00 – 16:00  
Chair: Shigeki Sagayama, University of Tokyo, Japan

**A Spectral-Temporal Method for Pitch Tracking**

Stephen A. Zahorian, Princy Dikshit, Hongbing Hu; Old Dominion University, USA

**Pitch Determination Using Aligned AMDF**

M. Shahidur Rahman, Hirobumi Tanaka, Tetsuya Shimamura; Saitama University, Japan

---

Nonmodal phonation occurs when glottal pulses exhibit non-uniform pulse-to-pulse characteristics such as irregular spacings, amplitudes, and/or shapes. The analysis of regions of such nonmodality has application to automatic speech, speaker, language, and dialect recognition. In this paper, we examine the usefulness of a technique called minimum-entropy deconvolution, or MED [1], for the analysis of pulse events in nonmodal speech. Our study presents evidence for both natural and synthetic speech that MED decomposes nonmodal phonation into a series of sharp pulses and a set of mixed-phase impulse responses. We show that the estimated impulse responses are quantitatively similar to those in our synthesis model. A hybrid method incorporating aspects of both MED and linear prediction is also introduced. We show preliminary evidence that the hybrid method has benefit over MED alone for composite impulse-response estimation by being more robust to short-time windowing effects as well as a speech aspiration noise component.
**Wed2A2O : Acoustic Modeling V — Novel Approaches**

**Oral Session, Allegheny Ballroom, center, 14:00 – 16:00**

**Chair:** Nelson Morgan, International Computer Science Institute, USA

---

**Syllable-Length Path Mixture Hidden Markov Models with Trajectory Clustering for Continuous Speech Recognition**

Yan Han, Lou Boves; Radboud Universiteit Nijmegen, The Netherlands

**Wed2A2O.1**

Recent research suggests that modeling coarticulation in speech is more appropriate at the syllable level. However, due to a number of additional factors that can affect the way syllables are articulated, creating multiple acoustic models per syllable might be necessary. Our previous research on longer-length multi-path models has proved that data-driven trajectory clustering to be an attractive approach to derive multi-path models. However, the use of single distribution with unvarying covariance to model a trajectory cluster may degrade its capability of detecting pronunciation variants. In this paper, we propose a new method, namely path mixture hidden Markov model, to alleviate the adverse effects of trajectory clustering. The improvement on performance observed in continuous speech recognition experiments show mixture model is a very effective approach.

**Acoustic Modeling for Spoken Dialogue Systems Based on Unsupervised Utterance-Based Selective Training**

Tobias Cincarek, Tomoki Toda, Hiroshi Saruwatari, Kiyohiro Shikano; Nara Institute of Science & Technology, Japan

**Wed2A2O.2**

The construction of high-performance acoustic models for certain speech recognition tasks is very costly and time-consuming, since it most often requires the collection and transcription of large amounts of task-specific speech data. In this paper acoustic modeling for spoken dialogue systems based on unsupervised selective training is examined. The main idea is to select those training utterances from an (untranscribed) speech data pool, so that the likelihood of a separate small (transcribed) development speech data set is maximized. If only the selected data are employed to retrain the initial acoustic models, a better performance is achieved than when retraining with all collected data. Using the proposed approach it is also possible to considerably reduce the costs for human-labeling of the speech data without compromising the performance. Furthermore, the method provides means for automatic task-adaptation of acoustic models, e.g. to adult or children speech. This is important, since detailed information about each automatically collected utterance is usually not available.

**GMM-Based Acoustic Modeling for Embedded Speech Recognition**

Christophe Lévy, Georges Linarès, Jean-François Bonastre; LIA, France

**Wed2A2O.3**

Speech recognition applications are known to require a significant amount of resources (training data, memory, computing power). However, the targeted context of this work — mobile phone embedded speech recognition system — only authorizes few KB of memory, few MIPS and usually small amount of training data. In order to fit the resource constraints, an approach based on a semi-continuous HMM system using a GMM-based state-independent acoustic modeling is proposed in this paper. A transformation is computed and applied to the global GMM in order to obtain each of the HMM state-dependent probability density functions. This strategy aims at storing only the transformation function parameters for each state and authorizes to decrease the amount of computing power needed for the likelihood computation. The proposed approach is evaluated on two tasks: a digit recognition task using the French corpus BDSON (which allows a Digit Error Rate of 2.5%) and a voice command task using French corpus VODIS (the Command Error Rate leads around 4.1%).

---

**Boosting HMM Performance with a Memory Upgrade**

Mathias De Wachter, Kris Demuynck, Dirk Van Compernolle; Katholieke Universiteit Leuven, Belgium

**Wed2A2O.4**

The state-of-the-art in automatic speech recognition is distinctly Markovian. The ubiquitous ‘beads-on-a-string’ approach, where sentences are explained as a sequence of words, words as a sequence of phones and phones as a sequence of acoustically stable states, is bound to lose a lot of dynamic information. In this paper we show that a combination with example-based recognition can be used to re-capture some of that information. A new approach to combine Hidden Markov Model (HMM) and phone-example-based continuous speech recognition is presented. Experiments show that the combination outperforms the HMM recognizer, and indicate that adding long-span information is especially beneficial.

**An Integrated Approach to Improve Speech Recognition Rate for Non-Native Speakers**

Y. Deng, X. Li, C. Kwan, R. Xu, B. Raj, Richard M. Stern, D. Williamson; 1 Carnegie Mellon University, USA; 2 Intelligent Automation Inc., USA; 3 Wright-Patterson AFB, USA

**Wed2A2O.5**

The current speech interfaces in many military applications may be adequate for native speakers. However, the recognition rate drops quite a lot for non-native speakers (people with foreign accents). This is mainly because the non-native speakers have large temporal and intra-phoneme variations when they pronounce the same words. This problem is also complicated by the presence of loud environmental noise such as tank noise, helicopter noise, etc. In this paper, we proposed a novel speech feature adaptation algorithm for continuous accent and environmental adaptation. This feature-based adaptation method is then integrated with conventional model-based maximum likelihood linear regression (MLLR) algorithm. Extensive experiments have been performed on the NATO non-native speech corpus with baseline acoustic model trained on native American English. The proposed feature-based adaptation algorithm improved the average recognition accuracy by 15%, while the MLLR model-based adaptation achieved 11% improvement. The combined adaptation achieved overall recognition accuracy improvement of 29.3%, and word error rate reduction of 31.8%.

**Bayesian Decision Tree State Tying for Conversational Speech Recognition**

Rusheng Hu, Yunxin Zhao; University of Missouri-Columbia, USA

**Wed2A2O.6**

This paper presents a new method of constructing phonetic decision trees (PDTs) for acoustic model state tying based on implicitly induced prior knowledge. Our hypothesis is that knowledge on pronunciation variation in spontaneous, conversational speech contained in a relatively large corpus can be used for building domain-specific or speaker-dependent PDTs. In the view of tree structure adaptation, this method leads to transformation of tree topology in contrast to keeping fixed tree structure as in traditional methods of speaker adaptation. A Bayesian learning framework is proposed to incorporate prior knowledge on decision rules in a greedy search of new decision trees, where the prior is generated by a decision tree growing process on a large seed set. Experimental results on the Telemedicine automatic captioning task demonstrate that the proposed approach results in consistent improvement in model quality and recognition accuracy.
Wed2A3O : Corpus-Based Synthesis
Oral Session, Allegheny Ballroom, right, 14:00 - 16:00
Chair: Antonio Bonafonte, Universitat Politècnica de Catalunya, Spain

Feature Extraction for Spectral Continuity Measures in Concatenative Speech Synthesis
Barry Kirkpatrick, Darragh O’Brien, Ronán Scaife; Dublin City University, Ireland
Wed2A3O.1
The quality of concatenative speech synthesis depends on the cost function employed for unit selection. Effective cost functions for spectral continuity are difficult to define and standard measures often do not accurately reflect human perception of discontinuity across a concatenated join. In this study the performance of a number of standard masks are compared for the task of detecting audible discontinuities in concatenated speech. Feature sets derived from the phase spectrum are also investigated. Feature extraction based on wavelet analysis is proposed to overcome some of the limitations of the standard masks tested. Receiver Operating Characteristic (ROC) curves are constructed for each measure from the results of a perceptual experiment and are used to rank the performance of each measure. Results indicate that phase spectra is comparable to magnitude spectra as a joint cost for spectral continuity. Measures based on wavelet transform coefficients outperform all other measures tested.

Decision Tree-Based Training of Probabilistic Concatenation Models for Corpus-Based Speech Synthesis
Shinsuke Sakai, Tatsuya Kawahara; Kyoto University, Japan
Wed2A3O.2
The measure of the goodness, or cost, of concatenating synthesis units plays an important role in concatenative speech synthesis. In this paper, we present a probabilistic approach to concatenation modeling in which the goodness of concatenation is represented as the conditional probability of observing the spectral shape of a unit given the previous unit and the current phonetic context. This conditional probability is modeled by a conditional Gaussian density whose mean vector has a form of linear transform of the past spectral shape. A phonetic decision-tree based parameter tying is performed to achieve a robust training that balances between model complexity and the amount of training data available. The concatenation models are implemented in a corpus-based speech synthesizer trained with a CMU Arctic database and the effectiveness of the proposed method was confirmed by a subjective listening test.

Constructions Stylistic Synthesis Databases from Audio Books
Yong Zhao1, Di Peng2, Lijuan Wang3, Min Chu1, Yining Chen1, Peng Yu1, Jun Guo2; Microsoft Research Asia, China; 2 Beijing University of Posts & Telecommunications, China; 3 Tsinghua University, China
Wed2A3O.3
In this paper, we explore how to construct stylistic TTS databases from audio books, in which a storyteller performs multiple roles. The goal is to identify and build a set of speech corpora, each of which not only portrays a representative voice style performed by the speaker, but also has sufficient sentences to synthesize natural speech using unit selection approach. We solve the problem in two procedures: first, by representing each role with Gaussian Mixture Models (GMM), all speech data are partitioned into a number of voice style clusters with a criterion that maximizes the likelihood of all utterances with respect to roles’ speaker models; then, pruning in terms of both acoustic and prosodic measures is followed to purify the clusters. The resulting 4 voice styles are subjectively interpreted as Neutral, Young, Elder and Adult, respectively. Perceptual experiments show that the proposed approach can synthesize speech with the recognizable voice styles with an average 72.5% identification rate, and the synthesized speech sounds better than those synthesized with utterances from a single role.

Expanding Phonetic Coverage in Unit Selection Synthesis Through Unit Substitution from a Donor Voice
Alistair Conkie, Ann K. Syrdal; AT&T Labs Research, USA
Wed2A3O.4
This paper describes experiments with synthetic voices using unit selection [1] concatenative synthesis where portions of the database audio recordings are modified for the purpose of producing a wider set of phonemes than is contained in the original voice recordings. Since it is known that performing global signal modification for the purposes of speech synthesis significantly reduces perceived voice quality [2] [3], the modifications that we perform are specifically confined to periodic portions of the signal that tend neither to cause concatenation discontinuities nor to convey much of the individual character or affect of the speaker.
We propose three methods to extend the phonetic coverage of unit selection voices (1) by modifying parts of a voice so that extra phones extracted from a donor voice can be added off line; (2) by extending the above methodology by using a harmonic plus noise model (HNM) [4] for speech representation in order to control aspects of the modification; (3) by combining recorded inventories from two voices so that at synthesis time selections can be made from either.
Experiments were conducted to evaluate the strengths and weaknesses of the three methods.

Unifying Unit Selection and Hidden Markov Model Speech Synthesis
Paul Taylor; University of Cambridge, UK
Wed2A3O.5
This paper presents a framework which can accommodate the two most widely used contemporary speech synthesis techniques, namely unit selection and hidden Markov models (HMMs). This is achieved by building a very general HMM where we have a network of states, each representing a single frame for a single unit. This network exactly mimics the behaviour of a unit selection system and is effectively memorising the data as an HMM. From this, we can merge states in the network so as to produce a synthesis system of any desired size. The paper discusses this technique as well as a statistical formulation of the join cost and a number of ways to represent the acoustic observations of the states.

CLUSTERGEN: A Statistical Parametric Synthesizer Using Trajectory Modeling
Alan W. Black; Carnegie Mellon University, USA
Wed2A3O.6
Unit selection synthesis has shown itself to be capable of producing high quality natural sounding synthetic speech when constructed from large databases of well-recorded, well-labeled speech. However, the cost in time and expertise of building such voices is still too expensive and specialized to be able to build individual voices for everyone. The quality in unit selection synthesis is directly related to the quality and size of the database used. As we require our speech synthesizers to have more variation, style and emotion, for unit selection synthesis, much larger databases will be required. As an alternative, more recently we have started looking for parametric models for speech synthesis, that are still trained from databases of natural speech but are more robust to errors and allow for better modeling of variation.
This paper presents the CLUSTERGEN synthesizer which is imple-
The tool thus supports the whole development chain starting from types can be developed with the help of the user modelling plug-in. We further explain how different dialog strategies for different user usability evaluations can be conducted at very early design stages. So that it contains enough information to be fully simulated without the usage of state machines gives us the possibility to fully specify the HMI which is in contrast to the traditional way of using flow-charts. The implementation of speech dialogs. These are specified using state machines, ensuring highly consistent interfaces. In this paper we focus on the development of multimodal interfaces. The HMI model captures all involved modalities, thus ensuring the constructs of Joint Intention Theory (JIT) along with first principles reasoning over a formal semantics of communicative acts [2, 3]. We try to show that STAPLE can subsume and extend the finite-state and frame-based dialogue approaches available today from commercial dialogue systems. In particular, we show how STAPLE can handle over-answering, dynamic environment changes and teamwork in a general domain independent manner.

Promt Selection with Reinforcement Learning in an AT&T Call Routing Application

Charles Lewis, Giuseppe Di Fabbrizio; AT&T Labs Research, USA

Reinforcement Learning (RL) algorithms provide a type of unsupervised learning that is especially well suited for the challenge of spoken dialogue systems (SDS) design. SDS are constantly subjected to new environments in the form of new groups of users, and RL provides an approach for automated learning that can adapt to new environments without costly supervision. In this paper, we describe some results from experiments with RL to select prompts for a call routing application. A simulation of the dialogue outcomes were used to experiment with different scenarios and demonstrate how RL can make a system more robust without supervision or developer intervention.

Developing Speech Dialogs for Multimodal HMIs Using Finite State Machines

Silke Goronzy, Raquel Mochales, Nicole Beringer; 3SOFT GmbH, Germany

We present a tool for model-based development of multimodal interfaces. The HMI model captures all involved modalities, thus ensuring highly consistent interfaces. In this paper we focus on the development of speech dialogs. These are specified using state machines, which is in contrast to the traditional way of using flow-charts. The usage of state machines gives us the possibility to fully specify the HMI so that it contains enough information to be fully simulated without the need to connect any target applications as well as for automatic target code generation. Due to the extensive simulation capabilities usability evaluations can be conducted at very early design stages. We further explain how different dialog strategies for different user types can be developed with the help of the user modelling plug-in. The tool thus supports the whole development chain starting from design studies to specification, development and testing over usability studies and target implementation.

Development of Advanced Dialog Systems with PATE

Norbert Pfleger, Jan Schehl; DFKI GmbH, Germany

Current commercial dialog systems show only limited capabilities with regard to the phenomena occurring in spontaneous, natural dialog. Many research prototypes, in contrast, are already able to deal with a great number of phenomena but lack the clarity and maintainability of commercial systems. In this paper we present a framework for developing advanced multimodal dialog systems designed to bridge this gap.

A Joint Intention-Based Dialogue Engine

Rajah Annamalai Subramanian, Philip Cohen; Natural Interaction Systems, USA

In this paper we discuss some of the advantages of using a joint intention-based interpreter for building spoken dialogue systems. We describe STAPLE [1], a joint-intention interpreter that enables a system to obtain team and communicative behavior automatically without having to program this behavior explicitly. With this approach there is no necessity for a programmer to indicate when a question should be posed, when information should be shared etc. The interpreter enables the agents to exhibit team-oriented dialogue by interpreting the constructs of Joint Intention Theory (JIT) along with first principles reasoning over a formal semantics of communicative acts [2, 3]. We try to show that STAPLE can subsume and extend the finite-state and frame-based dialogue approaches available today from commercial dialogue systems. In particular, we show how STAPLE can handle over-answering, dynamic environment changes and teamwork in a general domain independent manner.

MeMo: Towards Automatic Usability Evaluation of Spoken Dialogue Services by User Error Simulations

Sebastian Möller, Roman Englert, Klaus Engelbrecht, Verena Hafner, Anthony Jameson, Antti Oulasvirta, Alexander Raake, Norbert Reithinger; 1 Technische Universität Berlin, Germany; 2 DFKI GmbH, Germany

Proper usability evaluations of spoken dialogue systems are costly and cumbersome to carry out. In this paper, we present a new approach for facilitating usability evaluations which is based on user error simulations. The idea is to replace real users with simulations derived from empirical observations of users’ erroneous behavior. The simulated errors must cover both system-driven errors (e.g., due to poor speech recognition) as well as conceptual errors and slips of the user, because neither alone is predictive of perceived usability. The simulation is integrated into a workbench which produces reports of typical and rare errors, and which allows usability ratings to be predicted. If successful, this workbench will help designers in making choices between system versions and lower testing costs at early phases of development. Challenges to the approach are discussed and solutions proposed.
Wed2BuP : Modeling Speaker Emotional State
Poster Session, Butler Room, 14:00 – 16:00
Chair: Christer Gobl, Trinity College Dublin, Ireland

Synthesizing Breathiness in Natural Speech with Sinusoidal Modelling
Brett Matthews, Raimo Bakis, Ellen Eide; IBM T.J. Watson Research Center, USA
Wed2BuP.1
This paper discusses recent work in synthesizing a breathy quality in pre-recorded speech, which has applications in voice morphing and concatenative TTS. Previous work has shown that the breathy quality in speech is characterized in part by the presence of random noise in the upper region of the spectrum [1]. The sinusoidal modelling representation of speech facilitates making high-quality modifications to speech signals as well as modifying regions of the spectrum independently. We use sinusoidal modelling, along with techniques borrowed from analog communication systems to simulate aspiration noise in wideband speech signals above some lower cutoff frequency. Specifically, we use techniques based on amplitude modulation (AM) and phase modulation (PM), with the harmonics from the sinusoidal model of speech as carriers and lowpass random noise as the message signal. Formal listening tests were conducted and listeners rated the synthesized effect as “breathy” more often than in natural non-breathy speech, but significantly less often than in naturally breathy speech.

Voice GMM Modelling for FESTIVAL/MBROLA Emotive TTS Synthesis
Mauro Nicolao, Carlo Drioli, Piero Costi; ISTC-CNR, Italy
Wed2BuP.2
Voice quality is recognized to play an important role for the rendering of emotions in verbal communication. In this paper we explore the effectiveness of a processing framework for voice transformations finalized to the analysis and synthesis of emotive speech. We use a GMM-based model to compute the differences between an MBROLA voice and an anger voice, and we address the modification of the MBROLA voice spectra by using a set of spectral conversion functions trained on the data.

We propose to organize the speech data for the training in such way that the target emotive speech data and the diphone database used for the text-to-speech synthesis, both come from the same speaker. A copy-synthesis procedure is used to produce synthesis speech utterances where pitch patterns, phoneme duration, and principal speaker characteristics are the same as in the target emotive utterances. This results in a better isolation of the voice quality differences due to the emotive arousal.

Three different models to represent voice quality differences are applied and compared. The models are all based on a GMM representation of the acoustic space. The performance of these models is discussed and the experimental results and assessment are presented.

EmoVoice: A System to Generate Emotions in Speech
João P. Cabral, Luís C. Oliveira; L²F INESC-ID/IST, Portugal
Wed2BuP.3
Generating emotions in speech is currently a hot topic of research given the requirement of modern human-machine interaction systems to produce expressive speech.

We present the EmoVoice system, which implements acoustic rules to simulate seven basic emotions in neutral speech. It uses the pitch-synchronous time-scaling (PSTS) of the excitation signal to change the prosody and the most relevant glottal source parameters related to voice quality. The system also transforms other parameters of the vocal source signal to produce the irregular voicing quality. The correlation of the speech parameters with the basic emotions was derived from measurements of the glottal parameters and from results reported by other authors. The evaluation of the system showed that it can generate recognizable emotions but improvements are still necessary to discriminate some pairs of emotions.

Zhiyong Wu1, Shen Zhang2, Lianhong Cai1, Helen M. Meng1,2; Chinese University of Hong Kong, China; 1Tsinghua University, China
Wed2BuP.4
This paper describes our initial work in developing a real-time audio-visual Chinese speech synthesizer with a 3D expressive avatar. The avatar model is parameterized according to the MPEG-4 facial animation standard [1]. This standard offers a compact set of facial animation parameters (FAPs) and feature points (FPs) to enable realization of Chinese visemes and facial expressions (i.e. 27 target facial configurations). The Xface [2] open source toolkit enables us to define the influence zone for each FP and the deformation function that relates them. Hence we can easily animate a large number of coordinates in the 3D model by specifying values for a small set of FAPs and their FPs. FAP values for 27 target facial configurations were estimated from available corpora. We extended the dominance blending approach to effect animations for coarticulated visemes superimposed with expression changes. We selected six sentiment-carrying text messages and synthesized expressive visual speech (for all expressions, in randomized order) with neutral audio speech. A perceptual experiment involving 11 subjects shows that they can identify the facial expression that matches the text message’s sentiment 85% of the time.

Modeling the Acoustic Correlates of Expressive Elements in Text Genres for Expressive Text-to-Speech Synthesis
Hongwu Yang1, Helen M. Meng2, Lianhong Cai1; 1Tsinghua University, China; 2Chinese University of Hong Kong, China
Wed2BuP.5
This paper proposes a novel approach for describing the expressive elements in text genres and modeling their acoustic correlates for expressive text-to-speech synthesis (TTS). We apply the three-dimensional PAD (pleasure-displeasure, arousal-nonarousal and dominance-submissiveness) model in describing expressivity. In particular, we define a set of principles for annotating the P and A values of prosodic words found in texts from the tourist information domain. These text passages may be categorized into the descriptive genre (e.g. describing a beautiful scenic spot), the informative genre (e.g. presenting the opening hours of a museum) and the procedural genre (e.g. offering bus routes to a landmark). We choose the prosodic word as the basic unit for analysis since it bridges textual input with acoustic-visual Chinese speech synthesizer with a 3D expressive avatar. The avatar model is parameterized according to the MPEG-4 facial animation standard [1]. This standard offers a compact set of facial animation parameters (FAPs) and feature points (FPs) to enable realization of 20 Chinese visemes and 7 facial expressions (i.e. 27 target facial configurations). The Xface [2] open source toolkit enables us to define the influence zone for each FP and the deformation function that relates them. Hence we can easily animate a large number of coordinates in the 3D model by specifying values for a small set of FAPs and their FPs. FAP values for 27 target facial configurations were estimated from available corpora. We extended the dominance blending approach to effect animations for coarticulated visemes superimposed with expression changes. We selected six sentiment-carrying text messages and synthesized expressive visual speech (for all expressions, in randomized order) with neutral audio speech. A perceptual experiment involving 11 subjects shows that they can identify the facial expression that matches the text message’s sentiment 85% of the time.

Automatic Emotion Recognition of Speech Signal in Mandarin
Sheng Zhang1, P.C. Ching2, Fanrang Kong1; 1University of Science & Technology of China, China; 2Chinese University of Hong Kong, China
Wed2BuP.6
Traditionally, a simultaneous recognition process using the same feature set of a spoken utterance is used to classify the emotional state of the speaker in addition to its content. However, an analysis on the classification performance for every pair of emotions shows that different features have distinctive classification abilities for different emotions. Therefore, we propose an efficient emotion recognition process called cascade bisection (CB-process), which carries out emotion recognition by means of several bisecting steps and applies different feature sets for every step. This process is based on the features’ different abilities of classifying emotions. Through this, we can fully utilize the information extracted from features and achieve a better recognition performance. Five discrete emotional states, namely, neutral, anger, fear, joy, and sadness are distinguished from the input Mandarin speech. After extracting the acoustic features that contain information on short-time energy (amplitude), signal amplitude, and pitch, we derive the representation feature set for further use in the CB-process, which achieves better emotion recognition as demonstrated seen from the experimental results.

Feature Analysis for Emotion Recognition from Mandarin Speech Considering the Special Characteristics of Chinese Language

Yi-hao Kao, Lin-shan Lee; National Taiwan University, Taiwan

Emotion recognition from speech signals is regarded as a critical step toward intelligent human-machine interface. However, feature parameters useful for this purpose may have to do with the special structures of the language. In this paper we present a detailed analysis of the feature parameters for emotion recognition considering the characteristics of the Chinese language, primarily the monosyllabic structure and the tone behavior. The analysis is based on the feature parameters on three levels: frame-level, syllable-level, and word-level. The results show that the frame-level and syllable-level ones are good indicators, while taking the ensemble features on all three levels can yield a recognition accuracy of 90.0%. We also found that the pitch and power related features are the most important, and the fourth tone in Mandarin serves as the strongest indicator to emotions. All these findings are consistent with the characteristics of Mandarin Chinese.

Timing Levels in Segment-Based Speech Emotion Recognition

Björn Schuller, Gerhard Rigoll; Technische Universität München, Germany

Additional sub-phrase level information is believed to improve accuracy in speech emotion recognition systems. Yet, automatic segmentation is a challenge on its own considering word- or syllable boundaries. Further more clarification is needed which timing level leads to optimal results. In this paper we therefore quantitatively discuss three approaches to segment-level features based on 276 utterances of German language with duration as main predictor. The results indicate parameter evolution by emotion, and appear more consistent with duration as main predictor.

Analyzing Dialogue Data for Real-World Emotional Speech Classification

Ryuichi Nisimura, Souji Omae, Hideki Kawahara, Toshio Irino; Wakayama University, Japan

In order to obtain an understanding of the user's emotion in human-machine dialogues, an analysis of dialogical utterances in the real world was performed. This work comprises three major steps. (1) The actual conditions of 16 basic emotions were evaluated using Japanese child voices, which were collected through the field test of the public spoken dialogue system. (2) Two factors were derived by a factor analysis. The factors were defined as fundamental psychological factors representing "delightful" and "hateful" emotions. (3) The relationships between the factors and the physical acoustic features were investigated to establish a capability to sense a user's mental state for the dialogue system. In the experimental discriminations between the delightful and hateful emotions, a correct rate of 98.8% was achieved in classifying child's utterances by the SVM (Support Vector Machine) with 11 acoustic features.

Evolving Emotional Prosody

Cecilia Ovesdotter Alm, Xavier Llorà; University of Illinois at Urbana-Champaign, USA

Emotion is expressed by prosodic cues, and this study uses the active interactive Generic Algorithm to search a wide space for SAD and ANGRY parameters of intensity, F0, and duration in perceptual resynthesis experiments with users. This method avoids large recorded databases and is flexible for exploring prosodic emotion parameters. Solutions from multiple runs are analyzed graphically and statistically. Average results indicate parameter evolution by emotion, and appear more distinct for SAD. Solutions are quite successfully classified by CART, with duration as main predictor.

Vocal Emotion Recognition with Cochlear Implants

Xin Luo, Qian-Jie Fu, John J. Galvin III; House Ear Institute, USA

Besides conveying linguistic information, spoken language can also transmit important cues regarding the emotion of a talker. These prosodic cues are most strongly coded by changes in amplitude, pitch, speech rate, voice quality and articulation. The present study investigated the ability of cochlear implant (CI) users to recognize vocal emotions, as well as the relative contributions of spectral and temporal cues to vocal emotion recognition. An English sentence database was recorded for the experiment; each test sentence was produced according to five target emotions. Vocal emotion recognition was tested in 6 CI and 6 normal-hearing (NH) subjects. With unprocessed speech, NH listeners’ mean vocal emotion recognition performance was 90% correct, while CI users’ mean performance was only 45% correct. Vocal emotion recognition was also measured in NH subjects while listening to acoustic, sine-wave vocoder CI simulations. To test the contribution of spectral cues to vocal emotion recognition, 1, 2, 4, 8, and 16-channel CI processors were simulated; to test the contribution of temporal cues, the temporal envelope filter cutoff frequency in each channel was either 50 or 500 Hz. Results showed that both spectral and temporal cues significantly contributed to performance. With the 50-Hz envelope filter, performance generally improved as the number of spectral channels was increased. With the 500-Hz envelope filter, performance significantly improved only when the spectral resolution was increased from 1 to 2, and then from 2 to 16 channels. For all but the 16-channel simulations, increasing the envelope filter cutoff frequency from 50 Hz to 500 Hz significantly improved performance. CI users’ vocal emotion recognition performance was statistically similar to that of NH subjects listening to 1 - 8 spectral channels with the 50-Hz envelope filter, and to 1 channel with the 500-Hz envelope filter. The results suggest that, while spectral cues may contribute more strongly to recognition of linguistic information, temporal cues may contribute more strongly to recognition of emotional content coded in spoken language.
Emotion Detection in Infants' Cries Based on a Maximum Likelihood Approach
S. Matsunaga, S. Sakaguchi, M. Yamashita, S. Miyahara, S. Nishitani, K. Shinohara; Nagasaki University, Japan
Wed2BuP.12

This paper proposes a new procedure based on a maximum likelihood approach using hidden Markov models to detect infants' emotions through their cries. Our procedure uses stochastic acoustic models for each kind of emotion. The acoustic models are generated using infant's cries that are labeled segmentally according to their acoustic features. The procedure detects segment sequences with the highest likelihood among all kinds of emotions. The results of our preliminary recognition experiments on two emotions using three types of segment labeling show that the proposed procedure is applicable to emotion detection in an infant's cry and that the detailed transcription of acoustic segments is useful. In this paper, using detailed transcriptions, we broaden the experiments to include five emotions. Assuming the judgment of each infant's mother to be correct, we compared the result of the experiment and that of a subjective opinion test. We conducted the opinion test with the help of three child-rearing experts. Emotion recognition using the proposed procedure displays a favorable comparison with the judgment of our experts, showing the validity of the proposed procedure.

“Yeah Right”: Sarcasm Recognition for Spoken Dialogue Systems
Joseph Tepperman, David Traum, Shrikanth Narayanan; University of Southern California, USA
Wed2BuP.13

The robust understanding of sarcasm in a spoken dialogue system requires a reformulation of the dialogue manager's basic assumptions behind, for example, user behavior and grounding strategies. But automatically detecting a sarcastic tone of voice is not a simple matter. This paper presents some experiments toward sarcasm recognition using prosodic, spectral, and contextual cues. Our results demonstrate that spectral and contextual features can be used to detect sarcasm as well as a human annotator would, and confirm a long-held claim in the field of psychology — that prosody alone is not sufficient to discern whether a speaker is being sarcastic.

Identification of Confusion and Surprise in Spoken Dialog Using Prosodic Features
Rohit Kumar, Carolyn P. Rosé, Diane J. Litman; Carnegie Mellon University, USA; University of Pittsburgh, USA
Wed2BuP.14

Sensitivity to a user’s emotional state offers promise in improving the state of the art in spoken dialog systems. In this work, we attempt to detect the speaker’s states of confusion and surprise using prosodic features from his/her utterances. We have collected a corpus of utterances in realistic settings using an experimental methodology aimed at eliciting confusion and surprise from users. Classification experiments have yielded up to a 27.2% improvement over baseline performance using F0 and power features. We achieved the greatest success at classification of emotions that were most successfully elicited.

Analysis and Detection of Speech Under Sleep Deprivation
Tin Lay Nwe, Haizhou Li, Minghui Dong; Institute for Infocomm Research, Singapore
Wed2BuP.15

Stress has effect on speech characteristics and can influence the quality of speech. In this paper, we study the effect of Sleep-Deprivation (SD) on speech characteristics and classify Normal Speech (NS) and Sleep Deprived Speech (SDS). One of the indicators of sleep deprivation is ‘flattened voice’. We examine pitch and harmonic locations to analyse flatness of voice. To investigate, we compute the spectral coefficients that can capture the variations of pitch and harmonic patterns. These are derived using Two-Layer Cascaded-Subband Filter spread according to the pitch and harmonic frequency scale. Hidden Markov Model (HMM) is employed for statistical modeling. We use DCIEM map task corpus to conduct experiments. The analysis results show that SDS has less variation of pitch and harmonic pattern than NS. In addition, we achieve the relatively high accuracy for classification of Normal Speech (NS) and Sleep Deprived Speech (SDS) using proposed spectral coefficients.

Language, Gender, Speaking Style and Language Proficiency as Factors Influencing the Autonomous Vocalic Filler Production in Spontaneous Speech
Ioana Văsilescu, Martine Adda-Decker; LIMSI, France
Wed2BuP.16

This paper deals with the factors characterizing the production of autonomous vocalic filled pauses in large spontaneous speech corpora, namely language, gender, speaking style and language proficiency. Two types of corpora are analyzed: a corpus of broadcast news in French and American English and a corpus of short talks in a conference in English spoken by native and non-native speakers. Several acoustic and prosodic parameters are evaluated and correlated with each factor, namely timbre, pitch, duration and density. Results presented here show that the timbre is correlated with language and language proficiency, whereas the duration is linked both to gender and speaking style, the latter conditioning also the hesitation density in speech.

How to Handle Gender and Number Agreement in Statistical Language Models?
Caroline Lavecchia, Kamel Smaïli, Jean-Paul Haton; LORIA, France
Wed2CaP.1

The agreement in gender and number is a critical problem in statistical language modeling. One of the main difficulties in speech recognition of French language is the presence of misrecognized words due to the bad agreement (in gender and number) between words. Statistical language models do not treat this phenomena directly. This paper focuses on how to handle the issue of this agreement. We introduce an original model called Features-Cache (FC) to estimate the gender and the number of the word to predict. It is a dynamic variable-length Features-Cache. The size of the cache is automatically determined in accordance to syntagm delimiters. The main advantage of this model is that there is no need to any syntactic parsing: it is used as any other statistical language model. Several models have been carried out and the best one achieves an improvement of approximately 9 points in terms of perplexity. This model has been integrated in a speech recognition system based on JULIUS engine. Tests have been carried out on 280 sentences provided by AUPELF for the French automatic speech recognition evaluation campaign. This new model outperforms the baseline one, in terms of word error, by 3%.

Prosodic Features for a Maximum Entropy Language Model
Oscar Chan, Roberto Togneri; University of Western Australia, Australia
Wed2CaP.2
This paper presents an approach for incorporating prosodic knowledge into the language modelling component of a speech recogniser. We formulate features for a maximum entropy language model which capture various aspects of the relationships between prosody, syntax and the spoken word sequence. Maximum entropy is a powerful modelling technique, and well suited to modelling prosodic information. Tests conducted on the Boston University Radio Speech Corpus using this model showed improvements in perplexity, and n-best rescoring results also demonstrated small but statistically significant gains.

**Language Model Adaptation with a Word List and a Raw Corpus**

*Shinsuke Mori; IBM Japan Ltd., Japan*

*Wed2CaP.3*

In this paper, we discuss language model adaptation methods given a word list and a raw corpus. In this situation, the general method is to segment the raw corpus automatically using a word list, correct the output sentences by hand, and build a model from the segmented corpus. In this sentence-by-sentence error correction method, however, the annotation of handcrafted inference routines is lacking. We propose using dynamic Bayesian networks. Dynamic Bayesian networks are a generalization of the n-gram models and HMMs traditionally used in language modeling and speech recognition. Whereas those models use a single random variable to represent state, this results in a decrease of productivity. In this paper, we propose to concentrate on correcting the positions in which the words in the list appear by taking a word as a correction unit. This method allows us to avoid these problems and go directly to capturing the statistical behavior of specific words in the application. In the experiments, we used a variety of methods for preparing a segmented corpus and compared the language models by their speech recognition accuracies. The results showed the advantages of our method.

**Topic-Based Language Modeling with Dynamic Bayesian Networks**

*Pascal Wiggers, Léon J.M. Rothkrantz; Delft University of Technology, The Netherlands*

*Wed2CaP.4*

Although n-gram models are still the de facto standard in language modeling for speech recognition, more sophisticated models achieve better accuracy by taking additional information, such as syntactic rules, semantic relations or domain knowledge into account. Unfortunately, most of the effort in developing such models goes into the implementation of handcrafted inference routines. A generic mechanism to introduce background knowledge into a language model is lacking. We propose using dynamic Bayesian networks. Dynamic Bayesian networks are a generalization of the n-gram models and HMMs traditionally used in language modeling and speech recognition. Whereas those models use a single random variable to represent state, Bayesian networks can have any number of variables. As such they are particularly well-suited for the construction of models that take additional information into account. This paper discusses language modeling with Bayesian networks. Examples of Bayesian network implementations of well-known language models are given and a novel topic-based language model is presented.

**Speech Recognition of Foreign Out-of-Vocabulary Words Using a Hierarchical Language Model**

*Hirofumi Yamamoto1, Genichiro Kikui2, Satoshi Nakamura1, Yoshinori Sagiisaka1, 2; NICT, Japan; 2; ATR-SLC, Japan*

*Wed2CaP.5*

This paper proposes a new speech recognition scheme for foreign out-of-vocabulary words embedded in native-language speech. To recognize foreign names frequently observed in news speech or in translation speech, we adopted a hierarchical language model that had been successfully applied to OOV words covering native vocabularies. In this hierarchical language model, OOV vocabularies are modeled as a word-class model in the upper-layered model, and their statistical phonotactic constraints are modeled in the lower-layered model. Since extra statistics are needed to cover foreign words and their pronunciation differences, we have introduced two techniques. The first is to combine translation target language models and translation source statistics of OOVs using the hierarchical language model. The second is to automatically generate recognition target pronunciations from original pronunciations by syllable-to-syllable mapping. To confirm the validity of this recognition scheme, we have conducted speech recognition experiments using English speech including Japanese personal names as OOV words. The proposed method outperformed the existing algorithm using a lexicon consisting of all the words in the training set. Surprisingly, it achieved better OOV recognition results than the non-OOV condition where all the proper names in the test set are registered in the lexicon.

**Language Modeling of Chinese Personal Names Based on Character Units for Continuous Chinese Speech Recognition**

*Xinhui Hu1, Hirofumi Yamamoto1, Genichiro Kikui2, Yoshinori Sagisaka1, 1; NICT, Japan; 2; ATR-SLC, Japan*

*Wed2CaP.6*

In this paper, we analyze Chinese personal names to model their statistical phonotactic characteristics for continuous Chinese speech recognition. The analysis showed language-specific characteristics of Chinese personal names and strongly suggested the advantage of character-unit oriented modeling. A hierarchical language model was composed by reflecting statistical phonotactic characteristics of Chinese personal names as a lower intra-word model, and ordinary inter-word neighboring characteristics as an upper multi-class composite N-gram model. These two layers of models were trained independently using different language corpora. For the modeling of given names, the syllable without tone information was selected as the unit for training the bi-gram. The properties of either one or two characters of a given name were introduced to simplify the length constraint of the modeling process. For Chinese family names, we simply added them directly in the recognition lexicon, since their numbers are very restricted. The results from Chinese speech recognition experiments revealed that the proposed hierarchical language model greatly improved the identification accuracy of the Chinese given names compared with the conventional word-class N-gram model.

**A Syllable Based Continuous Speech Recognizer for Tamil**

*Lakshmi A., Hema A. Murthy; Indian Institute of Technology Madras, India*

*Wed2CaP.7*

This paper presents a novel technique for building a syllable based continuous speech recognizer when unannotated transcribed train data is available. We present two different segmentation algorithms to segment the speech and the corresponding text into comparable syllable like units. A group delay based two level segmentation algorithm is proposed to extract accurate syllable units from the speech data. A rule based text segmentation algorithm is used to automatically annotate the text corresponding to the speech into syllable units. Isolated syllable models are built using multiple frame size (MFS) and multiple frame rate (MFR) for all unique syllables by collecting examples from annotated speech. Experiments performed on Tamil language show that the recognition performance is comparable to recognizers built using manually segmented train data. These experiments suggest that system development cost can be reduced by using minimum manual effort if sentence level transcription of the speech data is available.

**Spontaneous Thai Speech Recognition**

*Monika Waszczyna1, Paisarn Charoenpornsawat2, Tanja Schultz2, 1; Multimodal Technologies Inc., USA; 2; Carnegie Mellon University, USA*

*Wed2CaP.8*
This paper expands previous work on Thai speech recognition, investigating pronunciation changes such as syllable and phoneme elisions as well as phoneme shifts in Thai spontaneous speech. We compare several approaches to model these effects in large vocabulary continuous speech recognition across multiple domains. This work includes experiments on two new speech databases that significantly alleviate the data sparseness problem of earlier publications. We found that given sufficient training data, a fully data driven approach using an allophone cluster tree yields the best results. Explicit modeling of pronunciation changes does not improve performance across domains.

**Automatic Generation of Statistical Language Models for Interactive Voice Response Applications**

Mithun Balakrishna\(^1\), Cyril Cerovic\(^1\), Dan Moldovan\(^1\), Ellis Cave\(^2\); \(^1\)University of Texas at Dallas, USA; \(^2\)Intervoice Inc., USA

This paper proposes a methodology to automatically generate statistical language models (SLMs) for the recognition of utterances in Interactive Voice Response (IVR) systems. The paper aims at creating SLMs for each IVR prompt \([1]\) with minimum amount of human intervention and prior knowledge regarding the expected user utterances at a particular prompt. A combination of prefiller patterns based on spontaneous speech utterances, WordNet \([2]\) and Roget's thesaurus based content word extraction and, world wide web based statistical validation is used to generate SLMs automatically. The created SLM not only reduces the manual labor involved in IVR application development but also focuses on minimizing the Word Error Rate (WER) and the Semantic Error Rate (SemER) of the ASR transcriptions. We use a WordNet \([2]\) lexical chain based semantic categorizer to classify ASR transcriptions into semantic categories representing each IVR prompt.

**Comparison of Keyword Spotting Methods for Searching in Speech**

Luboš Šmídl, Josef V. Psutka; University of West Bohemia, Czech Republic

This paper presents and discusses keyword spotting methods for searching in speech. In contrast with searching in text, the searching in speech or generally in multimedia data still represents a challenge. The aim of the paper is to present a keyword spotting (KWS) method based on a large vocabulary continuous speech recognition (LVCSR) system, based on phonetics decoder, and keyword spotting using a filler model. All the methods are evaluated and compared from various points of view — speed, quality, requirements on training data and so on. All experiments are done using a telephone-quality speech corpus. Furthermore, this paper presents a new block decision in filler model-based keyword spotting which brings the speedup of decision together with better detection.

**Speech and Speech Recognition During Dictation Corrections**

Keith Vertanen; University of Cambridge, UK

A natural way to correct errors made while dictating to a computer is to respeak portions of the original sentence. But often spoken corrections are themselves misrecognized, costing the user time and testing their patience. To better understand how users behave while correcting, I created a simulated dictation interface and fooled users into believing they were correcting errors by respeaking. I found that users not only hyperarticulate during corrections, but they do so preemptively before any misrecognition. Depending on the recognizer, hyperarticulation was found to cause relatively minor changes in error rate. The correction of isolated words or phrases was more troublesome, causing substantial recognition problems for an HTK recognizer. Dragon Naturally Speaking, on the other hand, performed slightly better on hyperarticulated speech and only degraded slightly on isolated corrections.
A Spoken Language Understanding Approach Using Successive Learners
Wei-Lin Wu, Ru-Zhan Lu, Hui Liu, Feng Gao; Shanghai Jiao Tong University, China
Wed2FoP.1
In this paper, we describe a novel spoken language understanding approach using two successive learners. The first learner is used to identify the topic of an input utterance. With the restriction of the recognized target topic, the second learner is trained to extract the corresponding slot-value pairs. The advantage of the proposed approach is that it is mainly data-driven and requires only minimally annotated corpus for training whilst retaining the understanding robustness and deepness for spoken language. Experiments have been conducted in the context of Chinese public transportation information inquiry domain. The good performance demonstrates the viability of the proposed approach.

Conversational Help Desk: Vague Callers and Context Switch
Osamuyimen Stewart, Juan Huerta, Ea-Ee Jan, Cheng Wu, Xiang Li, David Lubensky; IBM T.J. Watson Research Center, USA
Wed2FoP.2
Two salient properties of user behavior make Help Desk a unique speech application different from the more general transactional kind: (a) majority of users have only vague ideas about their problem, and (b) these users are likely to context-switch (change discourse topic) during the course of a dialog. We describe a conversational Help Desk natural language call routing application and show how the alignment of Voice User Interface (VUI), Grammar Development, and Application Architecture results in a conversational user interface that is able to guide the vague user in the most optimal way while being flexible to allow mid-discourse context switches. Usability evaluation confirms the peculiar user behavior and provides empirical evidence that user’s perception of time in a speech application can be influenced by the dialog; in this case, Help Desk users’ tend not to become impatient going through three-five dialog turns, as long as dialog is progressing toward problem-resolution.

Integrating Spoken Dialog and Question Answering: The Ritel Project
Sophie Rosset, Olivier Galibert, Gabriel Illouz, Aurélien Max; LIMSI, France
Wed2FoP.3
The Ritel project aims at integrating spoken language dialog and open-domain information retrieval to allow a human to ask general questions (e.g. “Who is currently presiding the French Senate?”) and refine her search interactively. This project is at the junction of several distinct research communities, and has therefore several challenges to tackle: real-time streamed speech recognition with very large vocabulary, open-domain dialog management, fast information retrieval from text, query cobuilding, communication between information retrieval and dialog components, and generation of natural sounding answers. In this paper, we present our work on the different components of Ritel, and provide initial results.

Rapid Simulation-Driven Reinforcement Learning of Multimodal Dialog Strategies in Human-Robot Interaction
Thomas Prommer, Hartwig Holzapfel, Alex Waibel; Universität Karlsruhe, Germany
Wed2FoP.4
In this work we propose a procedure model for rapid automatic strategy learning in multimodal dialogs. Our approach is tailored for typical task-oriented human-robot dialog interactions, with no prior knowledge about the expected user and system dynamics being present. For such scenarios, we propose the use of stochastic dialog simulation for strategy learning, where the user and system error models are solely trained through the initial execution of an inexpensive Wizard-of-Oz experiment. We argue that for the addressed dialogs, already a small data corpus combined with a low-conditioned simulation model facilitates learning of strong and complex dialog strategies. To validate our overall approach, we empirically show the supremacy of the learned strategy over a hand-crafted strategy for a concrete human-robot dialog scenario. To the authors’ knowledge, this work is the first to perform strategy learning from multimodal dialog simulation.

Software Architectures for Incremental Understanding of Human Speech
Gregory Aist 1, James Allen 1, Ellen Campana 1, Lucian Galescu 2, Carlos A. Gomez Gallo 1, Scott C. Stone 1, Mary Swift 1, Michael Tanenhaus 1, 1 University of Rochester, USA; 2 IHMC, USA
Wed2FoP.5
The prevalent state of the art in spoken language understanding by spoken dialog systems is both modular and whole-utterance. It is modular in that incoming utterances are processed by independent components that handle different aspects, such as acoustics, syntax, semantics, and intention / goal recognition. It is whole-utterance in that each component completes its work for an entire utterance prior to handing off the utterance to the next component. However, a growing body of evidence suggests that humans do not process language that way. Rather, people process speech by rapidly integrating constraints from multiple sources of knowledge and multiple linguistic levels incrementally, as the utterance unfolds. In this paper we describe ongoing work aimed at developing an architecture that will allow machines to understand spoken language in a similar way. This revolutionary approach is promising for two reasons: 1) it more accurately reflects contemporary models of human language understanding, and 2) it results in empirical improvements including increased parsing performance.

Lingua Machinae — An Unorthodox Proposal
Florian Schiel, Christoph Draxler, Marion Libossek; LMU München, Germany
Wed2FoP.6
Voice User Interfaces (VUI) are slowly emerging into today’s IT applications. A decade ago, members of the speech science community (as well as market analysts) predicted a much faster growth of commercial VUI usage and are now wondering why the acceptance was so moderate during the last five years. One reason for this is certainly the difficult economic situation since the crash of the new technology market. Another reason is the fact that people are slow, especially in the light of an uncertain economy, to leave well-trodden paths for new shores. In this paper we raise the unorthodox question whether VUIs might work more efficiently, be better accepted by users and have a greater commercial impact on future IT markets if there existed a widely agreed communication paradigm on how to use VUIs. For the moment we’d like to call this paradigm ‘Lingua Machinae’.
Evaluation of Content Presentation Strategies for an In-car Spoken Dialogue System

Heather Pon-Barr\textsuperscript{1}, Fulliang Weng\textsuperscript{1}, Sebastian Vargas\textsuperscript{2}; Robert Bosch Corp., USA; Stanford University, USA

Wed2FoP.7

In this paper we present a framework for managing information presentation in spoken dialogue systems. We describe a content optimization module that makes use of ontological relationships in information-seeking dialogues in order to organize knowledge base items and perform adjustments such as relaxing or tightening user constraints. We present the results of an experimental evaluation comparing two response strategies: (a) one that uses the content optimization module to offer suggestions and (b) one that gives no suggestions. The results indicate that giving such suggestions is preferred when a user query matches either no items or many items in the knowledge base, and may also lead to more efficient dialogues.

On Designing Context Sensitive Language Models for Spoken Dialog Systems

Vaibhava Goel, Ramesh Gopinath; IBM T.J. Watson Research Center, USA

Wed2FoP.8

In this paper we describe our approach to building dialog context sensitive language models for improving the recognition performance in spoken dialog systems. These methods were developed and successfully tested in the context of a large-scale commercially deployed system that takes in over ten million calls each month. Dialog sensitive language models are typically built by clustering dialog history into groups that elicit similar responses. A key innovation in this paper is to use an EM clustering procedure in lieu of a k-means clustering procedure that is typically used. The EM procedure results in clusters with higher log-probability which we argue leads to better recognition performance. Additionally, we empirically observe that the EM approach has much better worst case behavior than the k-means approach as it pertains to local optima.

Using SVM and Error-Correcting Codes for Multiclass Dialog Act Classification in Meeting Corpus

Yang Liu; University of Texas at Dallas, USA

Wed2FoP.9

Accurate classification of dialog acts (DAs) is important for many spoken language applications. Different methods have been proposed such as hidden Markov models (HMM), maximum entropy (Maxent), graphical models, and support vector machines (SVMs). In this paper, we investigate using SVMs for multiclass DA classification in the ICSI meeting corpus. We evaluate (1) representing DA tagging directly as a multiclass task, and (2) combining multiple binary classifiers via error correction output codes (ECOC). For the ECOC combination, different code matrices are utilized (e.g., the identity matrix, exhaustive code, BCH code, and random code matrix). We also compare using SVMs with our previous Maxent model. We find that for DA tagging, using multiple binary SVMs via ECOC outperforms a direct multiclass SVM, but neither achieves better performance than the Maxent model, possibly because of the small class set and the features currently used in the task.

A Multilingual Expectations Model for Contextual Utterances in Mixed-Initiative Spoken Dialogue

Hartwig Holzapfel, Alex Waibel; Universität Karlsruhe, Germany

Wed2FoP.10

This paper describes a model of generating expectations that are used to improve speech recognition and to resolve elliptical expressions in dialogue context. The algorithm is domain and language independent and part of the dialogue manager. We use the expectation model to weight a speech recognizer’s grammar rules in dialogue context which improves recognition rates significantly as shown in the evaluation. We explain what types of expectations the system can generate and give a classification of system actions based on speech act theory, explain the resolution of elliptical expressions and their interpretation in context, and evaluate the presented algorithm in a multilingual system with English and German speech recognition.

Dynamic Help Generation by Estimating User’s Mental Model in Spoken Dialogue Systems

Yuichiro Fukubayashi, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno; Kyoto University, Japan

Wed2FoP.11

In a speech interface, a gap between a user’s mental model and actual structures of systems tends to be large because the amount of information conveyed by speech is limited. We address dynamic help generation adapted to users, to decrease the gap between them. We defined a domain concept tree as an expression of a system’s actual structure. We estimated and maintained user’s knowledge about the system on the tree. Every node in the tree has values representing the degree to which a user understands the concepts corresponding to the nodes. The values are updated based on the content of user’s utterances and help messages the system gives. Help messages provided for users are determined by referring to the domain concept tree and identifying concepts the user does not understand. We evaluated our method by testing twelve novice subjects. Both the average time to complete tasks and the number of utterances significantly decreased because of the help messages provided by our method.

Dialog Act Tagging with Support Vector Machines and Hidden Markov Models

Dinoj Surendran, Gina-Anne Levow; University of Chicago, USA

Wed2FoP.12

We use a combination of linear support vector machines and hidden markov models for dialog act tagging in the HCRC MapTask corpus, and obtain better results than those previously reported. Support vector machines allow easy integration of sparse high-dimensional text features and dense low-dimensional acoustic features, and produce posterior probabilities usable by sequence labelling algorithms. The relative contribution of text and acoustic features for each class of dialog act is analyzed.
**Wed3A1O : Segmentation and VAD**

**Noise Robust Model-Based Voice Activity Detection**

Ángel de la Torre, Javier Ramírez, Carmen Benítez, José C. Segura, L. García, Antonio J. Rubio; Universidad de Granada, Spain

We propose a model-based VAD derived from the Vector Taylor Series (VTS) approach. A Gaussian mixture (trained with clean speech) is used in order to provide an appropriate decision rule for speech/non-speech detection. Additionally, VTS approach adapts the Gaussian mixture to noise conditions, yielding a stable performance for a wide range of SNRs. We have evaluated its ability for speech/non-speech detection and also its application for robust speech recognition. When compared to other VAD methods, the proposed VAD shows the best trade-off in speech/non-speech detection. When applied for Wiener Filtering and for frame dropping, the proposed VAD also provides the best recognition results.

**Auto-Segmentation Based VAD for Robust ASR**

Yu Shi, Frank K. Soong, Jian-Lai Zhou; Microsoft Research, Asia, China

An auto-segmentation based endpointing algorithm for robust ASR is proposed. The algorithm consists of two successive steps: (1) homogeneous segment partitioning and (2) segment clustering. The first step, due to its self-segmentation nature, does not need a noise model, and is applicable to different noises at various SNRs. The dynamic programming based segment partitioning, which can generate more homogeneous segments than individual frames for clustering, yields a more robust VAD mechanism. Experiments are performed on the AURORA2 digit database by comparing the new algorithm with the ETSI standard for DSR. Quantitative assessment of the new algorithm is performed via different evaluation criteria, including: ROC curves, speech/non-speech discrimination, and speech recognition performance.

**Improved Speech Activity Detection Using Cross-Channel Features for Recognition of Multiparty Meetings**

Kofi Boakye 1, Andreas Stolcke 2, 1International Computer Science Institute, USA; 2SRI International, USA

We describe the development of a speech activity detection system using an HMM-based segmenter for automatic speech recognition on individual headset microphones in multispeaker meetings. We look at cross-channel features (energy and correlation based) to incorporate into the segmenter for the purpose of addressing errors related to cross-channel phenomena such as crosstalk. Results demonstrate that these features provide a marked improvement (18% relative) over a baseline system using single-channel features as well as an improvement (8% relative) over our previous solution of separate speech activity detection and cross-channel analysis. In addition, the simple cross-channel energy features are shown to be more robust — and consequently better performing — than the more common correlation-based features.

**Evaluation of Voice Activity Detection by Combining Multiple Features with Weight Adaptation**

Yusuke Kida, Tatsuya Kawahara; Kyoto University, Japan

For noise-robust automatic speech recognition (ASR), we propose a novel voice activity detection (VAD) method based on a combination of multiple features. The scheme uses a weighted combination of four conventional VAD features: amplitude level, zero crossing rate, spectral information, and Gaussian mixture model (GMM) likelihood. The weights for combination are adaptively updated using minimum classification error (MCE) training. In this paper, we first investigate the effect of adaptation of the combination weights and GMM parameters, and demonstrate that the weights can be effectively adapted with a single utterance. Then, we present application of the method to ASR. It is confirmed that the proposed method significantly outperforms conventional methods in various noise conditions.

**Voice Activity Detection in Personal Audio Recordings Using Autocorrelogram Compensation**

Keansub Lee, Daniel P.W. Ellis; Columbia University, USA

This paper presents a novel method for identifying regions of speech in the kinds of energetic and highly-variable noise present in ‘personal audio’ collected by body-worn continuous recorders. Motivated by psychoacoustic evidence that pitch is crucial in the perception and organization of sound, we use a noise-robust pitch detection algorithm to locate speech-like regions. To avoid false alarms resulting from background noise with strong periodic components (such as air-conditioning), we add a new channel selection scheme to suppress frequency subbands where the autocorrelation is more stationary than encountered in voiced speech. Quantitative evaluation shows that these harmonic noises are effectively removed by this compensation technique in the domain of auto-correlogram, and that detection performance is significantly better than existing algorithms for detecting the presence of speech in real-world personal audio recordings.

**Discriminating Speech and Non-Speech with Regularized Least Squares**

Ryan Rifkin 1, Nima Mesgarani 2, 1Honda Research Institute USA Inc., USA; 2University of Maryland, USA

We consider the task of discriminating speech and non-speech in noisy environments. Previously, Mesgarani et. al [1] achieved state-of-the-art performance using a cortical representation of sound in conjunction with a feature reduction algorithm and a nonlinear support vector machine classifier. In the present work, we show that we can achieve the same or better accuracy by using a linear regularized least squares classifier directly on the high-dimensional cortical representation; the new system is substantially simpler conceptually and computationally. We select the regularization constant automatically, yielding a parameter-free learning system. Intriguingly, we find that optimal classifiers for noisy data can be trained on clean data using heavy regularization.
Automatic Grammar Correction for Second-Language Learners

John Lee, Stephanie Seneff; Massachusetts Institute of Technology, USA

A computer conversational system can potentially help a foreign-language student improve his/her fluency through practice dialogues. One of its potential roles could be to correct ungrammatical sentences. This paper describes our research on a sentence-level, generation-based approach to grammar correction: first, a word lattice of candidate corrections is generated from an ill-formed input. A traditional n-gram language model is used to produce a small set of N-best candidates, which are then reranked by parsing using a stochastic context-free grammar. We evaluate this approach in a flight domain with simulated ill-formed sentences. We discuss its potential applications in a few related tasks.

ASR-Based Corrective Feedback on Pronunciation: Does It Really Work?

Ambra Neri, Catia Cucchiarini, Helmer Strik; Radboud Universiteit Nijmegen, The Netherlands

We studied a group of immigrants who were following regular, teacher-fronted Dutch classes, and who were assigned to three groups using either a) Dutch CAPT, an ASR-based Computer Assisted Pronunciation Training (CAPT) system that provides feedback on a number of Dutch speech sounds that are problematic for L2 learners b) a CAPT system without feedback c) no CAPT system. Participants were tested before and after the training. The results show that the ASR-based feedback was effective in correcting the errors addressed in the training.

Evaluating Prosody of Mandarin Speech for Language Learning

Minghui Dong, Haizhou Li, Tin Lay Nwe; Institute for Infocomm Research, Singapore

This paper proposes an approach to automatically evaluate the prosody of Chinese Mandarin speech for language learning. In this approach, we grade the appropriateness of prosody of speech units according to a model speech corpus from a teacher’s voice. To this end, we build two models, which are the prosody model and the scoring model. The prosody model that is built from the teacher’s speech predicts the reference prosody for the learning text. The scoring model compares the student’s prosody with the reference prosody and gives a prosody rating score. Both the prosody model and the scoring model are built using regression tree. To make the two prosodies comparable, we transform the student’s prosody into the teacher’s prosody space. To build the scoring model, we derive from the corpus a reference data set, in which prosody rating is associated with prosody parameters. During speech evaluation, the student’s prosody is first transformed into the teacher’s prosody space and then evaluated by the scoring model. Experiments show that our model works well for speech of new speakers.

Spoken Language Technologies Applied to Digital Talking Books

Isabel Trancoso1, Carlos Duarte2, António Serralheiro3, Diamantino Caseiro1, Luis Carriço2, Céu Viana3,1, I2F INESC-ID/IST, Portugal; 2 LaSIGE/FCUL, Portugal; 3 I2F INESC-ID/Academia Militar, Portugal; 4 CLUL, Portugal

Digital Talking Books (DTBs) offer to visually impaired users an evolution of analogue talking books that mimics the interaction possibilities of print books. This paper describes a new DTB player which tries to improve the usability and accessibility of current players, through the combination of the possibilities offered by multimodal interaction and interface adaptability, and the integration of several language processing components. Besides the potential for a greater enjoyment of the reader in general, these modifications also pave the way to the use of DTBs in different domains, from e-inclusion to e-learning applications.

Building an English Speech Synthesis System from a Japanese ALS Patient’s Voice

Akemi Iida1, Jun Ito1, Shimpei Kajima2, Tsutomu Sugawara2,1 Tokyo University of Technology, Japan; 2 Sophia University, Japan

This paper reports on the development of an English speech synthesis system for a Japanese amyotropic lateral sclerosis patient as part of the project of developing a bilingual communication aid for this patient. The patient had a tracheotomy three years ago and anticipates the possibility of losing his phonatory function. His English speech database for Festival, a free speech synthesis system, was generated from his reading of a US diphone list. There were two problems with the recording. The first was the noise that the artificial ventilator made and the second was his difficulty in pronouncing English. Although the speaker’s English database was successfully built by Festvox and the voice was recognized as his voice, the utterance was unintelligible. We therefore proposed reconstructing the patient’s database by partially combining it with an English native speaker’s database. Results showed that the proposed approach can be promising for those facing this problem.

Multi-Modal System ICANDO: Intellectual Computer AssistantNt for Disabled Operators

Alexey Karpov1, Andrey Ronzhin1, Alexandre Cadiou2; 1 Russian Academy of Sciences, Russia; 2 ISEP, France

The paper describes a multi-modal system ICANDO (an Intellectual Computer AssistantNt for Disabled Operators) developed by Speech Informatics Group of SPIIRAS and intended for assistance to the persons without hands or with disabilities of their hands or arms in human-computer interaction. This system combines the modules for automatic speech recognition and head tracking in one multi-modal system. The developed system was applied for hands-free work with Graphical User Interface of MS Windows in such tasks as Internet communication and work with text documents. The architecture of the system, the methods for recognition and tracking, information fusion and synchronization, experimental conditions and the obtained results are described in the paper.
Wed3WeS : The Prosody of Turn-Taking and Dialog Acts

User Responses to Prosodic Variation in Fragmentary Grounding Utterances in Dialog

Gabriel Skantze, David House, Jens Edlund; KTH, Sweden
Wed3WeS.1

In a previous study we demonstrated that subjects could use prosodic features (primarily peak height and alignment) to make different interpretations of synthesized fragmentary grounding utterances. In the present study we test the hypothesis that subjects also change their behavior accordingly in a human-computer dialog setting. We report on an experiment in which subjects participate in a color-naming task in a Wizard-of-Oz controlled human-computer dialog in Swedish. The results show that two annotators were able to categorize the subjects' responses based on pragmatic meaning. Moreover, the subjects' response times differed significantly, depending on the prosodic features of the grounding fragment spoken by the system.

Analysis of Prosodic and Linguistic Cues of Phrase Finals for Turn-Taking and Dialog Acts

Carlos Toshinori Ishi, Hiroshi Ishiguro, Norihiro Hagita; ATR-IRC, Japan
Wed3WeS.2

This paper presents an analysis on the functions carried by phrase final tones in turn-taking and dialog acts, taking into account linguistic information about the part of speech (particles and auxiliary verbs) attributed to the morphemes at phrase finals. Natural conversational speech data are segmented in inter-pause units, and each utterance unit is arranged according to the phrase final morphemes. Turn-taking functions are annotated, and tones of each phrase final are described by acoustic-prosodic features. Analysis results show a relationship between tones and turn-taking functions in most of the morphemes, while no clear relationship is found in some classes of morphemes.

On Speaker-Specific Prosodic Models for Automatic Dialog Act Segmentation of Multi-Party Meetings
Jáčym Kolář¹, Elizabeth Shriberg², Yang Liu³;
¹International Computer Science Institute, USA; ²SRI International, USA; ³University of Texas at Dallas, USA
Wed3WeS.4

We explore speaker-specific prosodic modeling for dialog act segmentation of speech from the ICSI Meeting Corpus. We ask whether features beyond pauses help individual speakers, and whether some features help from prosody models trained on only their speech. We find positive results for both questions, although the second is more complex. Feature analysis reveals that duration is the most used feature type, followed by pause and pitch features. Results also suggest a difference between native and nonnative speakers in feature usage patterns. We conclude that features beyond pauses are useful for dialog act segmentation in natural conversation, and that for some speakers, speaker-specific training yields further gains.

A Case Study in the Identification of Prosodic Cues to Turn-Taking: Back-Channeling in Arabic
Nigel G. Ward, Yaffa Al Bayyari; University of Texas at El Paso, USA
Wed3WeS.5

Discovering and quantifying the prosodic signals that help manage turn-taking is difficult, in part because of the limitations of commonly used methods. This paper presents an integrated method that uses both perceptually-based analysis and quantitative analysis. The eight activities involved in the method — clarification of aims, problem formulation, corpus preparation, feature discovery, feature combination, hypothesis refinement, tuning, and evaluation — are illustrated using task of finding prosodic cues for back-channel feedback in Arabic.

 naïlon/ — Software for Online Analysis of Prosody
Jens Edlund, Mattias Heldner; KTH, Sweden
Wed3WeS.6

This paper presents naïlon/ — a software package for online real-time prosodic analysis that captures a number of prosodic features relevant for interaction control in spoken dialogue systems. The current implementation captures silence durations; voicing, intensity, and pitch; pseudo-syllable durations; and intonation patterns. The paper provides detailed information on how this is achieved. As an example application of naïlon/, we demonstrate how it is used to improve the efficiency of identifying relevant places at which a machine can legitimately begin to talk to a human interlocutor, as well as to shorten system response times.

From Reaction to Prediction: Experiments with Computational Models of Turn-Taking
David Schlangen; University of Potsdam, Germany
Wed3WeS.3

Deciding when to take (or not to take) the turn in a conversation is an important task. It has been stressed in the descriptive literature that such decisions must involve prediction, as they often seem to be made before a transition place has been reached. In computational systems, however, turn-taking is normally a reaction to parameters like pause length. In this paper, we report on experiments that try to bridge this gap. We describe an experiment (using controlled stimuli) that shows human performance at prediction of turn-taking decisions and then show that a model automatically induced from data can reach a similar level of performance. We then describe a series of experiments on spontaneous dialogue data where we combine pause thresholds with syntactic and prosodic information to make turn-taking decisions, successively reducing the pause threshold until reaction becomes prediction. All our classifiers improve significantly over the baselines; prediction however is shown to be the hardest task, and we discuss additional information sources that could improve it.
Conditional Random Fields for Hierarchical Segment Selection in Text-to-Speech Synthesis

Christian Weiss, Wolfgang Hess; Rheinische Friedrich-Wilhelms-Universität, Germany

This paper presents a hidden Markov model (HMM) based unit selection method for Text-To-Speech (TTS) synthesis application. The aim of this method is to build a corpus whose unit distribution approximates a given target distribution. Corpus selection can be expressed as a set covering problem, which is known to be NP-complete: we therefore resort to a heuristic approach, based on greedy algorithm. We propose the Kullback-Leibler divergence to guide the iterative selection of candidate sentences: indeed, this criterion gives the possibility to control the unit distribution at each step of the algorithm. We first show how to efficiently update, in an incremental manner, this criterion. We then present and discuss experimental results, where our selection algorithm is compared, for various unit sets, with alternative selection criteria.

HMM-Based Unit Selection Using Frame Sized Speech Segments

Zhen-Hua Ling, Ren-Hua Wang; University of Science & Technology of China, China

This paper presents a hidden Markov model (HMM) based unit selection method for concatenative speech synthesis system. Frame sized waveform segments are adopted as basic synthesis units here to increase the coverage rate of candidate units and the chance of finding appropriate ones. In training stage, a set of contextual dependent HMMs are trained with static and dynamic acoustic features. When synthesizing a sentence, the optimal frame sequence is searched out from speech corpus by maximizing the output probability of a sentence HMM constructed according to the contextual information of input text. Listening test proves that proposed method can achieve better performance of synthesized speech compared with the method using state sized units and cost function criterion.

The Target Cost Formulation in Unit Selection Speech Synthesis

Paul Taylor; University of Cambridge, UK

We review the various approaches that have been used to define the target cost in unit selection speech synthesis and show that there are a number of different and sometimes incompatible ways of defining this. We propose that this cost should be thought of as a measure of how similar two units sound to a human listener. We discuss the issue of what features should be used in unit selection and the pros and cons of using derived features such as F0. We then explore some algorithms used to calculate target costs and show that none are really ideal for the problem. Finally, we propose a new solution to this that uses a neural network to synthesise points in acoustic space around which we can build new clusters of units at run time.

Unit Selection and Its Relation to Symbolic Prosody: A New Approach

Daniel Tihelka, Jindřich Matoušek; University of West Bohemia, Czech Republic

Aiming at the improvement of the quality of synthetic speech generated by our native TTS ARTIC, we adopted the unit selection method. Our unit selection module is driven by prosody described solely by high-level symbolic features which are linked to the prosody of synthesized phrases through the phenomena of prosodic synonymy and homonymy. It was confirmed that such an approach not only generates speech with high naturalness but also keeps the richness of prosody. Our first version of this approach significantly increased the quality of the output speech, which was assessed by listeners as very close to natural.

The concept of prosodic synonymy and homonymy is, therefore, further extended and formally described in this paper, and its importance to the unit selection treatment is demonstrated. In addition, the difference of this concept from the concepts most frequently used is shown. Moreover, the first experiment following the formal definition of the problem presented in this paper has been carried out, proving that the whole concept is feasible.

Minimum Generation Error Criterion for Tree-Based Clustering of Context Dependent HMMs

Yi-Jian Wu, Wu Guo, Ren-Hua Wang; University of Science & Technology of China, China

Due to the inconsistency between HMM training and synthesis application in HMM-based speech synthesis, the minimum generation error (MGE) criterion had been proposed for HMM training. This paper continues to apply the MGE criterion for tree-based clustering of context dependent HMMs. As directly applying the MGE criterion results in an unacceptable computational cost, the parameter updating rules of the MGE criterion are simplified to rapidly update the parameters of clustered models, and an appropriate strategy by combining the MGE criterion with the ML criterion is designed to select the optimal question for tree node splitting. From the experiment results, the quality of synthetic speech was improved after applying the MGE criterion for HMM clustering.

Selective-LPC Based Representation of STRAIGHT Spectrum and Its Applications in Spectral Smoothing

Heng Kang, Wenjuan Liu; Chinese Academy of Sciences, China

In this paper we propose a new method to represent STRAIGHT spectrum. The new method provides STRAIGHT spectral parameters with the capability of interpolation and quantization, which is needed
for most speech manipulation, especially for spectral smoothing. The proposed method estimates 2-band selective-LPC whose spectral envelope fits the given STRAIGHT spectrum. With the interpolation properties of LSP, the estimated selective-LPC could be converted to LSP and then simply interpolated. We apply this representation in our spectral smoothing experiments and the results show that this method can get smooth spectral envelope over the segment boundaries. Listening tests prove that this algorithm effectively smooth speech boundaries with little quality degradation.

Towards a Comprehensive Investigation of Factors Relevant to Peak Alignment Using a Unit Selection Corpus

Matthias Jilka, Bernd Möbius; University of Stuttgart, Germany
Wed3BuP.8

This paper aims to demonstrate the use of a unit selection corpus, the IMS German Festival synthesis system [1], in carrying out a comprehensive investigation of factors influencing specific aspects of the phonetic realization of tonal categories. The study restricts itself to the alignment of peaks in H*L pitch accents in German. First results confirm not only well-known effects of syllable structure, e.g., peaks occurring relatively early when there is a sonorant onset or relatively late when there is a sonorant in the coda, but also attest to the special status of the nuclear pitch accent vs. accents occurring earlier in the intonation phrase. Furthermore, instances of H*L in syllables directly at the phrase boundaries (initial or final) are shown to behave significantly differently from those that are located farther away. A similar effect is observed when another pitch accent follows the H*L peak in the very next syllable as opposed to a distance of two or more syllables. In these cases it also matters whether a low or high target is following (the peaks occur relatively later when followed by a L target). The results should have the benefit of both describing the specific characteristics of the voice providing the corpus (allowing a more detailed phonetic realization of tonal categories during the synthesis process) and offering general insights into which factors are relevant to the alignment of H*L peaks in German.

Six Approaches to Limited Domain Concatenative Speech Synthesis

Robert J. Utama1, Ann K. Syrdal2, Alistair Conkie2; 1 CAIP, USA; 2 AT&T Labs Research, USA
Wed3BuP.9

This paper (this work constitute Robert Utama’s master thesis in the Electrical and Computer Engineering program in Rutgers University) describes the creation of 6 limited-domain Text-to-Speech (TTS) systems that are constrained to digit string and natural number domains (cardinal numbers only). Unit selection-based concatenative TTS systems were implemented in MATLAB to fulfill this goal. We evaluate and discuss various factors that can influence the naturalness or overall quality of the synthesized voice. Some of the factors studied are the length and type of the synthesis unit and the extent of co-articulation represented in the recorded speech database. In the end, we show that it is possible to create a high quality limited domain TTS system either with maximal or with carefully controlled minimal effects of co-articulation.

From Pre-Recorded Prompts to Corporate Voices: On the Migration of Interactive Voice Response Applications

V. Fischer, S. Kunzmann; IBM Deutschland Entwicklung GmbH, Germany
Wed3BuP.10

This paper describes our efforts towards the creation of corporate synthetic voices from low quality speech data, as it can typically be found on many Interactive Voice Response (IVR) units. In doing so, we first touch on several normalization techniques that aim on a better support of a highly automated voice construction process. Subsequently, we describe methods for the creation of enriched corporate voices which integrate speech recordings from different speakers in order to overcome problems arising from limited domain training data. Experiments are described which demonstrate the feasibility of the approach by comparing it to a less flexible solution that uses pre-recorded prompts in combination with a large footprint standard concatenative synthesizer. Results show that the enriched voices clearly outperform those voices build solely from IVR data, while achieving almost the same overall rating as the pre-recorded prompts solution.

Automatic Speech Segmentation with Multiple Statistical Models

Seung Seop Park, Jong Won Shin, Nam Soo Kim; Seoul National University, Korea
Wed3BuP.11

In this paper, we propose a novel approach to improve the performance of automatic speech segmentation techniques for concatenative text-to-speech synthesis. A number of automatic segmentation machines (ASMs) are simultaneously applied and the final boundary time marks are drawn from the multiple segmentation results. To identify the best time mark among those provided by the multiple ASMs, we apply a candidate selector trained over a set of manually-segmented speech database. The candidate selector defines a mapping from the phonetic boundary to the best ASM index which will output the time mark that may be closest to the manual segmentation result. The experimental results show that our approach dramatically improves the segmentation accuracy.

Evaluation of Perceptual Quality of Control Point Reduction in Rule-Based Synthesis

Kimmo Pärssinen, Marko Moberg; Nokia Technology Platforms, Finland
Wed3BuP.12

Text-to-speech implementations on embedded devices usually require low memory consumption and computational complexity. Due to its simplicity, formant synthesizer is still an attractive solution for some applications. The formant values and transitions are controlled by a set of rules, which assign control points for synthesis parameters. The paper investigates the possibility to reduce the number of control points for formant contours from four to two per phoneme. The reduced model contained only the values at the end of the offset transition and in the beginning of the offset transition. Various interpolation techniques were studied but linear interpolation was used for its simplicity. The 4- and 2-point models were compared in a listening evaluation test. The results show that the reduction of control points does not have any effect on the perceived quality. The dynamic, context dependent positioning of the two control points preserves the most essential information of formant contours.

Segment Connection Networks for Corpus-Based Speech Synthesis

Geert Coorman; Nuance Communications, Belgium
Wed3BuP.13

In this paper an efficient segment selection method for corpus-based speech synthesis is presented. Traditional unit-selectors use dynamic programming (DP) to find in a fully connected segment network the most appropriate segment sequence based on target- and concatenation costs. Instead of performing a full DP search, the presented unit-selector applies fast transformations on binary valued segment connection matrices. It is further also shown that this technique can be expanded to supra-segmental unit-selection without altering the segment size in the database.
Detection of a Third Speaker in Telephone Conversations

Uchechukuwa O. Ojoegbu1, Ananth N. Iyer1, Robert E. Yantorno1, Stanley J. Wenndt2; 1Temple University, USA; 2Air Force Research Laboratory, USA

Wed3CaP.1

Differentiating speakers participating in telephone conversations is a challenging task in speech processing because only short consecutive utterances can be examined for each speaker. Research has shown that, given only brief utterances (1 second or less), humans can recognize speakers with an accuracy of about 54% on average. The task becomes even more challenging when no information about the speakers is known a priori. In this paper, a technique for determining whether there are two or three speakers participating in a telephone conversation is presented. This approach assumes no knowledge or information about any of the participating speakers. The technique is based on comparing short utterances within the conversation and deciding whether or not they belong to the same speaker. The applications of this research include 3-way call detection and speaker and deciding whether or not they belong to the same speaker. This approach assumes no knowledge or information about any of the participating speakers. The technique is based on comparing short utterances within the conversation and deciding whether or not they belong to the same speaker. The applications of this research include 3-way call detection and speaker identification. The proposed technique tested on actual telephone conversations.

Improvement Speaker Clustering Using Global Similarity Features

Konstantin Biatov, Joachim Köhler; Fraunhofer IMK, Germany

Wed3CaP.2

In this paper global similarity features that improve speaker clustering based on standard bottom-up clustering are proposed. The novelty of this approach lies in the fact that it exploits the hypothesis that audio segments belonging to the same speaker should demonstrate similar global behavior, exhibit the same similarity and dissimilarity with all the other segments. Every segment is represented by a global similarity vector whose components are encoded by the distance between that segment and each of the other segments to be clustered. The distance between global similarity vectors is used for pre-segmentation of segment pairs having high global similarity for further merging. In this paper inter-segment distance for global similarity vectors based on Bayesian Information Criterion (BIC) and based on adapted cross likelihood ratio (CLR) are investigated. The evaluation, performed on artificially simulated conversations from the HTIMIT database and 23% error rate when tested on actual telephone conversations.

Voting for Two Speaker Segmentation

Balakrishnan Narayanaswamy, Rashmi Gangadharaiyah, Richard M. Stern; Carnegie Mellon University, USA

Wed3CaP.3

The process of locating the end points of each speakers voice in an audio file and then clustering segments based in speaker identity is called speaker segmentation. In this paper we present a method for two speaker segmentation, though it can be extended to more than two speakers. Most methods for speaker segmentation and clustering start with an initial computationally inexpensive speaker segmentation method, followed by a more accurate segment clustering. In this paper we describe a simple algorithm that improves the accuracy of the segment clustering while not increasing the computational complexity. Since the clustering is done iteratively, the improvement in each segment clustering step results in a significant overall increase in segmentation accuracy and cluster purity. We borrow ideas from speaker recognition to perform segment clustering by frame based voting. We look at each frame as an independent classifier deciding which speaker generated that segment. These ‘classifiers’ are combined by voting to make a decision as to which segments should be clustered together. This simple change leads to 56.9% decrease in error rates on a segmentation task for the SWITCHBOARD corpus.

Unsupervised Model Adaptation for Speaker Verification

Alexandre Preti, Jean-François Bonastre; LIA, France

Wed3CaP.4

This paper deals with unsupervised model adaptation for speaker recognition. Two adaptation schemes are proposed, the first one is based on a test by test model adaptation and the second one proposes a batch mode, where the adaptation is performed using a set of tests before computing the decision score for each of them. The experiments are conducted thanks to the NIST SRE 2005 database. This paper shows clearly the interest of unsupervised model adaptation when enough test data is available (batch mode) and the intrinsic difficulty of an online (test by test) adaptation mode.

A Quality Measure Method Using Gaussian Mixture Models and Divergence Measure for Speaker Identification

Rong Zheng, Shuwu Zhang, Bo Xu; Chinese Academy of Sciences, China

Wed3CaP.5

Previous work has demonstrated the promise of frame-level quality measure methods to robust speaker recognition. This paper explores the issues involved in applying soft estimates to quality measures as weighting factors in score computation. A quality measure algorithm using Gaussian mixture density and Jensen divergence measure is presented for traditional GMM-UBM scoring mechanism. Derivation and validation of the quality measurement are reported in this paper. We investigate the usefulness of different feature processing, different GMM-based quality models and incorporation of divergence measure for quality estimation. Comparison experiments performed on the NIST1999 SRE corpus show the effectiveness of the proposed methods.

Gammatone Auditory Filterbank and Independent Component Analysis for Speaker Identification

Yushi Zhang, Waleed H. Abdulla; University of Auckland, New Zealand

Wed3CaP.6

Feature extraction is the key procedure when aiming at robust speaker identification. The most commonly used feature extraction techniques work successfully only in clean or matched environments. Accurate speaker identification is made difficult due to a number of factors, with handset/channel mismatch and environmental noise being the most prominent. This paper presents a novel technique which based on Gammatone filterbank (GTF) and independent component analysis (ICA). The presented method first relies on the Gammatone filterbank to emulate the human cochlea frequency resolution. By using ICA, it extracts the dominant components from these frequency banks. The extracted features emphasis the difference in the statistical
structures among the speakers, which can model the distribution of the individuals. Compared to the commonly used techniques, such as linear predictive cepstral coefficients (LPCC), Mel-frequency cepstrum coefficients (MFCC) and perceptual linear predictive (PLP), the proposed method is more robust to additive noises and yields higher recognition rate in mismatch environments in a text-independent speaker identification system.

Study on Speaker Verification on Emotional Speech
Wei Wu, Thomas Fang Zheng, Ming-Xing Xu, Huan-Jun Bao; Tsinghua University, China
Wed3CaP.7
Besides background noise, channel effect and speaker's health condition, emotion is another factor which may influence the performance of a speaker verification system. In this paper, the performance of a GMM-UBM based speaker verification system on emotional speech is studied. It is found that speech with various emotions aggravates the verification performance. Two reasons for the performance aggravation are analyzed, they are mismatched emotions between the speaker models and the test utterances, and the articulating styles of certain emotions which create intense intra-speaker vocal variability. In response to the first reason, an emotion-dependent score normalization method is proposed, which is borrowed from the idea of Hnorm.

On the Fusion of Prosody, Voice Spectrum and Face Features for Multimodal Person Verification
M. Farrús, A. Garde, P. Ejarque, J. Luque, Javier Hernando; Universitat Politècnica de Catalunya, Spain
Wed3CaP.8
Multimodal person recognition systems normally use short-term spectral features as voice information. In this paper prosodic information is added to a system based on face and voice spectrum features. By using two fusion techniques, support vector machines and matcher weighting, different fusion strategies based on the fusion of monomodal scores in several steps are proposed. The performance of the system is clearly improved when the prosodic information is added and the best results are achieved when prosodic scores are previously fused and the resulting scores are fused again with spectral and facial scores. Speech and face scores have been obtained upon Switchboard-I and XM2VTS databases respectively.

An MRI Based Study of the Acoustic Effects of Sinus Cavities and Its Application to Speaker Recognition
Taran Pruthi, Carol Y. Espy-Wilson; University of Maryland, USA
Wed3CaP.9
The goal of this paper is to explore the effects of changes in velar coupling area and oral cavity configuration on the polos and zeros introduced in the nasalized vowel and nasal consonant spectra due to the sphenoidal and maxillary sinuses. MRI data for the vocal tract and nasal tract of one speaker was used to simulate the spectra of the nasalized vowels /a, æ, i, u/, and nasal consonants /m, n/ with different coupling areas. It is shown that during nasalized vowels, the frequencies of both polos and zeros due to the sinuses change with a change in the velar coupling area or the vowel. It is also shown that during nasal consonants, the zero frequencies are constant, and the pole frequencies are more stable as compared to nasalized vowels. This study, therefore, corroborates the use of nasal consonant spectra for speaker recognition and raises doubts on the potential benefits of using nasalization during vowels for that purpose.

Speaker Verification with Non-Audible Murmur Segments
Mariko Kajima1, Tomoko Matsui2, Hiromichi Kawanami1, Hiroshi Saruwatari1, Kiyohiro Shikano1;
1Nara Institute of Science & Technology, Japan; 2Institute of Statistical Mathematics, Japan
Wed3CaP.10
We propose a speaker verification method using non-audible murmur (NAM) segments, which are different from normal speech and hard for other people to catch them. To use NAM, we therefore take a text-dependent verification strategy in which each user utters her/his own keyword phrase and utilize not only speaker-specific but also keyword-specific acoustic information. We expect this strategy to yield a relatively high performance. NAM segments, which consist of multiple short-term feature vectors, are used as input vectors to capture keyword-specific acoustic information well. To handle segments with a large number of dimensions, we use the support vector machine (SVM). In experiments using NAM data of 19 male and 10 female speakers recorded in three different sessions, we achieved equal error rates of 0.04% (male) and 1.1% (female) when using 145-ms-long NAM segments. These rates are half or less those obtained with 25-ms-long input vectors.

Automatic Recognition of Speakers' Age and Gender on the Basis of Empirical Studies
Christian Müller; DFKI GmbH, Germany
Wed3CaP.11
This paper describes a system that exploits the paralinguistic information in the speech to estimate the speakers’ age and gender. Compared with previously published work, the so called AGENDER approach involves finer grained speaker classes and achieves a significantly higher classification accuracy. The introduction encompasses various application examples representing the actual AGENDER project context. Then hypotheses, method and a representative selection of results from extensive corpus analyses are presented, that build the empirical basis for the machine learning. Finally, the AGENDER approach on speaker classification is outlined, involving the comparison of different classification methods as well as evaluation results. The paper finishes with an outlook on extensions that are scheduled for the next project phase.

Text-Independent Speaker Identification in Birds
E.J.S. Fox, J.D. Roberts, M. Bennamoun; University of Western Australia, Australia
Wed3CaP.12
Speaker recognition is used to identify individual humans, but has rarely been applied to other species. To be applicable to the wide variety of bird species, text-independent speaker identification would be the most effective method. This is the first paper to report results of this technique in a species other than humans. Mel-frequency cepstral coefficients were extracted from recordings of three bird species and a multilayer perceptron was used as the classifier in each species. First, the song types used in training and testing were not controlled for, and these conditions gave an accuracy of 68–100%. Next the recordings of the wagtails and scrub-birds were split into their respective song types, a network was trained with one song type from each individual and tested with a different song type. With these purely text-independent conditions the accuracy was 71–90%.

Automatic Acoustic Identification of Insects Inspired by the Speaker Recognition Paradigm
Ilyas Potamitis1, Todor Ganchev2, Nikos Fakotakis2;
1Technological Educational Institute of Crete, Greece; 2University of Patras, Greece
Wed3CaP.13

This work reports our research efforts towards developing efficient equipment for the automatic acoustic recognition of insects. In particular, we discuss the characteristics of the acoustic patterns of a target insect family, namely the cricket family. To address the recognition problem we apply a feature extraction methodology that has been inspired by well documented tactics of speech processing, which were adapted here to the specifics of the sound production mechanism of insects, in combination with state-of-the-art speaker identification technology. We apply this approach to a large and well documented database of families and subfamilies of cricket sounds, and we report results that exceed 99% recognition accuracy on the levels of family and subfamily, and 94% on the level of a specific insect out of 105 species. We deem this equipment will be of practical benefit for non-intrusive acoustic environmental monitoring applications as it is directly expandable to other insect species.

**Wed3FoP : Multichannel Speech Enhancement/Speech Perception**

**Poster Session, Allegheny Ballroom Foyer, 16:30 - 18:30**

**Chair: Hyung-Min Park, KAIST, Korea**

---

**Improved Hybrid Microphone Array Post-Filter by Integrating a Robust Speech Absence Probability Estimator for Speech Enhancement**

**Junfeng Li**

**1Masato Akagi**

**2Yöiti Suzuki**

**1Tohoku University, Japan; 2JAIST, Japan**

**Wed3FoP.1**

To improve the performance of multi-channel speech enhancement algorithms, we previously proposed a hybrid Wiener post-filter for microphone arrays under the assumption of a diffuse noise field [4]. In considering the speech presence uncertainty, we further improve the hybrid post-filter presented before by integrating a novel robust estimator for the a priori speech absence probability, which makes full use of the correlation characteristics of the noises on different microphone pairs and hence offers the much more accurate speech absence probability estimates. The effectiveness of this improved hybrid post-filter was finally confirmed by the experiments using multi-channel recordings in various car environments.

---

**Soft Decision Combining for Dual Channel Noise Reduction**

**Timo Gerkmann, Rainer Martin; Ruhr-Universität Bochum, Germany**

**Wed3FoP.2**

We present a flexible dual channel noise reduction algorithm that combines the information of two microphone channels in the discrete Fourier transform domain. This algorithm can cope with substantially different signal-to-noise ratios at different time-frequency bins in both channels. The output is obtained by combining the outputs of two single channel filters and a dual channel filter weighted by the probability of speech presence. The algorithm has a low latency and does not require any additional information such as the microphone positions.

---

**An Improved Affine Projection Algorithm Based Crosstalk Resistant Adaptive Noise Canceller**

**Guo Chen, Vijay Parsa; University of Western Ontario, Canada**

**Wed3FoP.3**

This paper presents an improved affine projection algorithm based crosstalk resistant adaptive noise canceller (CRANC). The proposed CRANC consists of three adaptive filters, one of which is used for detecting speech activity and the others are employed to estimate the interference and crosstalk transfer functions, respectively. An efficient sequential speech detector is also proposed in this paper, which is completely built into the adaptive strategy and thus causes no delay in the input and output of the noise canceller. The proposed CRANC was tested and compared with the conventional CRANC in the white Gaussian noise environment. Experimental results show that the proposed CRANC achieved higher signal-to-noise ratio (SNR) improvements than the conventional CRANC.

---

**An Optimum Microphone Array Post-Filter for Speech Applications**

**Stamatis Leukimmitis, Dimitrios Dimitriadis, Petros Maragos; National Technical University of Athens, Greece**

**Wed3FoP.4**

This paper proposes a post-filtering estimation scheme for multichannel noise reduction. The proposed method extends and improves the existing Zelinski’s and, the most general and prominent, McCowan’s post-filtering methods that use the auto- and cross-spectral densities of the multichannel input signals to estimate the transfer function of the Wiener post-filter. A major drawback of these two speech enhancement algorithms is that the noise power spectrum at the beamformer’s output is over-estimated and therefore the derived filters are sub-optimal in the Wiener sense. The proposed method deals with this problem and can be considered as an optimal post-filter that is appropriate for a wide variety of different noise fields. In experiments over real-noise multichannel recordings, the proposed technique is shown to obtain a significant headstart over the other methods in terms of signal-to-noise ratio and speech degradation measures. In addition it is used for ASR experiments where promising preliminary results are presented.

---

**Multi-Microphone Periodicity Function for Robust F0 Estimation in Real Noisy and Reverberant Environments**

**Federico Flego, Maurizio Omologo; ITC-irst, Italy**

**Wed3FoP.5**

This paper outlines a new method to extract F0 from distant-talking speech signals acquired by a microphone network, which exploits the redundancy across the signals proceeding from each microphone, by jointly processing the different contributes. To this purpose, a multi-microphone periodicity function is derived from the magnitude spectrum computed on each microphone signal. This function allows to estimate F0 reliably, even under reverberant conditions, without the need of any post-processing or smoothing technique. Experiments, conducted on real lectures, showed that the proposed frequency-domain algorithm is more suitable than other time-domain based ones.

---

**A New Dual-Microphone Speech Enhancement Method for Oriented Noises**

**H.R. Abutalebi, M. Pourahmadi, M.R. Aghabozorgi; Yazd University, Iran**

**Wed3FoP.6**

In this paper, we examine a modified Cross-Talk Resistant Adaptive Noise Canceller (CTRANC) structure that contains a delay unit on the primary channel to solve the causality constraint of conventional CTRANC. This Asymmetric CTRANC (ACTRANC) structure allows flexible alignment of noise source with the sensors array when the speech source is fixed in its position. Inserting delay unit on the primary path releases the placement of the noise sources and consequently, improves the capability of the structure in noise reduction. This can also be interpreted as the releasing causality constraint on the adaptive filters. We have shown that the best delay value in the structure is proportional to the propagation delay between reference and primary microphones (in terms of sample). Objective evaluations and informal listening tests demonstrate superior performance of the proposed method rather than conventional CTRANC.
50 Years Late: Repeating Miller-Nichely 1955
Andrew Lovitt, Jont B. Allen; University of Illinois at Urbana-Champaign, USA
Wed3FoP.7
Portions of the procedure and analysis of the wide-band noise masking experiment in Miller-Nichely’s 1955 JASA paper (MN55) was repeated and a new set of data was collected in 2005. This classic paper is a commonly referenced work in which confusion matrices were collected for a set of consonant-vowels (CVs). From an analysis of the original results, they made conclusions about the robustness of various distinctive features when the CVs are degraded in masking noise. Our repeat experiment shows a number of similarities and differences. The two experiments show significantly different amounts of relative information transmitted for each distinctive feature. In the repeat experiment the voicing feature is more robust whereas the place feature is more robust.

New 20-Word Lists for Word Intelligibility Test in Japanese
Shuichi Sakamoto1, Tadahiro Yoshikawa1, Shigeaki Amano2, Yōiti Suzuki1, Tadahisa Kondo1;1 Tohoku University, Japan; 2 NTT Corporation, Japan
Wed3FoP.8
“Familiarity controlled Word lists 2003 (FW03)” is a set of 20 lists of 50 Japanese words each in four ranks of word familiarity (for a total of 4000 words). FW03 is useful for evaluating the ability of a person to hear human speech. However, it has been pointed out that the FW03 lists have too many words to evaluate the ability of a person to hear human speech in a short time. This can cause problems, such as when evaluating the ability of an elderly person to hear human speech during a medical exam. To remedy this problem, we propose new word lists that are a subset of FW03. The new lists were created by considering both phonetic balance and word familiarity. Each of the new word lists comprises only 20 words. This reduces the duration of an examination to around two-fifths of a conventional FW03-based examination. Word intelligibility tests suggest that these new lists provide word intelligibility scores nearly equal to those resulting from FW03.

Sparseness and Speech Perception in Noise
Guoping Li, Mark E. Lutman; University of Southampton, UK
Wed3FoP.9
Can we model speech recognition in noise by exploring higher order statistics of the combined signal? How will changes in these statistics affect speech perception in noise? This study addresses these questions in two experiments. One investigated the relationship between an established “glimpsing” model and the fourth order statistic, kurtosis. The glimpsing model [1] proposes that listeners can provide word intelligibility scores nearly equal to those resulting from FW03.

An Assessment of Automatic Speech Recognition as Speech Intelligibility Estimation in the Context of Additive Noise
Wei M. Liu, John S.D. Mason, Nicholas W.D. Evans, Keith A. Jellyman; University of Wales Swansea, UK
Wed3FoP.10
This paper investigates the potential applicability of automatic speech recognition (ASR) and 6 well-reported objective quality measures for the task of ranking intelligibility of speech degraded by different real life background noises. In a recent investigation ASR has been reported to give high subjective correlation with human assessment when tested with various system degradations. This paper extends this investigation in two directions. First, the usefulness of the measures in the context of different real-life noises is considered. Second, the direct correspondence between statistics computed by an ASR system and human perceived intelligibility is assessed. Subjective listening tests are carried out to provide ground truth. Results show that ASR and WSS (weighted spectral slope) are the only two measures out of the seven considered to give good correlation with human opinion. Specially noted is performance of ASR with correlations ranging from 0.77 to 0.90.

Underlying Quality Dimensions of Modern Telephone Connections
Marcel Wältermann1, Kirstin Scholz2, Alexander Raake3, Ulrich Heute4, Sebastian Möller1; 1 Ruhr-Universität Bochum, Germany; 2 Christian-Albrechts-Universität zu Kiel, Germany; 3 Technische Universität Berlin, Germany
Wed3FoP.11
It is the aim of the present paper to analyze the perceptual quality dimensions of modern telephone connections. Such connections differ from standard connections in their time-variant characteristics (e.g., due to Voice-over-IP transmission or due to noise reduction algorithms) and their user interfaces (e.g., hands-free terminals). With the help of two independent auditory experiments with subsequent multidimensional analyses, three perceptual dimensions were identified for a diverse set of stimuli. These dimensions were labeled “directness/frequency content”, “continuity”, and “noisiness”. Overall listening quality scores were collected in a separate experiment. A mapping of the obtained dimensions onto the overall listening quality scores by means of a linear model revealed that “continuity” appears to be the most important dimension in terms of overall listening quality.

An ERB Loudness Pattern Based Objective Speech Quality Measure
Guo Chen, Vijay Parsa, Susan Scollie; University of Western Ontario, Canada
Wed3FoP.12
This paper presents an objective speech quality measure which is based on loudness patterns using the equivalent rectangular band-width (ERB) scale. The proposed measure, called the loudness pattern distortion (LPD), is computed from the differences between the loudness patterns of the original and processed speech. The LPD measure takes into account the transmission through the outer and middle ear, the calculation of an excitation pattern from the physical spectrum, and the transformation of an excitation pattern to a loudness pattern. The effectiveness of the proposed measure was demonstrated by experimental evaluations in comparison with the standard ITU-T P.862 (PESQ) using three coded speech database of the ITU-T P-series Supplementary 23.
Thu1A1O : Diarization in ASR
Oral Session, Allegheny Ballroom, left, 10:30 - 12:30
Chair: Dirk Van Compernolle, Katholieke Universiteit Leuven, Belgium

A Spectral Clustering Approach to Speaker Diarization
Huaizhong Ning, Ming Liu, Hao Tang, Thomas S. Huang; University of Illinois at Urbana-Champaign, USA
Thu1A1O.1
In this paper, we present a spectral clustering approach to explore the possibility of discovering structure from audio data. To apply the Ng-Jordan-Weiss (NJW) spectral clustering algorithm to speaker diarization, we propose some domain specific solutions to the open issues of this algorithm: choice of metric; selection of scaling parameter; estimation of the number of clusters. Then, a postprocessing step -- "Cross EM refinement" -- is conducted to further improve the performance of spectral learning. In experiments, this approach has performance very similar to the traditional hierarchical clustering on the audio data of Japanese Parliament Panel Discussions, but it runs much faster than the latter.

BINSEG: An Efficient Speaker-Based Segmentation Technique
Jindrich Zdansky; Technical University of Liberec, Czech Republic
Thu1A1O.2
In this paper we present a new efficient approach to speaker-based audio stream segmentation. It employs binary segmentation technique that is well-known from mathematical statistic. Because integral part of this technique is hypotheses testing, we compare two well-founded (Maximum Likelihood, Informational) and one commonly used (BIC difference) approach for deriving speaker change-test statistics. Based on results of this comparison we propose both off-line and on-line speaker change detection algorithms (including way of effective training) that have merits of high accuracy and low computational costs. In simulated tests with artificially mixed data the on-line algorithm identified 95.7% of all speaker changes with precision of 96.9%. In tests done with 30 hours of real broadcast news (in 9 languages) the average recall was 74.4% and precision 70.3%.

Multi-Stream Speaker Diarization Systems for the Meetings Domain
Ascensión Gallardo-Antolín1, Xavier Anguera2, Chuck Wooters2;1 Universidad Carlos III de Madrid, Spain; 2 International Computer Science Institute, USA
Thu1A1O.3
In the context of speech and speaker recognition systems, it is well known that the combination of different feature streams can improve significantly their performance. However, the application of multi-stream (MS) techniques to speaker diarization systems has not been extensively studied. In this paper, we address this issue: we formulate different MS techniques, such as feature combination, probability combination and selection, for their specific application to the segmentation and clustering modules of a speaker diarization system. We evaluate the different methods proposed for the meetings domain (RT04s database and two different pairs of streams: first, MFCC and PLP and second, MFCC and prosodic features. For both types of multi-streams, results show that the MS probability combination approach applied to the segmentation stage clearly outperforms the single-stream, MS feature combination and MS selection systems.

Improved Performance Evaluation of Speech Event Detectors
Carla Lopes, Fernando Perdigão; Instituto de Telecomunicações, Portugal
Thu1A1O.4
The goal of event-based (EB) systems is the detection of the occurrence of important elements in the speech signal for different sound classes. In a speech recognition system, events can be combined to detect phones, words or sentences, or to identify landmarks to which a classifier or a decoder could be synchronized. The time boundaries of the events are then as important as the events themselves. Accordingly, the assessment of EB systems must take into account not only the correct identified sequence of events but also their correct time localization. Usually, only the token sequence or its boundaries are taken for evaluation. In this paper we propose an extension to standard recognition evaluation procedure, which combines recognition and segmentation performance. In our proposal, a modified Levenshtein algorithm is used in the alignment between labeled and recognized events, where the degree of overlapping between them is taken in the local distance definition. We evaluate our approach on a rule based event detector, using the TIMIT corpus and compare the results of the new evaluation procedure with standard metrics. The results show that accuracy drops if alignment is made as a function of the overlapping between labels; nevertheless the agreement with the labeled boundaries is significantly improved.

Speaker Diarization for Multiple Distant Microphone Meetings: Mixing Acoustic Features and Inter-Channel Time Differences
Jose M. Pardo, Xavier Anguera, Chuck Wooters; International Computer Science Institute, USA
Thu1A1O.5
Speaker diarization for recordings made in meetings consists of identifying the number of participants in each meeting and creating a list of speech time intervals for each participant. In recently published work [7] we presented some experiments using only TDOA values (Time Delay Of Arrival for different channels) applied to this task. We demonstrated that information in those values can be used to segment the speakers. In this paper we have developed a method to mix the TDOA values with the acoustic values by calculating a combined log-likelihood between both sets of vectors. Using this method we have been able to reduce the DER by 16.34% (relative) for the NIST RT05s set (scored without overlap and manually transcribed references) the DER for our devel06s set (scored with overlap and force-aligned references) by 21% (relative) and the DER for the NIST RT06s (scored with overlap and manually transcribed references) by 13% (relative).

Low-Complexity and Efficient Classification of Voiced/Unvoiced/Silence for Noisy Environments
Tuan Van Pham, Gernot Kubin; Graz University of Technology, Austria
Thu1A1O.6
This paper describes a low-complexity and efficient speech classifier for noisy environments. The proposed algorithm utilizes the advantage of time-scale analysis of the Wavelet decomposition to classify speech frames into voiced, unvoiced and silence classes. The classifier uses only one single multidimensional feature which is extracted from the Teager energy operator of the wavelet coefficients. The feature is enhanced and compared with quantile-based adaptive thresholds to detect phonetical classes. Furthermore, to save memory, the adaptive thresholds are replaced by a slope tracking method on the filtered feature. These algorithms are tested with the TIMIT database and additive white, car, factory noise, and compared with other methods to demonstrate their superior performance and robustness.
Thu1A2O : Language Model Adaptation, Refinement, and Evaluation

Oral Session, Allegheny Ballroom, center, 10:30 - 12:30
Chair: Andreas Stolcke, SRI International, USA

Unsupervised Language Model Adaptation Based on Automatic Text Collection from WWW

Motoyuki Suzuki, Yasutomo Kajiiura, Akinori Ito, Shozo Makino; Tohoku University, Japan
Thu1A2O.1

An n-gram trained by a general corpus gives high performance. However, it is well known that a topic-specialized n-gram gives higher performance than that of the general n-gram. In order to make a topic specialized n-gram, several adaptation methods were proposed. These methods use a given corpus corresponding to the target topic, or collect documents related to the topic from a database. If there is neither the given corpus nor the topic-related documents in the database, the general n-gram cannot be adapted to the topic-specialized n-gram.

In this paper, a new unsupervised adaptation method is proposed. The method collects topic-related documents from the world wide web. Several query terms are extracted from recognized text, and collected web pages given by a search engine are used for adaptation. Experimental results showed the proposed method gave 7.2 points higher word accuracy than that given by the general n-gram.

Unsupervised Language Model Adaptation Using Latent Semantic Marginals

Yik-Cheung Tam, Tanja Schultz; Carnegie Mellon University, USA
Thu1A2O.2

We integrated the Latent Dirichlet Allocation (LDA) approach, a latent semantic analysis model, into unsupervised language model adaptation framework. We adapted a background language model by minimizing the Kullback-Leibler divergence between the adapted model and the background model subject to a constraint that the marginalized unigram probability distribution of the adapted model is equal to the corresponding distribution estimated by the LDA model — the latent semantic marginals. We evaluated our approach on the RT04 Mandarin Broadcast News test set and experimented with different LM training settings. Results showed that our approach reduces the perplexity and the character error rates using supervised and unsupervised adaptation.

Unsupervised Language Model Adaptation for Mandarin Broadcast Conversation Transcription

David Mrva, Philip C. Woodland; University of Cambridge, UK
Thu1A2O.3

This paper investigates unsupervised language model adaptation on a new task of Mandarin broadcast conversation transcription. It was found that N-gram adaptation yields 1.1% absolute character error rate gain and continuous space language model adaptation done with PLSA and LDA brings 1.3% absolute gain. Moreover, using broadcast news language model alone trained on large data under-performs a model that includes additional small amount of broadcast conversations by 1.8% absolute character error rate. Although, broadcast news and broadcast conversation tasks are related, this result shows their large mismatch. In addition, it was found that it is possible to do a reliable detection of broadcast news and broadcast conversation data with the N-gram adaptation.

Language Model Adaptation for Tiny Adaptation Corpora

Dietrich Klakow; Saarland University, Germany
Thu1A2O.4

In this paper we address the issue of building language models for very small training sets by adapting existing corpora. In particular we investigate methods that combine task specific unigrams with longer range models trained on a background corpus. We propose a new method to adapt class models and show how fast marginal adaptation can be improved. Instead of estimating the adaptation unigram only on the adaptation corpus, we study specific methods to adapt unigram models as well. In extensive experimental studies we show the effectiveness of the proposed methods. As compared to FMA as described in [1] we obtain an improvement of nearly 60% for ten utterances of adaptation data.

Pronunciation Dependent Language Models

Andrej Ljolje; AT&T Labs Research, USA
Thu1A2O.5

Speech recognition systems are conventionally broken up into phonemic acoustic models, pronouncing dictionaries in terms of the phonemic units in the acoustic model and language models in terms of lexical units from the pronouncing dictionary. Here we explore a new method for incorporating pronunciation probabilities into recognition systems by moving them from the pronouncing lexicon into the language model. The advantages are that pronunciation dependencies across word boundaries can be modeled including contextual dependencies like geminates or consistency in pronunciation style throughout the utterance. The disadvantage is that the number of lexical items grows proportionally to the number of pronunciation alternatives per word and that language models which could be trained using text, now need phonetically transcribed speech or equivalent training data. Here this problem is avoided by only considering the most frequent words and word clusters. Those new lexical items are given entries in the dictionary and the language model dependent on the chosen pronunciation. The consequence is that pronunciation probabilities are incorporated into the language model and removed from the dictionary, resulting in an error rate reduction. Also, the introduction of pronunciation dependent word pairs as lexical items changes the behavior of the language model to approximate higher order n-gram language models, also resulting in improved recognition accuracy.

Improving Perplexity Measures to Incorporate Acoustic Confusability

Amit Anil Nanavati, Nitendra Raijput; IBM India Research Lab, India
Thu1A2O.6

Traditionally, Perplexity has been used as a measure of language model performance to predict its goodness in a speech recognition system. However this measure does not take into account the acoustic confusability between words in the language model. In this paper, we introduce Equivocality — modification of the perplexity measure for it to incorporate the acoustic features of words in a language. This gives an improved measuring criterion that matches much better with the recognition results than conventional Perplexity measure. The acoustic distance is used as a feature to represent the acoustic characteristic of the language model. This distance is measurable only with the acoustic model parameters and does not require any experimentation. We derive the Equivocality measure and calculate it for a set of grammars. Speech recognition experiments further justify the appropriateness of using Equivocality over Perplexity.
A recently developed fast MR imaging system is utilized for a study of emotional speech production. Speech utterances and corresponding mid-sagittal vocal tract images are simultaneously acquired by the MRI system. Neutral, angry, sad and happy emotions are simulated by a male American English speaker. The MRI system and analysis results are described in this report. In general articulation is found to be more active in terms of the rate of vocal tract shaping and the ranges of spectral parameter values in emotional speech. It is confirmed that angry speech is characterized by wider and faster vocal tract shaping. Moreover, angry speech shows the more prominent usage of the pharyngeal region than any other emotions. It is also observed that the average vocal tract length above the false vocal folds varies as a function of emotion and that happy speech exhibit relatively shorter length than other emotions. It is likely that this is due to the elevation of the larynx and that may facilitate the higher pitch and larger pitch range manipulation to encode happy emotional quality by the speaker.

Reconstructing Tongue Movements from Audio and Video
Hedvig Kjellström1, Olov Engwall2, Olle Bälter2; 1FOI, Sweden; 2KTH, Sweden
Thu1A3O.4
This paper presents an approach to articulatory inversion using audio and video of the user’s face, requiring no special markers. The video is stabilized with respect to the face, and the mouth region cropped out. The mouth image is projected into a learned independent component subspace to obtain a low-dimensional representation of the mouth appearance. The inversion problem is treated as one of regression; a non-linear regressor using relevance vector machines is trained with a dataset of simultaneous images of a subject’s face, acoustic features and positions of magnetic coils glued to the subject’s tongue. The results show the benefit of using both cues for inversion. We envisage this inversion method to be part of a pronunciation training system with articulatory feedback.

New Considerations for Vowel Nasalization Based on Separate Mouth-Nose Recording
Gang Feng, Cyril Kotenkov; ICP, France
Thu1A3O.5
In this paper an experimental setting is described which allows to separately record the speech signals emitted from the mouth and the nostrils. A series of transitions from an oral configuration to the nasopharyngeal configuration is then recorded with a speaker capable to control the movement of his velum while minimizing that of other articulators. The analysis of the recorded signals in the light of simulations with an acoustic model clearly shows that the formant evolution pattern observed in the mouth output is essentially caused by a modification of the oral tract shape due to the velum lowering and the pharyngeal constriction, while the connection with the nasal tract can be neglected in first approximation. On the other hand, the velum movement controls rather the nose output amplitude than its spectral pattern.

An Acoustic and Articulatory Study of Lombard Speech: Global Effects on the Utterance
Maëva Garnier1, Lucie Bailly1, Marion Dohen2, Pauline Welby2, Hélène Lavenbruck2; 1LAM, France; 2ICP, France
Thu1A3O.6
This study aims at characterizing the acoustic and articulatory modifications that occur in speech in noisy environments, and at examining them as compensatory strategies. Audio, EGG and video signals were recorded for a female native speaker of French. The corpus consisted of short sentences with a subject-verb-object (SVO) structure. The sentences were recorded in three conditions: silence, 85dB white noise, and 85dB cocktail party noise. Labial parameters were extracted from the video data. The analyses enabled us to examine the effect of the type of noise and to show that hyper-articulation concerns lip aperture and spreading rather than lip pinching. The analysis of the relationship between acoustic and articulatory parameters shows that this speaker especially adapts to noise not only by talking louder or increasing vowel recognition cues but also by increasing spectral emergence.

On Speech Variation and Word Type Differentiation by Articulatory Feature Representations
Louis ten Bosch, R. Harald Baayen, Mirjam Ernestus; Radboud Universiteit Nijmegen, The Netherlands
Thu1A3O.2
This paper describes ongoing research aiming at the description of variation in speech as represented by asynchronous articulatory features. We will first illustrate how distances in the articulatory feature space can be used for event detection along speech trajectories in this space. The temporal structure imposed by the cosine distance in articulatory feature space coincides to a large extent with the manual segmentations on phoneme level. The analysis also indicates that the articulatory feature representation provides better such alignments than the MFCC representation does. Secondly, we will present first results that indicate that articulatory features can be used to probe for acoustic differences in the onsets of Dutch singulars and plurals.

A Study of Emotional Speech Articulation Using a Fast Magnetic Resonance Imaging Technique
Sungbok Lee, Erik Bresch, Jason Adams, Abe Kazemzadeh, Shrikanth Narayanan; University of Southern California, USA
Thu1A3O.3
This study aims at characterizing the acoustic and articulatory modifications that occur in speech in noisy environments, and at examining them as compensatory strategies. Audio, EGG and video signals were recorded for a female native speaker of French. The corpus consisted of short sentences with a subject-verb-object (SVO) structure. The sentences were recorded in three conditions: silence, 85dB white noise, and 85dB cocktail party noise. Labial parameters were extracted from the video data. The analyses enabled us to examine the effect of the type of noise and to show that hyper-articulation concerns lip aperture and spreading rather than lip pinching. The analysis of the relationship between acoustic and articulatory parameters shows that this speaker especially adapts to noise not only by talking louder or increasing vowel recognition cues but also by increasing spectral emergence.

Voice Source Correlates of Prosodic Features in American English: A Pilot Study
Markus Iseli, Yen-Liang Shue, Melissa A. Epstein, Patricia Keating, Jody Kreiman, Abeer Alwan; University of California at Los Angeles, USA
Thu1A3O.1
In this paper, we examine the dependencies of voice source parameters \( F_0 \) (fundamental frequency), \( E_L \) (maximal glottal flow change), \( R_k \) (glottal symmetry/skew), \( L_i \) (value related to source spectral tilt) and \( H_1^* - H_2^* \) (difference of formant-corrected magnitudes of the first two source spectral harmonics) on prosodic features such as pitch accents, stress, and sentence type and the interdependencies of some of these measures. A small, carefully designed corpus containing a sentence in different prosodic configurations was used in this study. Statistical analysis was performed using two-way ANOVAs to test for the voice source parameter dependencies. Results show that \( F_0 \) is positively correlated with \( E_L \) and \( L_i \), and negatively correlated with \( H_1^* - H_2^* \). Stressed syllables showed lower values of \( R_k \) and \( H_1^* - H_2^* \) compared to stressless syllables. The effect of pitch accent can be seen as a combination of its \( F_0 \), and stress. Phrase-final syllables for interrogative sentences yielded a higher \( F_0 \) and lower \( R_k \) and \( H_1^* - H_2^* \) compared to declarative sentences. It was found that it is important to differentiate between tunes when analyzing prosodic features that involve tunes, such as pitch accent and probably boundary.
Thu1BuP 1: Voice Morphing
Poster Session, Butler Room, 10:30 – 12:30
Chair: Isabel Trancoso, L’F INESC-ID/IST, Portugal

Improving the Performance of HMM-Based Voice Conversion Using Context Clustering Decision Tree and Appropriate Regression Matrix Format
Long Qin, Yi-Jian Wu, Zhen-Hua Ling, Ren-Hua Wang; University of Science & Technology of China, China
Thu1BuP.1
To improve the performance of the HMM-based voice conversion system in which the LSP coefficient is introduced as the spectral representation, a model clustering technique to tie HMMs into classes for the model adaptation, considering the phonetic and linguistic contextual factors of HMMs, is adopted in this paper. Besides, due to the relationship between the LSP coefficients of adjacent orders, an appropriate format of the regression matrix is suggested according to the small amount of the adaptation training data. Subjective and objective tests prove that the source HMMs can be adapted more accurately using the proposed method, meanwhile the synthetic speech generated from the adapted model has better discrimination and speech quality.

Map-Based Adaptation for Speech Conversion Using Adaptation Data Selection and Non-Parallel Training
Chung-Han Lee, Chung-Hsien Wu; National Cheng Kung University, Taiwan
Thu1BuP.2
This study presents an approach to GMM-based speech conversion using maximum a posteriori probability (MAP) adaptation. First, a conversion function is trained using a parallel corpus containing the same utterances spoken by both the source and the reference speakers. Then a non-parallel corpus from a new target speaker is used for the adaptation of the conversion function which models the voice conversion between the source speaker and the new target speaker. The consistency among the adaptation data is estimated to select suitable data from the non-parallel corpus for MAP-based adaptation of the GMMs. In speech conversion evaluation, experimental results show that MAP adaptation using a small non-parallel corpus can reduce the conversion error and improve the speech quality for speaker identification compared to the method without adaptation. Objective and subjective tests also confirm the promising performance of the proposed approach.

Novel Method for Data Clustering and Mode Selection with Application in Voice Conversion
Jani Nurminen, Jilei Tian, Victor Popa; Nokia Research Center, Finland
Thu1BuP.3
Since the statistical properties of speech signals are variable and depend heavily on the content, it is hard to design speech processing techniques that would perform well on all inputs. For example, in voice conversion, where the aim is to transform the speech uttered by a source speaker to sound as if it was spoken by a target speaker, different types of inter-speaker relationships can be found from different types of speech segments. To tackle this problem in a robust manner, we have developed a novel scheme for data clustering and mode selection. When applied in the voice conversion application, the main idea of the proposed approach is to first cluster the target data to achieve a minimized intra-cluster variability. Then, a mode selector or a classifier is trained on aligned source-related data to recognize the target-based clusters. Auxiliary speech features can be used to enhance the classification accuracy, in addition to the source data. Finally, a separate conversion scheme is trained and used for each cluster. The proposed scheme is fully data-driven and it avoids the need to use heuristic solutions. The superior performance of the proposed scheme has been verified in a practical voice conversion system.

Text-Independent Cross-Language Voice Conversion
David Sändigmann 1, Harald Höge 1, Antonio Bonafonte 2, Hermann N ey 3, Julia Hirschberg 4; Siemens AG, Germany; 2 Universitat Politecnica de Catalunya, Spain; 3 RWTH Aachen University, Germany; 4 Columbia University, USA
Thu1BuP.4
So far, cross-language voice conversion requires at least one bilingual speaker and parallel speech data to perform the training. This paper shows how these obstacles can be overcome by means of a recently presented text-independent training method based on unit selection. The new method is evaluated in the framework of the European speech-to-speech translation project TC-Star and achieves a performance similar to that of text-dependent intra-lingual voice conversion.

Maximum Likelihood Voice Conversion Based on GMM with STRAIGHT Mixed Excitation
Yamato Ohtani, Tomoki Toda, Hiroshi Saruwatari, Kiyohi Shikano; Nara Institute of Science & Technology, Japan
Thu1BuP.5
The performance of voice conversion has been considerably improved through statistical modeling of spectral sequences. However, the converted speech still contains traces of artificial sounds. To alleviate this, it is necessary to statistically model a source sequence as well as a spectral sequence. In this paper, we introduce STRAIGHT mixed excitation to a framework of the voice conversion based on a Gaussian Mixture Model (GMM) on joint probability density of source and target features. We convert both spectral and source feature sequences based on Maximum Likelihood Estimation (MLE). Objective and subjective evaluation results demonstrate that the proposed source conversion produces strong improvements in both the converted speech quality and the conversion accuracy for speaker individuality.

Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion
Mikihiro Nakagiri, Tomoki Toda, Hideki Kashioka, Kiyohiro Shikano; Nara Institute of Science & Technology, Japan
Thu1BuP.6
The conversion method from Non-Audible Murmur (NAM) to ordinary speech based on the statistical voice conversion (NAM-to-Speech) has been proposed towards realization of “silent speech telephone.” Although NAM-to-Speech converts NAM to intelligible voices with similar quality to speech, there is still a large problem, i.e., difficulties of the $F_0$ estimation from unvoiced speech. In order to avoid this problem, we propose a conversion method from NAM to whisper that is a familiar and intelligible unvoiced speech (NAM-to-Whisper). Moreover, we enhance NAM-to-Whisper so that multiple types of body-transmitted unvoiced speech such as NAM and Body Transmitted Whisper (BTW) are accepted as input voices. We evaluate the performance of the proposed conversion method. Experimental results demonstrate that 1) intelligibility and naturalness of NAM are significantly improved by NAM-to-Whisper, 2) NAM-to-Whisper outperforms NAM-to-Speech, and 3) we can train a single conversion model successfully converting both NAM and BTW to the target voice.
An HMM-Based Singing Voice Synthesis System
Keiichi Saio, Heiga Zen, Yoshikiko Nankaku, Akinobu Lee, Keiichi Tokuda; Nagoya Institute of Technology, Japan
Thu1BuP.7
The present paper describes a corpus-based singing voice synthesis system based on hidden Markov models (HMMs). This system employs the HMM-based speech synthesis to synthesize singing voice. Musical information such as lyrics, tones, durations is modeled simultaneously in a unified framework of the context-dependent HMM. It can mimic the voice quality and singing style of the original singer. Results of a singing voice synthesis experiment show that the proposed system can synthesize smooth and natural-sounding singing voice.

Voice Conversion Based on Mixtures of Factor Analyzers
Yosuke Uto¹, Yoshikiko Nankaku¹, Tomoki Toda², Akinobu Lee¹, Keiichi Tokuda¹; ¹Nagoya Institute of Technology, Japan; ²Nara Institute of Science & Technology, Japan
Thu1BuP.8
This paper describes the voice conversion based on the Mixtures of Factor Analyzers (MFA) which can provide an efficient modeling with a limited amount of training data. As a typical spectral conversion method, a mapping algorithm based on the Gaussian Mixture Model (GMM) has been proposed. In this method two kinds of covariance matrix structures are often used: the diagonal and full covariance matrices. GMM with diagonal covariance matrices requires a large number of mixture components for accurately estimating spectral features. On the other hand, GMM with full covariance matrices needs sufficient training data to estimate model parameters. In order to cope with these problems, we apply MFA to voice conversion. MFA can be regarded as intermediate model between GMM with diagonal covariance and with full covariance. Experimental results show that MFA can improve the conversion accuracy compared with the conventional GMM.

Efficient Gaussian Mixture Model Evaluation in Voice Conversion
Jilei Tian, Jani Nurminen, Victor Popa; Nokia Research Center, Finland
Thu1BuP.9
Voice conversion refers to the adaptation of the characteristics of a source speaker's voice to those of a target speaker. Gaussian mixture model (GMM) based voice conversion has been found to be efficient in the voice conversion task. The GMM parameters are estimated from a training set with the goal to minimize the mean square error (MSE) between the transformed and target vectors. Obviously, the quality of the GMM model plays an important role in achieving better voice conversion quality. This paper presents a very efficient approach for the evaluation of GMM models directly from the model parameters without using any test data, facilitating the improvement of the transformation performance especially in the case of embedded implementations. Though the proposed approach can be used in any application that utilizes GMM based transformation, we take voice conversion as an example application throughout the paper. The proposed approach is experimented with in this context and evaluated against an MSE based evaluation method. The results show that the proposed method is in line with all subjective observations and MSE results.

Constrained Structural Maximum a posteriori Linear Regression for Average-Voice-Based Speech Synthesis
Yuji Nakano, Makoto Tachibana, Junichi Yamagishi, Takao Kobayashi; Tokyo Institute of Technology, Japan
Thu1BuP.10
This paper proposes a constrained structural maximum a posteriori linear regression (CSMAPLR) algorithm for further improvement of speaker adaptation performance in HMM-based speech synthesis. In the algorithm, the concept of structural maximum a posteriori (SMAP) adaptation is applied to estimation of transformation matrices of the constrained MLLR (CMLLR), where recursive MAP-based estimation of the transformation matrices from the root node to lower nodes of context decision tree is conducted. We incorporate the algorithm into HSMM-based speech synthesis system and show that CSMAPLR adaptation utilizes both of the advantage of CMLLR and SMAPLR adaptation from the result of objective evaluation test. We also show that CSMAPLR adaptation provides more similar synthetic speech to the target speaker than CMLLR and SMAPLR adaptation from the result of subjective evaluation test.

Frequency Warping Based on Mapping Formant Parameters
Zhi-Wei Shuang¹, Raimo Bakis², Slava Shechtman³, Dan Chazan³, Yong Qin¹; ¹IBM China Research Lab, China; ²IBM T.J. Watson Research Center, USA; ³IBM Haifa Research Lab, Israel
Thu1BuP.11
We propose a novel method of generating a frequency warping function by mapping formant parameters of the source speaker and the target speaker. Alignment and selection process are performed to ensure that the mapping formants can represent speakers' difference well. This approach requires only a very small amount of training data for generating the warping function, which can greatly facilitate its application. It can also achieve high quality of the converted speech with successfully converting a speaker's identity. A practical voice morphing system has been built based on this approach. And experimental results show its effectiveness.

Automatic Phonetic Segmentation by Using a SPM-Based Approach for a Mandarin Singing Voice Corpus
Cheng-Yuan Lin, J.-S. Roger Jang; National Tsing Hua University, Taiwan
Thu1BuP.12
This paper proposes a score predictive model (SPM) based approach to integrate two segmentation results obtained by HMM and DTW for a Mandarin singing voice corpus. The SPM can predict the score of a boundary according to its corresponding 14 dimensional feature vector. In order to verify the performance of the proposed method, several experiments were performed. The experimental results demonstrate the feasibility of the proposed approach.

A Comparison of Singing Evaluation Algorithms
Partha Lal; University of Edinburgh, UK
Thu1BuP.13
This paper describes a system that compares user renditions of short sung clips with the original version of those clips. The F0 of both recordings was estimated and then Viterbi-aligned with each other. The total difference in pitch after alignment was used as a distance metric and transformed into a rating out of ten, to indicate the user how close he or she was to the original singer.

An existing corpus of sung speech was used for initial design and optimisation of the system. We then collected further development and evaluation corpora — these recordings were judged for closeness to an original recording by two human judges. The rankings assigned by those judges were used to design and optimise the system. The design was then implemented and deployed as part of a telephone-based entertainment application.
Improving Speech Recognition of Two Simultaneous Speech Signals by Integrating ICA BSS and Automatic Missing Feature Mask Generation

Ryu Takeda, Shun’ichi Yamamoto, Kazunori Komatani, Tetsuya Ogata, Hiroshi G. Okuno; Kyoto University, Japan

Thu1CaP.1

Robot audition systems require capabilities for sound source separation and the recognition of separated sounds, since we hear a mixture of sounds in our daily lives, especially mixed of speech. We report a robot audition system with a pair of omni-directional microphones embedded in a humanoid that recognizes two simultaneous talkers. It first separates the sound sources by Independent Component Analysis (ICA) with the single-input multiple-output (SIMO) model. Then, spectral distortion in the separated sounds is then estimated to generate missing feature masks. Finally, the separated sounds are recognized by missing-feature theory (MFT) for Automatic Speech Recognition (ASR). The novel aspects of our system involve estimates of spectral distortion in the temporal-frequency domain in terms of feature vectors and based on estimates error in SIMO-ICA signals. The resulting system outperformed the baseline robot audition system by 7%.

Missing-Feature Reconstruction for Band-Limited Speech Recognition in Spoken Document Retrieval

Wooil Kim, John H.L. Hansen; University of Texas at Dallas, USA

Thu1CaP.2

In spoken document retrieval, it is necessary to support a variety of audio corpora from sources that have a range of conditions (e.g., channels, microphones, noise conditions, recording media, etc.). Varying band-limited speech represents one of the most challenging factors for robust speech recognition. The missing-feature reconstruction method shows the effectiveness in recognition of the speech corrupted by additive noise. However, it has a problem when applied to the band-limited speech reconstruction, since it assumes that the observations in the unreliable regions are always greater than the latent original clean speech. In this study, we propose to modify the current way to calculate the marginal probability for reconstruction into the computation depending only on the reliable components. To detect the cut-off regions from incoming speech, the blind mask estimation scheme is proposed, which employs the synthesized band-limited speech model without training database. Experimental results on Aurora 2.0 and actual band-limited speech (NGSW corpus) indicate that the proposed method is effective in improving recognition accuracy of the band-limited speech. Through combining with an adaptation method, 22.17% of relative improvement is obtained on NGSW.

Incremental Learning of MAP Context-Dependent Edit Operations for Spoken Phone Number Recognition in an Embedded Platform

Hahn Koo¹, Yan Ming Cheng²;¹ University of Illinois at Urbana Champaign, USA, ²Motorola Labs, USA

Thu1CaP.3

Error-corrective post-processing (ECPP) has great potential to reduce speech recognition errors beyond that obtained by speech model improvement. ECPP approaches aim to learn error-corrective rules to directly reduce speech recognition errors. This paper presents our investigation into one such approach, incremental learning of maximum a posteriori (MAP) context-dependent edit operations. Limiting our dataset to spoken telephone number recognition output, we have evaluated this approach in an automotive environment using an embedded speech recognizer in a mobile device. We have found that a reduction of approximately 44–49% in speech recognition string errors can be achieved after learning.

Development and Evaluation of Speech Database in Automotive Environments for Practical Speech Recognition Systems

Yasunari Obuchi, Nobuo Hataoka; Hitachi Ltd., Japan

Thu1CaP.4

Aiming at practical speech recognition systems, we are developing speech databases representing the situation in which the application is used, and evaluating various techniques using the database. Such methodology is expected to contribute to bridge the expectations of the developers and the reactions of the users. We start with the applications in automotive environments, or car navigation systems more precisely. During the data collection, special attention was paid to maintain the spontaneousness of the speaker, to cover failed utterances, and to use the hardware setup suitable for microphone array techniques. After the database is prepared, various techniques are evaluated. In some cases, oracle information is used to find the upper limit of the improvement of a specific module. In other cases, typical improving algorithms are tested. Recognition experiments using two separate decoders indicate that endpoint detection, feature normalization, speaker adaptation, and parallel decoding are promising fields. We also present some modifications of parallel decoding to reduce the computational cost and to realize practical applications.

An Effective and Efficient Utterance Verification Technology Using Word N-Gram Filler Models

Dong Yu, Yun-Cheng Ju, Alex Acero; Microsoft Research, USA

Thu1CaP.5

In this paper we propose a novel, effective, and efficient utterance verification (UV) technology for access control in the interactive voice response (IVR) systems. The key of our approach is to construct a context-free grammar by using the secret answer to a question and a word N-gram based filler model. The N-gram filler provides rich alternatives to the secret answer and can potentially improve the accuracy of the UV task. It can also absorb carrier words used by callers and thus can improve the robustness. We also propose using a predictor based on the best alternative to calculate the confidence. We show detailed experimental results on a tough UV test set that contains 930 positive and 930 negative cases and discuss types of questions that are suitable for the UV task. We demonstrate that our approach can achieve a 2.14% equal error rate (EER) on average and 0.8% false accept rate if the false reject rate is 2.6% and above. This is a 49% EER reduction compared with the approaches using acoustic fillers, and a 72% EER reduction compared with the posterior probability based confidence measurement.

An Effective Bispectrum Phase Entropy-Based Algorithm for VAD

J.M. Górriz, Javier Ramírez, C.G. Puntonet, José C. Segura; Universidad de Granada, Spain

Thu1CaP.6

In this paper we propose a novel Voice Activity Detection (VAD) algorithm, based on the integrated bispectrum function (IBI), for improving Automated Speech Recognition (ASR) systems that work in noisy environments. In particular we use the combination of two features, IBI magnitude and IBI phase to formulate a robust and smoothed decision rule for speech/pause discrimination. The analysis performed on the new combined feature highlighted: i) the advantages of each individual feature, while compensating the drawback of each other, and ii) the higher ability for endpoint detection given by a lower variance of the decision function in pause/speech frames. The experiments conducted on the Spanish SpeechDat-Car database showed...
that the proposed algorithm outperforms ITU G.729, ETSI AMR1 and AMR2 and ETSI AFE standards as well as other recently reported VAD methods in speech/non-speech detection performance.

**Two-Step Unsupervised Speaker Adaptation Based on Speaker and Gender Recognition and HMM Combination**

Petr Cerva, Jan Nouza, Jan Silovsky; Technical University of Liberec, Czech Republic

Thu1Cap.7

In this paper, we present a new strategy for unsupervised speaker adaptation. In our approach, the adaptation is performed in two steps for each test utterance. In the first online step, we utilize speaker and gender identification, a set of speaker dependent (SD) hidden Markov models (HMMs) and our own fast linear model combination approach to create a proper model for the first speech recognition pass. After that the recognized phonetic transcription of the utterance is used for maximum likelihood (ML) estimation of more accurate weights for the final model combination step. Our experimental results on different types of broadcast programs show that the proposed method is capable to reduce the word error rate (WER) relatively by more than 17%.

**CENSREC2: Corpus and Evaluation Environments for In Car Continuous Digit Speech Recognition**

Satoshi Nakamura¹, Masakiyo Fujimoto², Kazuya Takeda³; ¹NICT, Japan; ²NTT Corporation, Japan; ³Nagoya University, Japan

Thu1Cap.8

This paper introduces a common database and an evaluation framework for connected digit speech recognition in real driving car environments, CENSREC-2, as an outcome of IPSJ-SIG SLP Noisy Speech Recognition Evaluation Working Group. Speech data of CENSREC-2 was collected using two microphones, a close-talking microphone and a hands-free microphone, under three car speeds and four car conditions. CENSREC-2 provides four evaluation environments which are designed using speech data collected in these car conditions.

**Detection of Word Fragments in Mandarin Telephone Conversation**

Cheng-Tao Chu, Yun-Hsuan Sung, Yuan Zhao, Daniel Jurafsky; Stanford University, USA

Thu1Cap.9

We describe preliminary work on the detection of word fragments in Mandarin conversational telephone speech. We extracted prosodic, voice quality, and lexical features, and trained Decision Tree and SVM classifiers. Previous research shows that glottalization features are instrumental in English fragment detection. However, we show that Mandarin fragments are quite different than English; 90% of Mandarin fragments are followed immediately by a repetition of the fragmentary word. These repetition fragments are not glottalized, and they have a very specific distribution; the 12 most frequent words (‘you’, ‘I’, ‘that’, ‘have’, ‘then’, etc.) cover 50% of the tokens of these fragments. Thus rather than glottalization, we found the most useful feature for Mandarin fragment detection was the identity of the neighboring character (word or morpheme). In an oracle experiment using the true (reference) neighboring words as well as prosodic and voice quality features, we achieved 80% accuracy in Mandarin fragment detection.

**A DTW-Based Dissimilarity Measure for Left-to-Right Hidden Markov Models and Its Application to Word Confusability Analysis**

Qiang Hao, Wei Li; University of Hong Kong, China

Thu1Cap.10

We propose a dynamic time-warping (DTW) based distortion measure for measuring the dissimilarity between pairs of left-to-right continuous density hidden Markov models with state observation densities being mixture of Gaussians. The local distortion score required in DTW is defined as an approximate Kullback-Leibler divergence (KLD) between two Gaussian mixture models (GMMs). Several approximate KLDs are studied and compared for pairs of GMMs with different properties, and one of them is identified for being used in our DTW-based HMM dissimilarity measure. In an experiment of identifying automatically the subsets of confusable Putonghua (Mandarin Chinese) syllables, it is observed that the result based on the proposed HMM dissimilarity measure is highly consistent with the one based on syllable recognition confusion matrix obtained on a testing data set.

**Multi-Flow Block Interleaving Applied to Distributed Speech Recognition Over IP Networks**

Angel M. Gómez, Juan J. Ramos-Muñoz, Antonio M. Peinado, Victoria Sánchez; Universidad de Granada, Spain

Thu1Cap.11

Interleaving has shown to be a useful technique to provide robust distributed speech recognition over IP networks. This is due to its ability to disperse consecutive losses. However, this ability is related to the delay introduced by the interleaver. In this work, we propose a novel multi-flow block interleaver which exploits the presence of several streams and allows to reduce the involved delay. Experimental results have shown that this interleaver approximates the performance of end-to-end interleavers but with a fraction of their delay. As disadvantage, this interleaver must be placed in a common node where more than one flow are available.

**Moving Speech Recognition from Software to Silicon: The In Silico Vox Project**

Edward C. Lin, Kai Yu, Rob A. Rutenbar, Tsuhan Chen; Carnegie Mellon University, USA

Thu1Cap.12

To achieve much faster decoding, or much lower power consumption, we need to liberate speech recognition from the artificial constraints of its current software-only form, and move the essential computations directly into silicon. There are vast efficiencies waiting to be unlocked in this application — we need the proper architecture to do so. We report results from a first-generation hardware architecture simulated at bit-level, and a complete, working FPGA-based prototype. Simulation results show that rather modest hardware designs, running 10-20X slower than conventional processors, can already decode at 0.6 xRT, running the standard 5K Wall Street Journal benchmark.

**A Study on Detection Based Automatic Speech Recognition**

Chengyuan Ma, Yu Tsao, Chin-Hui Lee; Georgia Institute of Technology, USA

Thu1Cap.13

We propose a new approach to automatic speech recognition based on word detection and knowledge-based verification. Given an utterance, we first design a collection of word detectors, one for each lexical item in the vocabulary. Some pruning strategies are used to eliminate unlikely word candidates. Then these detected words are combined into word strings. The proposed approach is different from the conventional maximum a posteriori decoding method, and it is a critical component in building a bottom-up, detection-based speech recognition system in which knowledge in acoustics, speech and language can easily be incorporated into pruning unlikely word hypotheses and rescoring. The proposed approach was evaluated on a connected digit task using phone models trained from the TIMIT corpus. When compared with state-of-the-art connected digit recognition algorithms, we found the proposed detection based framework works well even no digit samples were used for training the detectors and recognizers. Other knowledge based constraints, such as manner and place of
artificial detectors, can be incorporated into this detection-based approach to improve the robustness and performance of the overall system.

**Novel Time Domain Multi-Class SVMs for Landmark Detection**

Rahul Chitturi\(^1\), Mark Hasegawa Johnson\(^2\); \(^1\)University of Texas at Dallas, USA; \(^2\)University of Illinois at Urbana Champaign, USA

Thu1CaP.14

The training of precise speech recognition models depends on accurate segmentation of the phonemes in a training corpus. Segmentation is typically performed using HMMs, but recent speech recognition work suggests that the transient acoustic features characteristic of manner-class phoneme boundaries (landmarks) may be more precisely localized using acoustic classifiers specifically designed for the task of landmark detection. This paper makes an empirical exploration of new features which suit Landmark Detection and the application of Multi-class SVMs that are capable of improving the time alignment of phoneme boundaries proposed by Binary SVMs and HMM-based speech recognizer. On a standard benchmark data set (A database of Telugu — Official Indian Language, spoken by 75 million people), we achieve a new state-of-the-art performance, reducing RMS phone boundary alignment error from 32ms to 22ms.

**Towards Automatic Parameter Extraction of Command-Response Model for Cantonese**

Raymond W.M. Ng\(^1\), Tan Lee\(^1\), Wentao Gu\(^2\); \(^1\)Chinese University of Hong Kong, China; \(^2\)University of Tokyo, Japan

Thu1FoP.1

Command-response model is one of the quantitative models capable of illustrating the voice fundamental frequency (F0) characteristics of different languages. In recent years, the model has been adapted to tone languages including Cantonese. This paper presents an automatic optimization procedure to enhance parameter extraction of command-response model for short phrases of Cantonese. We conduct a parameter extraction test on 128 speech segments produced by three speakers, all of which are modeled by single phrase command in the command-response model. The extracted parameters attained accuracy compatible with that obtained from the conventional manual estimation.

**A Model for the f0 Reset in Corpus-Based Intonation Approaches**

Francisco Campillo\(^1\), Jan P.H. van Santen\(^2\), Eduardo R. Banga\(^1\); \(^1\)University of Vigo, Spain; \(^2\)Oregon Health & Science University, USA

Thu1FoP.2

Concatenative intonation systems model the intonation contours as the concatenation of small natural units extracted from suitable contexts. The special characteristics of this type of models make it difficult to include some factors whose effect overcomes the phonemic group domain. In this paper one of those factors, the f0 reset related to the sentence internal pauses, is addressed.

**Generating German Intonation with a Trainable Prosodic Model**

Gérard Bailly, Jan Gorisch; ICP, France

Thu1FoP.3

A trainable prosodic model called SFC (Superposition of Functional Contours), proposed by Holm and Bailly, is here confronted to German intonation. Training material is the publicly available Siemens Synthesis Corpus that provides spoken utterances for high-quality speech synthesis. We describe the labeling framework and first evaluation results that compares the original prosody of test sentences of this corpus with their prosodic rendering by the proposed model and state-of-the-art systems available on-line on the web.

**Incorporating Second-Order Information into Two-Step Major Phrase Break Prediction for Korean**

Seungwon Kim\(^1\), Jinsik Lee\(^1\), Byeongchang Kim\(^2\), Gary Geunbae Lee\(^1\); \(^1\)POSTECH, Korea; \(^2\)Catholic University of Daegu, Korea

Thu1FoP.4

In this paper, we present a new phrase break prediction method that integrates second-order information into general maximum entropy model. The phrase break prediction problem was mapped into a classification problem in our research. The features we used for the prediction of phrase breaks are of several layers such as local features (part-of-speech (POS) tags, a lexicon, lengths of eojeos and location of juncture in the sentence), global features (chunk label derived from a eojeo parse tree) and second-order features (distance probability of previous and next phrase break). These three features were combined and used in the experiments, and we were able to generate good performance especially in the major phrase break prediction.

**Totally Data-Driven Duration Modeling Based on Generalized Linear Model for Mandarin TTS**

Lifu Yi, Jian Li, Xiaoyan Lou, Jie Hao; Toshiba China R&D Center, China

Thu1FoP.5

This paper proposes a totally data-driven duration modeling method for Mandarin TTS, which uses Generalized Linear Models (GLM) to model duration and stepwise regression to automatically select the attribute set with statistical measurements. This method can get a good tradeoff between model complexity and goodness of fit. Besides, speaking rate is introduced as a new modeling attribute, which not only achieves higher performance but also provides a novel approach to adjust speaking rate when synthesizing. We also propose to use \( R^2 \) to fairly evaluate the modeling performances on different databases, since \( R^2 \) refers to the fraction of corresponding variance explained by a model. Experiments show the performance of GLM is significantly higher than that of CART. With our much smaller models and corpus, the proposed method also achieves comparable results reported by other excellent researches.

**Segmental Duration Modeling in Turkish**

Ozlem Ozturk\(^1\), Tolga Ciloglu\(^2\); \(^1\)Dokuz Eylul University, Turkey; \(^2\)Middle East Technical University, Turkey

Thu1FoP.6

Naturalness of synthetic speech highly depends on appropriate modeling of prosodic aspects. Mostly, three prosody components are modeled: segmental duration, pitch contour and intensity. In this study, we present our work on modeling segmental duration in Turkish using machine-learning algorithms, especially Classification and Regression Trees (CART). The models predict phone durations based on attributes such as phone identity, neighboring phone identities, lexical stress, position of syllable in word, part-of-speech (POS) information, word length in number of syllables and position of word in utterance extracted from a speech corpus of approximately 700 sentences. Obtained models predict segment durations better than mean duration approximations (∼0.77 Correlation Coefficient, CC, and 20.4 ms Root-Mean Squared Error, RMSE). Attributes phone identity, neighboring phone identities, lexical stress, syllable type, POS, phrase break information, and location of word in the phrase constitute best predictor set for phoneme duration modeling.
Lexical Stress in Continuous Speech Recognition
Rogier C. van Dalen, Pascal Wiggers, Léon J.M. Rothkrantz; Delft University of Technology, The Netherlands
Thu1Fop.7
Human listeners use lexical stress for word segmentation and disambiguation. We look into using lexical stress for large-vocabulary speech recognition for the Dutch language. It appears that beside vowels, consonants should be taken into account. By introducing stressed phonemes, and features for spectral bands and the fundamental frequency, we reduce the word error rate by 2.6%.

Improving Tone Recognition with Combined Frequency and Amplitude Modelling
Siwei Wang, Gina-Anne Levow; University of Chicago, USA
Thu1Fop.8
To improve tone recognition in continuous speech, we propose a strategy focusing on separating regions influenced by tonal coarticulation from regions that more closely approximate canonical tone production. Given a syllable segmentation, this approach employs amplitude and pitch information to generate an improved sub-syllable segmentation and feature representation. This sub-syllable segmentation is derived from the convex hull of the amplitude-pitch plot. Our approach achieves a 15% improvement using our segmentation strategy over a simple time-only segmentation. Finally, a future extension with sequential labelling is discussed.

Latent Prosodic Modeling (LPM) for Speech with Applications in Recognizing Spontaneous Mandarin Speech with Disfluencies
Che-Kuang Lin, Lin-shan Lee; National Taiwan University, Taiwan
Thu1Fop.9
In this paper, a new approach of Latent Prosodic Modeling (LPM) for analyzing the prosody of speech is presented. Based on a set of newly defined prosodic characters, prosodic terms, documents, and the Probabilistic Latent Semantic Analysis (PLSA) framework, prosody can be modeled using a set of prosodic states representing various latent factors such as speakers, speaking rate, utterance modality, intonation behavior, etc. In terms of some probabilistic relationships with the observed prosodic features. Organizing the training data based on this new model may also produce more delicate classification models for various speech processing applications considering the prosody. In the initial application example, we showed the use of this model on the task of disfluency IP detection for spontaneous Mandarin speech recognition, and improved IP detection accuracy and speech recognition performance were obtained in the experiments.

Tone Recognition of Continuous Speech of Standard Chinese Using Neural Network and Tone Nucleus Model
Keikichi Hirose, Hui Hu, Xiaodong Wang, Nobuaki Minematsu; University of Tokyo, Japan
Thu1Fop.10
A method is developed for recognizing lexical tone types of Standard Chinese syllables in continuous speech. Neural network (four-layered perceptron) is adopted as classifier. The method includes two steps: first recognizing tone types using prosodic features of voiced part, and then re-recognizing by viewing only on tone nucleus, which is a portion of the syllable showing rather stable fundamental frequency ($F_0$) contour regardless of tone types of the preceding and following syllables. The voiced part (or tone nucleus) is divided into 20 segments, and $F_0$, delta-$F_0$, $F_0$ slope and short-term energy of each segment are served as inputs to the neural network. In order to cope with tone coarticulation, prosodic feature parameters for the last 5 segments of the preceding syllable and the initial 5 segments of the following syllable are included in the neural network inputs. Information on syllable length is also added to the inputs. Tone recognition experiment was conducted for a female speaker’s utterances included in HKU96 corpus. The average recognition rate was 86.5% including neutral tone syllables, when the tone nucleus model was not used. It increased to 86.9%, when the model was used. The obtained rate is higher by more than 3 points as compared to that obtained by the hidden-Markov-model-based tone recognizer developed by the authors formerly.

Prosodic Feature Generation for Back-Channel Prediction
Thamar Solorio, Olac Fuentes, Nigel G. Ward, Yaffa Al Bayyari; University of Texas at El Paso, USA
Thu1Fop.11
Using prosodic information to predict when back-channels are appropriate in spontaneous dialogs has become somewhat of a reference problem for automatic discovery techniques. Here we present experiments with two ideas: the use of features derived from randomly generated pitch and energy filters, and the use of instance-based learning, specifically the Locally Weighted Linear Regression (LWLR) algorithm. For the task of predicting possible back-channel locations in Arabic [6], we obtain 22% precision and 51% recall, which is as good as that obtained using a laboriously developed and hand-tuned rule.

On the Sufficiency and Redundancy of Pitch for TRP Projection
Wieneke Wesseling, R.J.J.H. van Son, Louis C.W. Pols; University of Amsterdam, The Netherlands
Thu1Fop.12
In two Reaction Times (RT) experiments, subjects were asked to respond with minimal responses to prerecorded dialogs and impoverished versions of these dialogs, containing either only intonation and pause information, hummed stimuli, or no periodic component at all, whispered stimuli. For the hummed stimuli, response delays and, especially, variances were higher than the original recordings. Responses to mid-frequency pitch utterance-ends were significantly longer than responses to low pitch utterance-ends, suggesting that our subjects fell back to reacting to pauses when presented with hummed utterances ending in a mid-frequency tone. This suggests that, in contrast to low or high end-tones, intonation contours that end in a mid-frequency tone might not contain any useful information for predicting end-of-utterance Turn Relevant Places (TRPs). We conclude that just the intonation and pauses of a conversation contain sufficient information for projection of TRPs. However this information is measurably impoverished with respect to original to an extent that increases the “processing” time by 10%. No difference was found between whispered and original speech. This lack of any effect of removing all periodic sound components from the speech signal indicates that in natural speech the pitch signal itself might be redundant for predicting TRPs.
Thu2A10 : Discriminative Training
Oral Session, Allegheny Ballroom, left, 14:00 - 16:00
Chair: Olivier Siohan, IBM T.J. Watson Research Center, USA

Hypothesis Spaces for Minimum Bayes Risk Training in Large Vocabulary Speech Recognition
Matthew Gibson, Thomas Hain; University of Sheffield, UK
Thu2A10.1
The Minimum Bayes Risk (MBR) framework has been a successful strategy for the training of hidden Markov models for large vocabulary speech recognition. Practical implementations of MBR must select an appropriate hypothesis space and loss function. The set of word sequences and a word-based Levenshtein distance may be assumed to be the optimal choice but use of phoneme-based criteria appears to be more successful. This paper compares the use of different hypothesis spaces and loss functions defined using the system constituents of word, phone, physical triphone, physical state and physical mixture component. For practical reasons the competing hypotheses are constrained by sampling. The impact of the sampling technique on the performance of MBR training is also examined.

Minimum Divergence Based Discriminative Training
Jun Du1, Peng Liu2, Frank K. Soong2, Jian-Lai Zhou2, Ren-Hua Wang1; 1 University of Science & Technology of China, China; 2 Microsoft Research Asia, China
Thu2A10.2
We propose to use Minimum Divergence (MD) as a new measure of errors in discriminative training. To focus on improving discrimination between any two given acoustic models, we refine the error definition in terms of Kullback-Leibler Divergence (KLD) between them. The new measure can be regarded as a modified version of Minimum Phone Error (MPE) but with a higher resolution than just a symbol matching based criterion. Experimental recognition results show the new MD based training yields relative word error rate reductions of 57.8% and 6.1% on TIDigits and Switchboard databases, respectively, in comparison with the ML trained baseline systems. The recognition performance of MD is also shown to be consistently better than that of MPE.

Solving Large Margin Estimation of HMMS via Semidefinite Programming
Xinwei Li, Hui Jiang; York University, Canada
Thu2A10.3
In this paper, we propose to use a new optimization method, i.e., *semidefinite programming (SDP)*, to solve large margin estimation (LME) problem of continuous density hidden Markov models (CDHMM) for speech recognition. First of all, we introduce a new constraint into the LME to guarantee the boundedness of the margin of CDHMM. Secondly, we show that the LME problem under this new constraint can be formulated as an SDP problem under some relaxation conditions and it can be solved very efficiently by using some fast optimization algorithms specially designed for SDP. The new method is evaluated in a continuous digit string recognition task by using the TIDIGITS database. Experimental results clearly demonstrate that the new SDP-based method outperforms the previously proposed optimization methods using gradient descent search in both recognition accuracy and convergence speed. With the SDP-based optimization method, the best LME models achieve 0.53% in string error rate and 0.18% in WER on the TIDIGITS task. To our best knowledge, this is the best result ever reported in this task.

Use of Incrementally Regulated Discriminative Margins in MCE Training for Speech Recognition
Dong Yu, Li Deng, Xiaodong He, Alex Acero; Microsoft Research, USA
Thu2A10.4
In this paper, we report our recent development of a novel discriminative learning technique which embeds the concept of discriminative margin into the well established minimum classification error (MCE) method. The idea is to impose an incrementally adjusted “margin” in the loss function of MCE algorithm so that not only error rates are minimized but also discrimination “robustness” between training and test sets is maintained. Experimental evaluation shows that the use of the margin improves a state-of-the-art MCE method by reducing 17% digit errors and 19% string errors in the TIDigits recognition task. The string error rate of 0.55% and digit error rate of 0.19% we have obtained are the best-ever results reported on this task in the literature.

Soft Margin Estimation of Hidden Markov Model Parameters
Jinyu Li, Ming Yuan, Chin-Hui Lee; Georgia Institute of Technology, USA
Thu2A10.5
We propose a new discriminative learning framework, called soft margin estimation (SME), for estimating parameters of continuous density hidden Markov models. The proposed method makes direct usage of the successful ideas of soft margin in support vector machines to improve generalization capability, and of decision feedback learning in minimum classification error training to enhance model separation in classifier design. We attempt to incorporate frame selection, utterance selection and discriminative separation in a single unified objective function that can be optimized with the well-known generalized probabilistic descent algorithm. We demonstrate the advantage of SME in theory and practice over other state-of-the-art techniques. Tested on a connected digit recognition task, the proposed SME approach achieves a string accuracy of 99.33%. To our knowledge, this is the best result ever reported on the TIDIGITS database.

Discriminative Models for Spoken Language Understanding
Ye-Yi Wang, Alex Acero; Microsoft Research, USA
Thu2A10.6
This paper studies several discriminative models for spoken language understanding (SLU). While all of them fall into the conditional model framework, different optimization criteria lead to conditional random fields, perceptron, minimum classification error and large margin models. The paper discusses the relationship amongst these models and compares them in terms of accuracy, training speed and robustness.
Thu2A3O : Speech Synthesis
Oral Session, Allegheny Ballroom, right, 14:00 – 16:00
Chair: Paul Taylor, University of Cambridge, UK

Evaluating a Virtual Speech Cuer
G. Gibert, Gérard Bailly, F. Elisei; ICP, France
Thu2A3O.1
This paper presents the virtual speech cuer built in the context of the ARTUS project aiming at watermarking hand and face gestures of a virtual animated agent in a broadcasted audiovisual sequence. For deaf televiewers that master cued speech, the animated agent can be then superimposed — on demand and at the reception — on the original broadcast as an alternative to subtitling. The paper presents the multimodal text-to-speech synthesis system and the first evaluation performed by deaf users.

Intelligibility of Machine Translation Output in Speech Synthesis
Laura Mayfield Tomokiyo, Kay Peterson, Alan W. Black, Kevin A. Lenz; Cepstral LLC, USA
Thu2A3O.2
One use of text-to-speech synthesis (TTS) is as a component of speech-to-speech translation systems. The output of automatic machine translation (MT) can vary widely in quality, however. A synthetic voice that is extremely intelligible on naturally-occurring text may be far less intelligible when asked to render text that is automatically generated. In this paper, we compare the quality of synthesis of naturally-occurring text and its MT counterpart. We find that intelligibility of TTS on MT output is significantly lower than on either naturally-occurring text or semantically unpredictable sentences, and explore the reasons why.

A Technique for Controlling Voice Quality of Synthetic Speech Using Multiple Regression HSMM
Makoto Tachibana, Takashi Nose, Junichim Yamagishi, Takao Kobayashi; Tokyo Institute of Technology, Japan
Thu2A3O.3
This paper describes a technique for controlling voice quality of synthetic speech using multiple regression hidden semi-Markov model (HSMM). In the technique, we assume that the mean vectors of output and state duration distribution of HSMM are modeled by multiple regression with a parameter vector called voice quality control vector. We first choose three features for controlling voice qualities, that is, “smooth voice – nonsmooth voice,” “warm – cold,” “high-pitched – low-pitched,” and then we attempt to control voice quality of synthetic speech for these features. From the results of several subjective tests, we show that the proposed technique can change these features of voice quality intuitively.

Learning from Errors in Grapheme-to-Phoneme Conversion
Tatyana Polyakova, Antonio Bonafonte; Universitat Politècnica de Catalunya, Spain
Thu2A3O.4
In speech technology it is very important to have a system capable of accurately performing grapheme-to-phoneme (G2P) conversion, which is not an easy task especially if talking about languages like English where there is no obvious letter-phone correspondence. Manual rules so widely used before are now leaving the way open for the machine learning techniques and language independent tools.

In this paper we present an extension of the use of transformation-based error-driven algorithm to G2P task. A set of explicit rules was inferred to correct the pronunciation for U.S. English, Spanish and Catalan using well-known machine-learning techniques in combination with transformation based algorithm. All methods applied in combination with transformation rules significantly outperform the results obtained by these methods alone.

Eigenvoice Conversion Based on Gaussian Mixture Model
Tomoki Toda, Yamato Ohtani, Kiyohiro Shikano; Nara Institute of Science & Technology, Japan
Thu2A3O.5
This paper describes a novel framework of voice conversion (VC). We call it eigenvoice conversion (EVC). We apply EVC to the conversion from a source speaker’s voice to arbitrary target speakers’ voices. Using multiple parallel data sets consisting of utterance-pairs of the source and multiple pre-stored target speakers, a canonical eigenvoice GMM (EV-GMM) is trained in advance. That conversion model enables us to flexibly control the speaker individuality of the converted speech by manually setting weight parameters. In addition, the optimum weight set for a specific target speaker is estimated using only speech data of the target speaker without any linguistic restrictions. We evaluate the performance of EVC by a spectral distortion measure. Experimental results demonstrate that EVC works very well even if we use only a few utterances of the target speaker for the weight estimation.

Generating Time-Constrained Audio Presentations of Structured Information
Brian Langner, Rohit Kumar, Arthur Chan, Lingyun Gu, Alan W. Black; Carnegie Mellon University, USA
Thu2A3O.6
Presenting complex information in an understandable manner using speech is a challenging task to do well. Significant limitations, both in the generation process and from the human listeners’ capabilities, typically make for poorly understood speech. This work examines possible strategies for producing understandable spoken complex information working within those limitations, as well as identifying ways to improve systems to reduce the limitations' impact. We discuss a simple user study that explores these strategies with complex structured information, and describe a spoken dialog system that will make use of this work to provide a speech interface to structured information in a more understandable manner.
**Thu2WeO : Multimodal Processing**

**Oral Session, Westmoreland Room, 14:00 – 16:00**

**Chair: Jürgen Fritsch, Multimodal Technologies Inc., USA**

**Multimodal Authentication Using Qualitative Support Vector Machines**

*F. Alsaaed, A. Arriyaeinina, L. Meng, A. Malegaonkar; University of Hertfordshire, UK*

*Thu2WeO.1*

This paper proposes an approach to enhancing the accuracy of multimodal biometrics in uncontrolled environments. Variation in operating conditions results in mismatch between the training and test material, and thereby affects the biometric authentication performance regardless of this being unimodal or multimodal. The paper proposes a technique to reduce the effects of such variations in multimodal fusion. The proposed technique is based on estimating the quality aspect of the test scores and then passing these aspects into the Support Vector Machine either as features or weights. Since the fusion process is based on the learning classifier of Support Vector Machine, the technique is termed Support Vector Machine with Quality Measurement (SVM-QM). The experimental investigation is conducted using face and speech modalities. The results clearly show the benefits gained from learning the quality aspects of the biometric data used for authentication.

**Adaptive Multimodal Fusion by Uncertainty Compensation**

*Vassilis Pitsikalis, Athanassios Katsamanis, George Papandreou, Petros Maragos; National Technical University of Athens, Greece*

*Thu2WeO.2*

While the accuracy of feature measurements heavily depends on changing environmental conditions, studying the consequences of this fact in pattern recognition tasks has received relatively little attention to date. In this work we explicitly take into account feature measurement uncertainty and we show how classification rules should be adjusted to compensate for its effects. Our approach is particularly fruitful in multimodal fusion scenarios, such as audio-visual speech recognition, where multiple streams of complementary time-evolving features are integrated. For such applications, provided that the measurement noise uncertainty for each feature stream can be estimated, the proposed framework leads to highly adaptive multimodal fusion rules which are widely applicable and easy to implement. We further show that previous multimodal fusion methods relying on stream weights fail under our scheme under certain assumptions; this provides novel insights into their applicability for various tasks and suggests new practical ways for estimating the stream weights adaptively. The potential of our approach is demonstrated in audio-visual speech recognition using either synchronous or asynchronous models.

**Effects of Familiarity with Faces and Voices on Second-Language Speech Processing: Components of Memory Traces**

*Debra M. Hardison; Michigan State University, USA*

*Thu2WeO.3*

Familiarity with a talker’s voice and face was found to facilitate processing of second-language speech. This advantage is accentuated when visual cues are limited to either the mouth and jaw area, or eyes and upper cheek areas of a talker’s face. Findings are compatible with a multiple-trace model of bimodal speech processing.

**Automatic Metadata Generation and Video Editing Based on Speech and Image Recognition for Medical Education Contents**

*Satoshi Tamura1, Koji Hashimoto1, Jiong Zhu1, Satoru Hayamizu1, Hirotsugu Asai2, Hideki Tanahashi2, Makoto Kanagawa1; 1Gift University, Japan; 2Gift Prefectural Research Institute of Manufacturing Information Technology, Japan; 3Sanyo Electric Co. Ltd., Japan*

*Thu2WeO.4*

This paper reports a metadata generation system as well as an automatic video edit system. The metadata are information described about the other data. In the audio metadata generation system, speech recognition using general language model (LM) and specialized LM is performed to input speech in order to obtain segment (event group) and audio metadata (event information) respectively. In the video edit system, visual metadata obtained by image recognition and audio metadata are combined into audio-visual metadata. Subsequently, multiple videos are edited to one video using the audio-visual metadata. Experiments were conducted to evaluate event detection of the systems using medical education contents, ACLS and BLS. The audio metadata system achieved about a 78% event detection correctness. In the edit system, an 87% event correctness was obtained by audio-visual metadata, and the survey proved that the edited video is appropriate and useful.

**Analysis of Correlation Between Audio and Visual Speech Features for Clean Audio Feature Prediction in Noise**

*Ibrahim Almajai, Ben Milner, Jonathan Darch; University of East Anglia, UK*

*Thu2WeO.5*

The aim of this work is to examine the correlation between audio and visual speech features. The motivation is to find visual features that can provide clean audio feature estimates which can be used for speech enhancement when the original audio signal is corrupted by noise. Two audio features (MFCCs and formants) and three visual features (active appearance model, 2-D DCT and cross-DCT) are considered with correlation measured using multiple linear regression. The correlation is then exploited through the development of a maximum a posteriori (MAP) prediction of audio features solely from the visual features. Experiments reveal that features representing broad spectral information have higher correlation to visual features than those representing finer spectral detail. The accuracy of prediction follows the results found in the correlation measurements.

**TDA: A New Trainable Trajectory Formation System for Facial Animation**

*Oxana Govokhina1, Gérard Bailly1, Gaspard Breton2, Paul Bagshaw2; 1ICP, France; 2France Télécom, France*

*Thu2WeO.6*

A new trainable trajectory formation system — named TDA — for facial animation is here proposed that dissociates parametric spaces and methods for movement planning and execution. Movement planning is achieved by HMM-based trajectory formation. This module essentially plans configurations of lip geometry (aperture, spreading and protrusion). Movement execution is performed by concatenation of multi-represented diphones. This module is responsible for selecting and concatenating detailed facial movements that best obey to the target kinematics of the geometry previously planned. Movement planning ensures that the essential visual characteristics of visemes are reached (lip closing for bilabials, rounding and opening for palatal fricatives, etc) and that appropriate coarticulation is planned. Movement execution grafts phonetic details and idiosyncratic articulatory strategies (dissymmetries, importance of jaw movements, etc) to the planned gestural score. This planning scheme is compared to alternative planning strategies using articulatory modeling and motion capture data.
Thu2BuP : Speech Analysis
Poster Session, Butler Room, 14:00 – 16:00
Chair: Joseph Picone, Mississippi State University, USA

Modeling of Speech Signals Based on Bessel-Like Orthogonal Transform
Giorgio Biagetti, Paolo Crippa, Claudio Turchetti; Università Politecnica delle Marche, Italy

Thu2BuP.1
In this paper a novel modeling technique for speech signals, based on the source-filter model of speech production and on orthogonal transform theory, is presented. The proposed approach models the frequency response of such filter, by projection onto a basis of damped Bessel functions, which have been chosen for their similarity to the signal to be modeled. In such a way an orthogonal transform pair is defined which provides a simple and effective methodology for the extraction of model parameters, and its effectiveness in the case of voiced speech has been demonstrated by synthesizing natural sounding speech signals with the aid of only a few extracted parameters.

Glottal Closure and Opening Detection for Flexible Parametric Voice Coding
Pamornpol Jinchitra; Stanford University, USA

Thu2BuP.2
The knowledge of glottal closure and opening instants (GCC/GCO) is useful for many speech analysis applications. A Pitch-synchronous waveform encoding of voice is one such application. In this paper, a dynamic programming is employed to solve for the global close/open phase segmentation based on the polynomial parametric waveform of the derivative glottal waveform and its quasi-periodicity. Not only does the algorithm identify GCIs, but also the elusive GOSs, and as a by-product, the parameters of the glottal excitation waveform. The results show its effectiveness compared with a classical method. Its application to parametric voice encoding which allows for simple time-pitch scaling as well as voicing quality conversion is demonstrated.

Independent Components for Acoustic Modeling
Jan Trmal, Jan Vaněk, Luděk Müller, Jan Zelinka; University of West Bohemia, Czech Republic

Thu2BuP.3
In the paper, we present a comparative study of several methods used nowadays in the field of feature and information extraction. We compared several Independent Component Analysis (ICA) algorithms together with the commonly used Principal Component Analysis (PCA) algorithm in two real-world tasks. The first task was a Voice Activity Detection (VAD), the second is Speaker Verification and Recognition (SVR). The VAD system as well as the SVR system benefited from the ICA decompositions. Moreover, a brief comparison of the information extraction ability is described.

Pitch-Scale Modification Using the Modulated Aspiration Noise Source
Daryush Mehta1; Thomas F. Quatieri2; 1HST, USA; 2Massachusetts Institute of Technology, USA

Thu2BuP.4
Spectral harmonic/noise component analysis of spoken vowels shows evidence of noise modulations with peaks in the estimated noise source component synchronous with both the open phase of the periodic source and with time instants of glottal closure. Inspired by this observation of natural modulations and of fullband energy in the aspiration noise source, we develop an alternate approach to high-quality pitch-scale modification of continuous speech. Our strategy takes a dual processing approach, in which the harmonic and noise components of the speech signal are separately analyzed, modified, and re-synthesized. The periodic component is modified using standard modification techniques, and the noise component is handled by modifying characteristics of its source waveform. Since we have modeled an inherent coupling between the periodic and aspiration noise sources, the modification algorithm is designed to preserve the synchrony between temporal modulations of the two sources. The reconstructed modified signal is perceived in informal listening to be natural-sounding and typically reduces artifacts that occur in standard modification techniques.

Max-Gabor Analysis and Synthesis of Spectrograms
Tony Ezzat, Jake Bouvrie, Tomaso Poggio; Massachusetts Institute of Technology, USA

Thu2BuP.5
We present a method that analyzes a two-dimensional magnitude spectrogram \( S(f,t) \) into its local constituent spectro-temporal amplitudes \( A(f,t) \), frequencies \( F(f,t) \), orientations \( \phi(f,t) \), and phases \( \psi(f,t) \). The method operates by performing a two-dimensional local Gabor-like analysis of the spectrogram, retaining only the parameters of the 2D-Gabor filter with maximal amplitude response within the local region. We demonstrate the technique over a wide variety of speakers, and show how the spectrograms in each case may be adequately reconstructed using the parameters of the Max-Gabor analysis. Finally, we discuss the nature of the extracted Max-Gabor parameters.

Monitoring of the Natural Voice Variations in Open and Closed Phases with Frequency Warped ARMA Modeling
Pedro J. Quintana-Morales, Juan L. Navarro-Mesa, Antonio G. Ravelo-García, Fernando D. Lorenzo-Garcia; Universidad de Las Palmas de Gran Canaria, Spain

Thu2BuP.6
The objective of this paper is to propose the use of a speech model with psychoacoustical information to distinguish between the open and closed phases of the vocal folds, in order to monitor formants as phonetic speech characteristics. Taking a frequency warped ARMA model for each one of the phases, the aim is to integrate the information of various consecutive periods in which the poles and the zeros that form the vocal tract can be considered common due to its slow variation. To analyze the capacity of phonetic monitoring, different phonetic voice transitions of speech registers are used from a database that provides information on glottal closings. First, we show the dependency of the warping factor. Then, the consistency and reliability of the method is demonstrated, as well as its better performance compared to no warped ones. And finally, we will see the well behavior against moderate noise.

Speech Analyzer Using a Joint Estimation Model of Spectral Envelope and Fine Structure
Hirokazu Kameoka, Jonathan Le Roux, Nobutaka Ono, Shigeki Sagayama; University of Tokyo, Japan

Thu2BuP.7
We have been working on a new speech analyzer based on a parametric representation of speech governed by the \( F_0 \) parameter, towards practical human-machine interfaces. As a precise estimation of the frequency response of the vocal tract from a real speech signal requires the power of each component of the harmonic structure to be accurately estimated, one hopes to have a high-precision estimation of \( F_0 \). At the same time, under the empirical constraint that speech spectral envelopes are usually smooth in the power domain, half pitch errors can be significantly avoided. Therefore, \( F_0 \) and the envelope should be estimated jointly rather than separately through an optimal estimation of the spectral envelope and the spectral fine structure. In this article, we introduce a new speech analysis method using a spectral model with a composite function of envelope and fine structure models.
An Investigation of Manifold Learning for Speech Analysis
Andrew Errity, John McKenna; Dublin City University, Ireland
Thu2BuP.8
Due to the physiological constraints of articulatory motion the speech apparatus has limited degrees of freedom. As a result, the range of speech sounds a human is capable of producing may lie on a low dimensional submanifold of the high dimensional space of all possible sounds. In this study a number of manifold learning algorithms are applied to speech data in an effort to extract useful low dimensional structure from the high dimensional speech signal. The ability of these manifold learning algorithms to separate vowels in a low dimensional space is evaluated and compared to a classical linear dimensionality reduction method. Results indicate that manifold learning algorithms outperform classical methods in low dimensions and are capable of discovering useful manifold structure in speech data.

An Incremental Algorithm for Signal Reconstruction from Short-Time Fourier Transform Magnitude
Jake Bouvrie, Tony Ezzat; Massachusetts Institute of Technology, USA
Thu2BuP.9
We present an algorithm for reconstructing a time-domain signal from the magnitude of a short-time Fourier transform (STFT). In contrast to existing algorithms based on alternating projections, we offer a novel approach involving numerical root-finding combined with explicit smoothness assumptions. Our technique produces high-quality reconstructions that have lower signal-to-noise ratios when compared to other existing algorithms. If there is little redundancy in the given STFT, in particular, the algorithm can produce signals which also sound significantly better perceptually, as compared to existing work.

Automatic Assignment of Anchoring Points on Vowel Templates for Defining Correspondence Between Time-Frequency Representations of Speech Samples
Toru Takahashi, Masashi Nishi, Toshio Irino, Hideki Kawahara; Wakayama University, Japan
Thu2BuP.10
The automatic assignment of anchoring points is proposed to define the correspondence between the time-frequency representations of speech sounds for speech morphing, speech texture mapping, and so on. The correspondence is modeled as a set of segmental bilinear function. These model parameters are called anchoring points. Although, the correspondence significantly affects the quality of such manipulated speech sounds as morphed and texture mapped speech sounds, anchoring points were manually aligned on time-frequency representations.

Anchoring points should be placed at auditorily important locations. When a spectrogram is presented as a time-frequency representation, auditorily important locations are given by formant frequencies around vowel transitions. The central idea of the proposed method is to prepare vowel template spectra with pre-assigned anchoring points in advance and to deform one of the templates to match the input speech spectrum. Finally, anchoring points on the input spectrum are copied from pre-assigned anchoring points.

Experimental results suggest that the naturalness of morphed speech based on the proposed automatic assignment method has equivalent quality to STRAIGHT synthetic speech samples.

Nonlinear Dynamical Invariants for Speech Recognition
S. Prasad, S. Srinivasan, M. Pannuri, G. Lazarou, Joseph Picone; Mississippi State University, USA
Thu2BuP.11
There is growing interest in modeling nonlinear behavior in the speech signal, particularly for applications such as speech recognition. Conventional tools for analyzing speech data use information from the power spectral density of the time series, and hence are restricted to the first two moments of the data. These moments do not provide a sufficient representation of a signal with strong nonlinear properties. In this paper, we investigate the use of features, known as invariants, that measure the nonlinearity in a signal. We analyze three popular measures: Lyapunov exponents, Kolmogorov entropy and correlation dimension. These measures quantify the presence (and extent) of chaos in the underlying system that generated the observable. We show that these invariants can discriminate between broad phonetic classes on a simple database consisting of sustained vowels using the Kullback-Leibler divergence measure. These features show promise in improving the robustness of speech recognition systems in noisy environments.

Thu2CaP : Advances in Noisy ASR
Poster Session, Cambria Room, 14:00 - 16:00
Chair: Lin-shan Lee, National Taiwan University, Taiwan

Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition
Shih-Hsiang Lin, Yao-Ming Yeh, Berlin Chen; National Taiwan Normal University, Taiwan
Thu2CaP.1
The performance of current automatic speech recognition (ASR) systems radically deteriorates when the input speech is corrupted by various kinds of noise sources. Quite a few of techniques have been proposed to improve ASR robustness in the past several years. Histogram equalization (HEQ) is one of the most efficient techniques that have been used to compensate the nonlinear distortion. In this paper, we explored the use of the data fitting scheme to efficiently approximate the inverse of the cumulative density function of training speech for HEQ, in contrast to the conventional table-lookup or quantile based approaches. Moreover, the temporal average operation was also performed on the feature vector components to alleviate the influence of sharp peaks and valleys that were caused by non-stationary noises. Finally, we also investigated the possibility of combining our approaches with other feature discrimination and decorrelation methods. All experiments were carried out on the Aurora-2 database and task. Encouraging results were initially demonstrated.

Missing Data Mask Models with Global Frequency and Temporal Constraints
Sébastien Demange, Christophe Cerisara, Jean-Paul Haton; LORIA, France
Thu2CaP.2
Missing data recognition has been developed in order to increase noise robustness in automatic speech recognition. Many different factors, including the speech decoding process itself, shall be considered to locate the masks. In this work, we are considering Bayesian models of the masks, where every spectral feature is classified as reliable or masked, and is independent from the rest of the signal. This classification strategy can produce unrelated small “spots”, while experiments suggest that oracle reliable and unreliable features tend to be clustered into time-frequency blocks. We call this undesired effect: the “checkerboard” effect.

In this paper, we propose a new Bayesian missing data classifier that
integrates frequency and temporal constraints in order to reduce, or avoid, this "checkerboard" effect. The proposed classifier is evaluated on the Aurora2 connected digit corpora. Integrating such constraints in the missing data classification leads to significant improvements in recognition accuracy.

Multi-Stream ASR: An Oracle Perspective
Hemant Misra, Jithendra Vepa, Hervé Bourlard; IDIAP Research Institute, Switzerland
Thu2CaP.3
Multi-stream based automatic speech recognition (ASR) systems are usually shown to outperform single stream systems, especially in noisy test conditions. And, indeed, there is a trend today in ASR towards using more and more acoustic features combined at the input (early integration, possibly preceded by some linear or nonlinear transformation) or later in the recognition process (e.g., at the level of likelihoods, then referred to as late integration). However, to guarantee optimal exploitation of such multi-stream systems, we need to use features that are as much complementary as possible, while also using the best combination weights of those streams. In practice, it is never clear whether we fully exploit the potential of the available streams. This present paper investigates an 'oracle' test to provide some insight in these issues. Although not providing us with an absolute performance upper bound, oracle is shown to indicate the complimentary of the feature streams used, and to provide a reasonable reference target to evaluate combination strategies. The oracle analysis is supported by results obtained on Numbers95 database using different feature streams and entropy based combination method.

A Weight Estimation Method Using LDA for Multi-Band Speech Recognition
Koji Iwano, Kaname Kojima, Sadaoki Furui; Tokyo Institute of Technology, Japan
Thu2CaP.4
This paper proposes a band-weight estimation method using Linear Discriminant Analysis (LDA) for multi-band automatic speech recognition (ASR). In our scheme, a spectral domain feature, SPEC, is modeled using a multi-stream HMM technique. This paper also proposes the use of Output Likelihood Normalization (OLN) in combination with the LDA-based weight-estimation method in order to adjust the relative weights of individual word (phoneme) models. Experiments were conducted using Japanese connected digit speech in various kinds of noise and SNR conditions. Experimental results show that the proposed LDA-based method is effective in all noise conditions. The results also confirm that the combination of OLN with the LDA-based method further increases noise robustness of the multi-band ASR. Furthermore, comparing the results of LDA applied to the SPEC and MFCC features respectively, it can be seen that greater performance gains are achieved with the former case than with the latter; this means that SPEC within a multi-band speech recognition framework can more effectively deal with the noise contamination than MFCC.

Powered Cepstral Normalization (P-CN) for Robust Features in Speech Recognition
Chang-wen Hsu, Lin-shan Lee; National Taiwan University, Taiwan
Thu2CaP.5
Cepstral normalization has been popularly used as a powerful approach to produce robust features for speech recognition. Good examples of approaches in this family include the well known Cepstral Mean Subtraction (CMS) and Cepstral Mean and Variance Normalization (CMVN), in which either the first or both the first and the second moments of the Mel-frequency Cepstral Coefficients (MFCCs) are normalized. In this paper, an improved approach of Powered Cepstral Normalization (P-CN) is proposed to normalize the MFCC parameters in the \( r \)-th powered domain, where \( r > 1.0 \). The basic idea is that when the MFCC parameters are raised to the \( r \)-th power, the harmful parts of environmental disturbances may be more emphasized than the speech features which are relatively smooth. Therefore performing the normalization in the domain of the \( r \)-th power may be more helpful. But the value of \( r \) should not be too large because in that case the environmental disturbances may be exaggerated and further corrupt the speech features. This approach is computationally simple and efficient. Initial experimental results on AURORA 2.0 testing environment showed that significant improvements in recognition rates are consistently obtainable under all different noisy conditions.

Robust Automatic Speech Recognition for Accented Mandarin in Car Environments
Pei Ding, Lei He, Xiang Yan, Jie Hao; Toshiba China R&D Center, China
Thu2CaP.6
This paper addresses the issues of robust automatic speech recognition (ASR) for accented Mandarin in car environments. A robust front-end is proposed, which adopts a Minimum Mean-Square Error (MMSE) estimator to suppress the background noise in frequency domain, and then implements spectrum smoothing both in time and frequency index to compensate those spectrum components distorted by the noise over-reduction. In the context of Mandarin speech recognition, a special adverse factor is the diversification of Chinese dialects, i.e. the pronunciation difference among dialects decreases the recognition performance if the acoustic models are trained with an unmatched accented database. We propose to train the models with multiple accented Mandarin databases to solve this problem. Evaluation results of isolated phrase recognition show that the proposed front-end can obtained the average error rate reduction (ERR) of 58.3% and 9.7% for artificial car noisy speech and real in-car speech respectively, when compared with the baseline in which no noise compensation technology is used. The efficiency of the proposed model training scheme is also proved in the experiments.

A Robust Feature Extraction Based on the MTF Concept for Speech Recognition in Reverberant Environment
Xugang Lu, Masashi Unoki, Masato Akagi; JAIST, Japan
Thu2CaP.7
This paper proposes a robust feature extraction method for automatic speech recognition (ASR) systems in reverberant environment. In this method, a sub-band power envelope inverse filtering algorithm based on the modulation transfer function (MTF), that we have previously proposed, is incorporated as a front-end processor for ASR. The impulse response of the room acoustics is assumed to be exponential decay modulated white noise, and speech is assumed to be temporal modulated white noise in each sub-band. Therefore, the impulse response of the environment does not need to be measured. Testing demonstrated that this algorithm can restore the temporal power envelope of reverberant speech in subbands and thus reduce the loss of speech intelligibility caused by reverberation. Testing of its ability to recognize digitized Japanese speech was done by using reverberant speech created by simple convolution of the room acoustics and speech. The algorithm had a 32.1% higher error reduction rate (on average, for reverberation times from 0.1 to 2.0 s) compared with the traditional cepstral mean normalization (CMN) of the auditory power spectrum based method (AFCC).

Clean Speech Feature Estimation Based on Soft Spectral Masking
Young Joon Kim\(^1\), Woohyung Lim\(^2\), Nam Soo Kim\(^2\); \(^1\) ETRI, Korea; \(^2\) Seoul National University, Korea
Thu2CaP.8
In this paper, we first analyze the problems of speech and noise contamination process in noise-masking point of view, and propose a new approach to estimate degree of noise masking effect on clean speech distribution model based on sequential noise estimation. Sequential noise estimation is performed frame-by-frame using interacting
multiple model (IMM) algorithm, so that real-time implementation is possible. After applying IMM algorithm, degree of noise masking effect named as noise masking probability (NMP) is calculated. Estimation of clean speech spectrum in noisy environments is performed by controlling the advantages of log spectrum domain and those of linear spectrum domain algorithm based on NMP. We have performed recognition experiments under noise conditions using the AURORA2 database which is developed for a standard reference of speech recognition performance. Simulation results show that this approach is effective when noise masking effect is dominated at low SNR.

Robust Speech Recognition by Modifying Clean and Telephone Feature Vectors Using Bidirectional Neural Network
Mansoor Vali, Seyed Ali Seyed Salehi, Kazem Karimi; Amirkabir University of Technology, Iran
Thu2CaP.9
In this paper we present a new method for nonlinear compensation of distortions, e.g. channel effects and additive noise, in clean and telephone speech recognition. A Bidirectional Neural Network (Bidi-NN) was developed and implemented in order to modify distorted input feature vectors and improve the overall recognition accuracy. Distorted components in feature vectors were estimated in accordance with the latent knowledge in the hidden layer of the neural network. This knowledge is obtained by training with clean and telephone speech, simultaneously and is mostly induced by phonemic content and less influenced by the irrelevant variations in signal. An MLP neural network was trained with these modified feature vectors. Comparing the achieved results with a reference model that was trained with unmodified feature vectors, demonstrate significant improvement in clean and telephone speech recognition accuracy.

Silence Energy Normalization for Robust Speech Recognition in Additive Noise Environment
Chung-fu Tai, Jeh-weth Hung; National Chi Nan University, Taiwan
Thu2CaP.10
The energy parameter has been widely used as an extension to the basic features of mel-frequency cepstral coefficients (MFCCs) to improve the recognition accuracy in speech recognition. In this paper, a simple and effective approach for energy normalization for silence (non-speech) portions in an utterance is proposed. This approach, named as silence energy normalization (SEN), uses the high-pass filtered log-energy as the feature for speech/non-speech classification, and then the log-energy of non-speech frames is set to be a small constant while that of speech frames is kept unchanged. In the experiments conducted on AURORA2 database, we showed that SEN provides an averaged word error rate reduction of 34.9% and 44.6% for Test Sets A and B, respectively, when compared with the baseline processing. It was also shown that SEN outperforms similar approaches like energy subtraction (ES) and feature vector selection (FVS). Finally, we showed that SEN can be integrated with cepstral mean and variance normalization (CMVN), to achieve further improved recognition performance.

Handling Convolutional Noise in Missing Data
Automatic Speech Recognition
Maarten Van Segbroeck, Hugo Van hamme; Katholieke Universiteit Leuven, Belgium
Thu2CaP.11
Missing Data Techniques have already shown their effectiveness in dealing with additive noise in automatic speech recognition systems. For real-life deployments, a compensation for linear filtering distortions is also required. Channel compensation in speech recognition typically involves estimating an additive shift in the log-spectral or cepstral domain. This paper explores a maximum likelihood technique to estimate this model offset while some data are missing. Recognition experiments on the AURORA2 recognition task demonstrate the effectiveness of this technique. In particular, we show that our method is more accurate than previously published methods and can handle narrow-band data.

Noisy Speech Recognition Based on Selection of Multiple Noise Suppression Methods Using Noise GMMs
Norihide Kitaoka, Souta Hamaguchi, Seiichi Nakagawa; Toyohashi University of Technology, Japan
Thu2CaP.12
To achieve high recognition performance for a wide variety of noise and for a wide range of signal-to-noise ratio, this paper presents integration methods of four noise reduction algorithms: spectral subtraction with smoothing of time direction, temporal domain SVD-based speech enhancement, GMM-based speech estimation and KLT-based comb-filtering. In this paper, we proposed two types of combination methods of noise suppression algorithms: selection of front-end processor and combination of results from multiple recognition processes. Recognition results on the AURORA-2 task showed the effectiveness of our proposed methods.

Using Posterior-Based Features in Template Matching for Speech Recognition
Guillermo Aradilla, Jihendra Vepa, Hervé Bourlard; IDIAP Research Institute, Switzerland
Thu2CaP.13
Given the availability of large speech corpora, as well as the increasing of memory and computational resources, the use of template matching approaches for automatic speech recognition (ASR) have recently attracted new attention. In such template-based approaches, speech is typically represented in terms of acoustic vector sequences, using spectral-based features such as MFCC of PLP, and local distances are usually based on Euclidean or Mahalanobis distances. In the present paper, we further investigate template-based ASR and show (on a continuous digit recognition task) that the use of posterior-based features significantly improves the standard template-based approaches, yielding to systems that are very competitive to state-of-the-art HMMs, even when using a very limited number (e.g., 10) of reference templates. Since those posteriors-based features can also be interpreted as a probability distribution, we also show that using Kullback-Leibler (KL) divergence as a local distance further improves the performance of the template-based approach, now beating state-of-the-art of more complex posterior-based HMMs systems (usually referred to as “Tandem”).

Hypothesis-Based Feature Combination of Multiple Speech Inputs for Robust Speech Recognition in Automotive Environments
Yasunari Obuchi, Nobuo Hataoka; Hitachi Ltd., Japan
Thu2CaP.14
In a microphone array system, feature combination in the MFCC domain can improve speech recognition accuracy. Multiple microphones provide different feature parameters such as MFCCs even if they have similar speech and noise signals, because of the phase difference and transmission characteristics. In this paper, we investigate how the recognition performance changes when we average multiple MFCC feature vectors. In addition, we extend Hypothesis-Based Feature Combination, which we formerly proposed for dual-microphone systems, to multi-input systems. Experimental results show that variance re-scaling is necessary when we combine multiple inputs with Cepstral Mean Normalization (CMN), in both MFCC average and HBFC. However, we can obtain better results without variance re-scaling if we use Mean and Variance Normalization (MVN) with MFCC average or HBFC. In the experiments using the database collected in a real automotive
Continuous Time-Frequency Masking Method for Blind Speech Separation with Adaptive Choice of Threshold Parameter Using ICA
Zbyněk Koldovský, Jan Nouza, Jan Kolorenc; Technical University of Liberec, Czech Republic
Thu2FoP.1
We propose a novel method for blind speech separation using continuous time-frequency masking. The method is equipped with an adaptive choice of a threshold parameter that is based on utilization of ICA methods. We present a direct application that consists in the speech segregation for automatic transcription of spoken broadcasts disturbed by background music. Experimental results show improved performance in comparison with traditionally used binary masking methods.

Multistage Convolutive Blind Source Separation for Speech Mixture
Yanxue Liang, Ichiro Hagiwara; Tokyo Institute of Technology, Japan
Thu2FoP.2
Blind source separation for convolutive mixture of speech signals has been addressed in many literatures. However, widely applied Multichannel Blind Deconvolution (MBD) method suffers whitening effect or arbitrary filtering problem which results in dramatic decrease of Automatic Speech Recognition system's performance. In present paper, a new MBD based multistage method is proposed, in which contributions of each source to every microphone are final goal rather than original signals. In detail, MBD is first implemented using entropy maximization criterion combined with Natural Gradient (NG) algorithm, then compensation matrix is constructed, based on which sources are recovered to its contribution to every microphone, i.e., whitening effect or arbitrary filtering problem has been transformed to fixed filtering problem. After compensation processing, for a certain source, it becomes Single Input and Multi-Output (SIMO) problem. Thus, not only spatial quality of source can be preserved, but also SIMO blind deconvolution can be further applied to fully recover temporal structure of speech signal. Finally, experiment shows validity and superiority over other methods in both spectra preservation efficiency and speed.

Detection and Separation of Speech Events in Meeting Recordings
Futoshi Asano, Jun Ogata; AIST, Japan
Thu2FoP.3
When applying automatic speech recognition (ASR) to meeting recordings including spontaneous speech, the performance of ASR is greatly reduced by the overlap of speech events. In this paper, a method of separating the overlapping speech events using an adaptive beamforming (ABF) framework is proposed. The main feature of this method is that all the necessary information for the adaptation of ABF, including microphone calibration, is obtained from meeting recordings based on the results of speech event detection. The performance of the separation is evaluated via ASR using real meeting recordings.

Audio Person Tracking in a Smart-Room Environment
Alberto Abad, Carlos Segura, Dušan Macho, Javier Hernando, Climent Nadeu; Universitat Politècnica de Catalunya, Spain
Thu2FoP.4
Reliable measures of speaker positions are needed for computational perception of human activities taking place in a smart-room environment. In this work, it is described the development process and the experiments conducted in the design and implementation of an Audio Person Tracking system for smart-room environments. The proposed system is based on the SRP-PHAT algorithm, as it is known to perform robustly in most environmental conditions. Novelties proposed are aimed to enhance the accuracy of the system independently on the application scenario and to reduce the computational complexity.

Tracking and Beamforming for Multiple Simultaneous Speakers with Probabilistic Data Association Filters
Tobias Gehrig, Ulrich Klee, John W. McDonough, Shajith Ikbal, Matthias Wölfl, Christian Fügen; Universität Karlsruhe, Germany
Thu2FoP.5
In prior work, we developed a speaker tracking system based on an extended Kalman filter using time delays of arrival (TDOAs) as acoustic features. While this system functioned well, its utility was limited to scenarios in which a single speaker was to be tracked. In this work, we remove this restriction by generalizing the IEKF, first to a probabilistic data association filter, which incorporates a clutter model for rejection of spurious acoustic events, and then to a joint probabilistic data association filter (JPDADF), which maintains a separate state vector for each active speaker. In a set of experiments conducted on seminar and meeting data, the JPDADF speaker tracking system reduced the multiple object tracking error from 20.7% to 14.3% with respect to the IEKF system. In a set of automatic speech recognition experiments conducted on the output of a 64 channel microphone array which was beamformed using automatic speaker position estimates, applying the JPDADF tracking system reduced word error rate from 67.3% to 66.0%. Moreover, the word error rate on the beamformed output was 13.0% absolute lower than on a single channel of the array.

Modeling the Precedence Effect for Binaural Sound Source Localization in Noisy and Echoic Environments
Martin Heckmann, Tobias Rodemann, Björn Schölling, Frank Joublin, Christian Goerick; Honda Research Institute Europe GmbH, Germany
Thu2FoP.6
We present a new way of modelling the Precedence Effect to enable the robust measurement of localization cues (ITD and IID) in echoic environments. Based on this we developed a localization system which is inspired by the auditory system of mammals. It uses a Gammatone filter bank for preprocessing and extracts the ITD cue via zero crossings (IID calculation is straightforward). The mapping between the cue values and the different angles is learned offline which facilitates the adaptation to different head geometries. The performance of the system is demonstrated by localization results for two simultaneous speakers and the mixture of a speaker, music, and fan noise in a normal meeting room. A real-time demonstrator of the system is presented in [1].

Using a Differential Microphone Array to Estimate the Direction of Arrival of Two Acoustic Sources
Fotios Talantzis1, Anthony G. Constantinides2, Lazaros C. Polymenakos1; 1 Athens Information Technology, Greece; 2 Imperial College London, UK
Thu2FoP.7
When applying automatic speech recognition (ASR) to meeting recordings including spontaneous speech, the performance of ASR is greatly reduced by the overlap of speech events. In this paper, a method of separating the overlapping speech events using an adaptive beamforming (ABF) framework is proposed. The main feature of this method is that all the necessary information for the adaptation of ABF, including microphone calibration, is obtained from meeting recordings based on the results of speech event detection. The performance of the separation is evaluated via ASR using real meeting recordings.
We present a system that estimates the direction of arrival of two competing acoustic sources using two closely spaced receivers that form a differential microphone array. The main advantage of the proposed array topology is that null steering can be essentially performed by adapting a set of two scalars. The direction of arrival estimation relies on the successful estimation of the relative delays between the microphone signals using the decorrelation constraint. Processing is performed in real-time by operating on blocks of recorded data. We examine the performance of the system for different block sizes and investigate its robustness in environments of strong multipath reflections where algorithms often fail to distinguish between the true direction of arrival and that of a dominant reflection. The overall performance of the system is compared to the simple omni-directional array topology. The results indicate that the examined framework can track the two directions of arrival adequately.

**Speaker Localization Based on Oriented Global Coherence Field**

Alessio Bratti, Maurizio Omologo, Piergiorgio Svaizer; ITC-irst, Italy

Thu2FoP.8

This paper proposes a new speaker localization method that is based on a preliminary estimation of the head orientation. The basic information on which the estimation is accomplished is called Oriented Global Coherence Field (OGCF).

The new algorithm is shown to be significantly more robust than the traditional ones so far explored. Its robustness is also due to an effective speech activity detection, implicitly performed by a thresholding technique applied to OGCF information. To show the performance of the proposed system, experiments were conducted on the NIST RT-05 Spring Evaluation source localization task, which is based on real recordings of lectures in noisy and reverberant environments.

**Performance Evaluation of Three Features for Model-Based Single Channel Speech Separation Problem**

M.H. Radfar$^1$, R.M. Dansereau$^1$, A. Sayadayan$^2$; $^1$Carleton University, Canada; $^2$Amirkabir University of Technology, Iran

Thu2FoP.9

This paper addresses the efficiency of three features for the model-based single channel speech separation problem. The separability of three features: log spectrum, modulated lapped transform (MLT) coefficients, and a fusion of pitch and envelop information are evaluated using a VQ-based speech separation technique. At the core of this approach are two trained codebooks of the quantized feature vectors of speakers, whereby the main evaluation for separation is performed. The experiments are conducted in two different scenarios: speaker-dependent and speaker independent. The results show that the log spectrum outperforms the other features for speaker-dependent scenario. However, for the speaker-independent scenario, the best results are obtained from applying the pitch-envelope feature.

**Single-Channel Speech Separation Using Sparse Non-Negative Matrix Factorization**

Mikkel N. Schmidt, Rasmus K. Olsson; Technical University of Denmark, Denmark

Thu2FoP.10

We apply machine learning techniques to the problem of separating multiple speech sources from a single microphone recording. The method of choice is a sparse non-negative matrix factorization algorithm, which in an unsupervised manner can learn sparse representations of the data. This is applied to the learning of personalized dictionaries from a speech corpus, which in turn are used to separate the audio stream into its components. We show that computational savings can be achieved by segmenting the training data on a phoneme level. To split the data, a conventional speech recognizer is used. The performance of the unsupervised and supervised adaptation schemes result in significant improvements in terms of the target-to-masker ratio.

**Adaptive Speech Enhancement for Speech Separation in Diffuse Noise**

Rong Hu, Yunxin Zhao; University of Missouri-Columbia, USA

Thu2FoP.11

An adaptive enhancement method is proposed to improve recognition accuracy on the outputs of blind speech separation (BSS) system based on adaptive decorrelation filtering (ADF) in diffuse noise. A divide and conquer strategy is taken to deal with the noise effects on both system adaptation and ADF outputs. First, fast noise compensation (NC) is performed for filter adaptation, forcing ADF to focus on the task of separation; then, output noises are reduced by conventional speech enhancement, such as spectral subtraction or subspace methods. To make stationary-noise reduction techniques fit for output noises with time-varying properties caused by ADF adaptations, a fast adaptive procedure is developed to map known stationary input noise statistics to output. Separation and recognition experiments were conducted for both real and simulated diffuse noises, based on TIMIT speech data and impulse response data from a room with reverberation time $T_{50} = 0.3$ sec. The proposed techniques significantly improved phone recognition accuracy of ADF results.

**A Probabilistic Graphical Model for Microphone Array Source Separation Using Rich Pre-Trained Source Models**

H.T. Attias; Golden Metallic Inc., USA

Thu2FoP.12

Voice based computing applications, such as phone communication and speech recognition, use microphone arrays to capture voice from a human speaker. In many environments of interest, however, sounds from other sources interfere with the speaker’s voice, posing severe problems for subsequent processing. This paper describes a new framework for treating this problem, and presents and demonstrates a new algorithm for the cancellation of interfering sounds. Our framework combines techniques from statistical machine learning with ideas from speech and audio processing. An important feature involves training rich probabilistic models on data from different types of relevant sound sources. Those source models are then incorporated into a larger probabilistic model of the observed microphone data. Using that model we derive our algorithm, which is of the expectation-maximization type and infers from data the clean sound of separate individual sources. We report very good results on data recorded in different environments.

**Geometrically Constrained Permutation-Free Source Separation in an Undercomplete Speech Unmixing Scenario**

Erik Visser; SoftMax, USA

Thu2FoP.13

Frequency domain blind source separation (BSS) problems are typically solved in each frequency bin independently and therefore require additional measures to resolve the resulting permutation problem. In this paper, a frequency domain methodology is presented based on a recently introduced extension of Independent Component Analysis (ICA) to multi-variate components which uses a multi-variate activation function to model dependencies between frequency bins and therefore inherently manages to align most of the permutations. Since the latter approach shows slow convergence behaviour and is prone to converging to local optima, additional geometric constraints are used here to force the BSS algorithm to separate sources with a consistent direction of arrival (DOA) over all frequencies into a minimum number of output channels. DOA information is obtained from a priori knowl-
edge or from subband analysis of partially separated source signals. The methodology is illustrated in an undercomplete acoustic source separation scenario with 3 speakers and 4 microphones.

**Highly Directional Multi-Beam Audio Loudspeaker**

_Dirk Olszewski, Klaus Linhard; DaimlerChrysler AG, Germany_

_Thu2FoP.14_

This work presents an audio beam loudspeaker that possesses the capability of emitting multiple audio beams into different directions at the same time. Ultrasonic loudspeakers applying parametric array technology are used to generate highly directional audible sound beams. The loudspeaker itself appears as a single device and can be mounted under the ceiling of a room in order to supply selected room segments with highly directional audio information such as music or speech.

Due to the fact that parametric array technology requires very high ultrasound levels, high-power transducers have to be used. Because phased array technology applied on those transducers for beam steering does not work successfully, an alternative approach based on pre-steered segments is presented in this work.
Yi, Lifu ....................... Mon2A3O.5 64
Yingthawornsuk, T. .......... Mon3CaP.4 78
Yip, Michael C.W. ............. Tue1BuP.4 88
Ynoguti, Carlos Alberto ..... Mon1CaP.9 59
Yoma, Nestor Becerra ....... Mon3A1O.5 72
Yoshikawa, Tadahiro ........ Wed3FoP.8 143
You, Hong ........................ Mon3FoP.8 81
You, Jiali ..................... Tue3BuP.15 109
Young, Steve ................ Mon1A1O.5 52
Yu, Dong ........................ Mon3BuP.6 76
Yu, Kai ........................ Thu1CaP.12 150
Yu, Peng ........................ Wed2A3O.3 126
Yuan, Jiahong ............... Mon3A3O.1 74
Yun, Hwan Sik ................ Tue3FoP.2 111
Zahorian, Stephen A. ........ Wed2A1O.5 124
Zamalloa, Maider .......... Tue1CaP.2 89
Zavarehei, Esfandiar ........ Mon2A1O.2 62
Zdansky, Jindrich .......... Wed1CaP.13 122
Železný, Miloš ............... Mon3WeS.6 75
Zelinka, Jan .................. Thu2BuP.3 156
Zeller, Marcus .............. Tue1A2O.1 84
Zen, Heiga ................... Tue2BuP.13 99
Žgank, Andrej ............... Mon1CaP.12 59
Zhai, Lu-Feng ................ Mon2CaP.5 68
Zhang, Bing .................. Wed1A2O.5 115
Zhang, Chi .................... Mon3FoP.5 80
Zhang, Le ..................... Mon3BuP.1 75
Zhang, Qi ..................... Tue2A3O.5 96
Zhang, Rong .................. Mon3A2O.4 73
Zhang, Shen .................. Wed2BuP.4 128
Zhang, Sheng ................ Wed2BuP.6 128
Zhang, Shilei ................ Wed1FoP.1 122
Zhang, Shuwu ................. Wed1FoP.1 122
Zhang, Ying ................... Tue1A1O.6 83
Zhang, Yushi .................. Mon3CaP.6 140
Zhang, Zhengyou ............ Tue1CaP.7 90
Zhao, David Y. ............... Mon1CaP.10 59
Zhao, Li ....................... Mon1WeS.2 55
Zhao, Yong ................... Tue1A1O.2 83
Zhu, Xiaodan ................. Wed1WeS.2 117
Zimmerman, M. .............. Mon1BuP.4 56
Zitouni, Imed ................. Mon1BuP.11 57
Zoerner, Eric ................. Mon1CaP.8 59
Zollmann, Andreas ......... Tue1A1O.6 83
Zolnay, András ............. Mon2BuP.1 66
Zweig, G. .................... Tue3A2O.1 104

Z

Zahorian, Stephen A. ........ Wed2A1O.5 124
Zamalloa, Maider .......... Tue1CaP.2 89
Zavarehei, Esfandiar ........ Mon2A1O.2 62
Zdansky, Jindrich .......... Wed1CaP.13 122
Železný, Miloš ............... Mon3WeS.6 75
Zelinka, Jan .................. Thu2BuP.3 156
Zeller, Marcus .............. Tue1A2O.1 84
Zen, Heiga ................... Tue2BuP.13 99
Žgank, Andrej ............... Mon1CaP.12 59
Zhai, Lu-Feng ................ Mon2CaP.5 68
Zhang, Bing .................. Wed1A2O.5 115
Zhang, Chi .................... Mon3FoP.5 80
Zhang, Le ..................... Mon3BuP.1 75
Zhang, Qi ..................... Tue2A3O.5 96
Zhang, Rong .................. Mon3A2O.4 73
Zhang, Shen .................. Wed2BuP.4 128
Zhang, Sheng ................ Wed2BuP.6 128
Zhang, Shilei ................ Wed1FoP.1 122
Zhang, Shuwu ................. Wed1FoP.1 122
Zhang, Ying ................... Tue1A1O.6 83
Zhang, Yushi .................. Mon3CaP.6 140
Zhang, Zhengyou ............ Tue1CaP.7 90
Zhang, Zhiping ............... Mon1WeS.2 55
Zhao, David Y. ............... Mon1CaP.10 59
Zhao, Li ....................... Mon1WeS.2 55
Zhao, Yong ................... Tue1A1O.2 83
Zhu, Xiaodan ................. Wed1WeS.2 117
Zimmerman, M. .............. Mon1BuP.4 56
Zitouni, Imed ................. Mon1BuP.11 57
Zoerner, Eric ................. Mon1CaP.8 59
Zollmann, Andreas ......... Tue1A1O.6 83
Zolnay, András ............. Mon2BuP.1 66
Zweig, G. .................... Tue3A2O.1 104