

Tutorials

Monday, August 27, 2007

University of Antwerp building, Rodestraat 14
Rooms R008, R012, R013 and R124

Voice Quality in Vocal Communication

Christophe d'Alessandro (LIMSI-CNRS, Paris)

Morning Tutorial (9:00 – 12:30)

Overview

Until recently, voice quality and its functions in speech communication have been only marginally considered in the speech communication community. However, there is today some evidence that voice quality settings and voice quality modulations are playing a central role in human voice-based communication, i.e. speech, singing and other kinds of expressive vocalizations.

A challenging problem is the relationship between voice quality and prosody. Prosodic parameters are usually restricted to pitch, duration, pauses and some sort of intensity parameter. Voice quality is not usually considered in the prosodic field, which is more focused on intonation, accentuation and rhythm. However, synthesis of expressive speech has demonstrated that convincing natural sounding results are impossible to obtain without dealing with voice quality parameters.

Vocal expression of emotions and attitudes is one of the main domains of interest for voice quality research. Although it has been studied for a long time in psychology, it can be considered as an emerging research domain not only for speech recognition and synthesis, but also for speech coding and other areas of speech communication.

Voice quality is still a rather fuzzy concept: what is the timbre of a voice? How to measure and quantify vocal effort? What are the domains of variation of every day speech? What are the physical and perceptive correlates of voice quality? What are the relationships between voice quality and others aspects of prosody?

These matters and others will be addressed in the tutorial.

Presenter

Christophe d'Alessandro received the B.S. degree in Mathematics, the M.S and the Ph.D degrees in Computer Science from Paris VI University, in 1983, 1984 and 1989, respectively. He has been a permanent Researcher at LIMSI, a laboratory of the CNRS (French National Agency for Scientific Research), since October 1989. Prior to joining the CNRS, Dr. d'Alessandro has been a Lecturer in computer science at Paris XI University from October 1987 to October 1989. He also graduated in music, and he has been titular Organist at Sainte-Elisabeth in Paris, France, since 1992. Christophe d'Alessandro is « Directeur de Recherche » at the CNRS. He is the head of the “Situating Perception” group in the “Human-Machine Communication” department at LIMSI. His research interests include text-to-speech synthesis, signal processing for speech analysis and synthesis, perception and synthesis of intonation in speech and singing, voice source analysis and synthesis, speech synthesis assessment, gesture control of synthesis, musical acoustics. Christophe d'Alessandro edited a book on text-to-speech conversion ‘Synthèse de la parole à partir du texte’. He also organized the first international workshop on Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)

The Modulation Spectrum and Its Application to Speech Science and Technology

Les Atlas (Univ. of Washington), Steven Greenberg (Silicon Speech; Univ. of California, Berkeley; Technical Univ. of Denmark) and Hynek Hermansky (IDIAP, Martigny)

Morning Tutorial (9:00 – 12:30)

Overview

This tutorial describes the biological, mathematical and engineering bases of the modulation spectrum, which encapsulates many perceptually relevant properties of speech in the range between 50 and 1000 ms. The modulation spectrum reflects slow energy fluctuations associated with the opening and closing of the lips, the jaw and other speech articulators. These fluctuations are differentially distributed across the *acoustic frequency* spectrum. The constellation of modulation patterns across frequency and time is referred to as the *Complex Modulation Spectrum (CMS)*, in which both magnitude and phase are important. CMS-related features are beginning to be used in a variety of applications, including automatic speech recognition, speech synthesis, audio coding and auditory prostheses, and are likely to play an important role in audio/speech technology over the coming decade.

Presenters

Les Atlas received his M.S. and Ph.D. degrees in Electrical Engineering from Stanford University in 1979 and 1984, respectively. He joined the University of Washington in 1984, where he is currently a Professor of Electrical Engineering. His research is in digital signal processing, with specializations in acoustic analysis, time-frequency representations, as well as signal recognition and coding. Professor Atlas received a National Science Foundation Presidential Young Investigator Award and a 2004 Fulbright Senior Research Scholar Award. He was General Chair of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, and Chair of the IEEE Signal Processing Society Technical Committee on Theory and Methods. He is a Fellow of the IEEE “for contributions to time-varying spectral analysis and acoustical signal processing.”

Steven Greenberg received the A.B. in Linguistics from the University of Pennsylvania and a Ph.D. in Linguistics (with a strong minor in Neuroscience) from the University of California, Los Angeles. He has been a scientist at the International Computer Science Institute (Berkeley, CA). He is a Visiting Professor at the Centre for Applied Hearing Research, Technical University of Denmark and is also affiliated with the Center for New Music and Audio Technology, University of California, Berkeley. Dr. Greenberg’s research focuses on the interface between the science and technology of spoken language. His studies of human speech perception have examined the role of the modulation spectrum in understanding spoken language.

Hynek Hermansky is a senior researcher and director of research at IDIAP in Martigny, Switzerland, and serves as Professor at the Swiss Federal Institute of Technology in Lausanne, Switzerland. He has been working in speech processing for over 30 years, previously at the University of Tokyo, Panasonic Technologies in Santa Barbara, California, U S WEST Advanced Technologies, and has been a Professor and Director of the Center for Information Processing at OHSU Portland, Oregon. He is a Fellow of IEEE for “Invention and development of perceptually-based speech processing methods”, and a Member of the Editorial Board of Speech Communication and of Phonetica. He holds a Dr.Eng. Degree from the University of Tokyo, and a Dipl. Ing. Degree from Brno University of Technology, Czech Republic.

Spoken Language Processing by Mind and Machine

Roger K. Moore (Univ. of Sheffield) and Anne Cutler (Max-Planck-Institute for Psycholinguistics, Nijmegen)

Morning Tutorial (9:00 – 12:30)

Overview

This tutorial will compare and contrast spoken language processing as performed by machines with the corresponding processes performed by human beings. Theories of human speech perception, production, cognition and discourse will be discussed alongside algorithms for automatic speech recognition, synthesis, understanding and dialogue. Attention will be given to the latest research attempts to unify these areas within a common framework.

Great strides have been made in our understanding of how human beings use and process speech as well as in our ability to create and implement practical applications that incorporate voice-based interaction. However, our current level of knowledge is quite modest in comparison to the advanced communicative skills of the average human being, and the full potential of truly ubiquitous SL technology may not be able to be realised using today's models and algorithms. It is therefore timely for the different research communities to develop a modest understanding of each other's progress.

Presenters

Roger Moore studied Computer and Communications Engineering at the University of Essex and was awarded the B.A. (Hons.) degree in 1973. He subsequently received the M.Sc. and Ph.D. degrees from the same university in 1975 and 1977 respectively, both theses being on the topic of automatic speech recognition. In 1985 he became head of the newly created 'Speech Research Unit' (SRU) and subsequently rose to the position of Senior Fellow in the 'Defence and Evaluation Research Agency' (DERA). Since 2004 he has been Professor of Spoken Language in the 'Speech and Hearing' Research Group (SPandH) at Sheffield University. Prof. Moore is a Fellow of the Institute of Acoustics and a Visiting Professor in the Department of Phonetics and Linguistics at UCL. He is past Chairman of the EAGLES working party on spoken language resources, and Editor of the 'Handbook of Standards and Resources for Spoken Language Systems'. Prof. Moore served as President of the 'International Speech Communication Association' (ISCA) from 1997 to 2001 and President of the Permanent Council of the 'International Conferences on Spoken Language Processing' (ICSLP) from 1996 to 2001. In 1994 he was awarded the prestigious UK Institute of Acoustics Tyndall medal for "distinguished work in the field of speech research and technology" and in 1999 he was presented with the NATO RTO Scientific Achievement Award for "repeated contribution in scientific and technological cooperation".

Anne Cutler is a director of the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, Professor of Comparative Psycholinguistics at the Radboud University Nijmegen, and Professor in MARCS Auditory Laboratories at the University of Western Sydney. She studied in Australia, Germany and the US, and before moving to Nijmegen she worked in the UK, chiefly with the Medical Research Council in Cambridge. Her research centres on how language-specific phonological structure affects the recognition of spoken language. Recent research has included studies of the recognition of Dutch, English, German, French, Japanese, Korean, Finnish, Cantonese, Sesotho, Spanish, Telugu and Arabic.

Processing Morphologically-Rich Languages

Katrin Kirchhoff (Univ. of Washington) and Ruhi Sarikaya (IBM)

Morning Tutorial (9:00 – 12:30)

Overview

Morphologically-rich languages like Arabic, Turkish, Finnish, Korean, etc., present significant challenges for speech processing, natural language processing and machine translation. These languages are characterized by highly productive morphological processes (inflection, agglutination, compounding) that may produce a very large number of word forms for a given root form. Modelling each form as a separate word leads to a number of problems for speech and language processing applications, including:

1. increase in dictionary size
2. poor language model (LM) probability estimation
3. higher out-of-vocabulary (OOV) rate
4. inflection gap for machine translation.

Large-scale speech and language processing systems require more advanced modelling techniques to address these problems:

- automatic decomposition of complex word forms into smaller units
- methods for optimizing the selection of units at different levels of processing
- diacritization/vowelization (for Arabic)
- pronunciation modelling for morphologically-rich languages
- morphologically-rich languages in speech synthesis
- novel probability estimation techniques that avoid data sparseness problems
- creating data resources and annotation tools for morphologically-rich languages

Presenters

Dr. Kirchhoff is a Research Assistant Professor in the EE Department at the University of Washington. Her research interests are in automatic speech recognition and natural language processing, with an emphasis on multilingual applications. She has published over 50 refereed conference papers, journal papers, and book chapters in these areas and is co-editor of a recent book on “Multilingual Speech Processing”. In 2002 she led a team effort on developing Novel Models for Arabic Speech Processing at the Johns-Hopkins Summer Research Workshop. Dr. Kirchhoff has served on numerous conference and workshop committees and is a member of the Editorial Board of the Speech Communication journal.

Dr. Ruhi Sarikaya is a research staff member in the Human Language Technologies Group at IBM T.J. Watson Research Center. He received the B.S. degree from Bilkent University, Turkey in 1995, M.S. degree from Clemson University, SC in 1997 and the Ph.D. degree from Duke University, Durham, NC in 2001 all in electrical and computer engineering. At IBM he has received several prestigious awards for his work including an Outstanding Technical Achievement Award and a Research Division Award. Prior to joining IBM in 2001 he was a researcher at the Center for Spoken Language Research (CSLR) at the University of Colorado at Boulder for two years. He also spent the summer of 1999 at the Panasonic Speech Technology Laboratory, Santa Barbara, CA. He has served in the organizing committee of ASRU’05.

Voice Transformation

Yannis Stylianou (University of Crete)
Afternoon Tutorial (14:00 – 17:30)

Overview

Voice Transformation refers to the various modifications one may apply to the sound produced by a person, speaking or singing. A major application area of Voice Transformation is that of concatenating speech synthesis. Voice Transformation is a flexible, could be a simple, and efficient way to bring the variety needed in the current Text-To-Speech synthesis systems based on concatenation of units (large or small). Recently, because the speech quality produced by concatenating speech synthesis systems has considerably been improved, the demand for high quality Voice Transformation algorithms grew a lot.

In this tutorial we will provide a description of various ways in which someone can modify voice and provide details on how to implement these modifications using a simple, although quite efficient, parametric model based on a harmonic representation of speech. Discussing quality issues of current voice transformation algorithms in conjunction with properties of the speech production and perception systems we will try to pave the way for more natural Voice Transformation algorithms in the future.

Presenter

Yannis Stylianou is Associate Professor at University of Crete, Department of Computer Science since 2002. He received the Diploma of Electrical Engineering from NTUA, Athens, Greece in 1991 and the M.Sc. and Ph.D. degrees in Signal Processing from ENST, Paris, France in 1992 and 1996, respectively. From 1996 until 2001 he was with AT&T Labs Research (Murray Hill and Florham Park, NJ, USA) as a Senior Technical Staff Member. In 2001 he joined Bell-Labs Lucent Technologies, in Murray Hill, NJ, USA. He was Associate Editor for the IEEE Signal Processing Letters. Currently is Associate Editor for the EURASIP Journal on Speech and Audio Processing and Vice-Chairman of the COST Action 2103: “Advanced Voice Function Assessment” developing algorithms for voice quality control. He holds 8 patents.

A Mathematical Theory of Speech Signals – Beyond the Linear Model

Gernot Kubin and Erhard Rank (Graz Univ. of Technology)

Afternoon Tutorial (14:00 – 17:30)

Overview

This tutorial will introduce the speech modeling problem in a systematic way and review signal theoretic concepts from both the deterministic, dynamical systems perspective and the stochastic, information-theory perspective. It will show how to apply these advanced concepts to speech signals in cases where conventional linear approaches fail, and it will use these building blocks to develop a fullfledged oscillator-plus-noise model of speech signals that can be adapted automatically to sustained speech sounds. Finally, specific applications to continuous speech are discussed where the new signal theoretic methods have led to competitive engineering solutions.

The presentation will be organized in six units of approx. 30 minutes each:

1. Speech modeling as a signal theoretic problem
2. Deterministic theory – Nonlinear dynamical systems from fading-memory filters to chaotic oscillators
3. Stochastic theory – Cyclostationarity, higher-order statistics and information theory
4. First steps in nonlinear speech modeling – Where linear models fail and nonlinear models prevail
5. Speech analysis and synthesis by nonlinear prediction of the speech wave – How to automatically identify oscillator-plus-noise models for sustained speech sounds
6. Selected applications to continuous speech – Error concealment, time-scale modification, and pathological voice augmentation.

Presenters

Gernot Kubin has worked on speech analysis, synthesis, coding for mobile and IP telephony, error concealment, watermarking, enhancement and augmentation, echo cancellation, resource collection, as well as the recognition of speech, speakers, and regional varieties over the past 25 years (including affiliations with TU Vienna, Philips Research Eindhoven, AT&T Bell Laboratories Murray Hill, KTH Stockholm, Global IP Sound Stockholm/San Francisco) and, in particular, on nonlinear speech processing since 1990. He is currently full professor and head of the Signal Processing and Speech Communication Laboratory at Graz University of Technology, Graz, Austria, scientific director of the Christian Doppler Laboratory for Nonlinear Signal Processing and of the Competence Network for Advanced Speech Technology COAST, and he is a member of the board of the Vienna Telecommunications Research Centre FTW.

Erhard Rank has worked in speech processing, synthesis and recognition, in particular in hybrid concatenative and model based speech synthesis, synthesis of emotional speech, nonlinear model based synthesis of speech and musical signals, and noise reduction/speech enhancement. He was affiliated with the Austrian Research Institute for Artificial Intelligence ÖFAI, the Vienna Telecommunications Research Centre FTW, and as research and teaching assistant at TU Vienna and TU Graz. In his PhD thesis (2005) he developed an oscillator model for the automatic identification and re-synthesis of speech sounds. He is currently with the Signal Processing and Speech Communication Laboratory at Graz University of Technology, Graz, Austria.

Talking to Computers: from Speech Sounds to Human Computer Interaction

Giuseppe Riccardi and Sebastian Varge (Univ. of Trento)

Afternoon Tutorial (14:00 – 17:30)

Overview

This tutorial will provide an introduction to dialogue systems by addressing general design issues (application scenarios, HCI, devices and multi-modality) and outlining the available choices for the different modules of typical dialogue systems. We will systematically describe both commercially deployed systems and research prototypes along various dimensions, including architectural choices, motivating theoretical ideas, application domains, use of statistical versus rule-based methods, training material used in corpus-based components and evaluation methods and results. We will provide information about available software toolkits that can give the participants starting points to conduct experimental research and implement their own dialogue systems. In addition, we will describe current research issues and challenges, for example the use of 'thick pipelines' and corpus-based methods.

The tutorial will be structured as follows:

1. *Introduction*: historical remarks on first and second generation dialogue systems. Motivation for using spoken/multimodal dialogue systems; discussion of application scenarios in brief.
2. *Theoretical framework*: detailed description of individual modules (Speech Recognition, Natural Language Understanding, Dialogue Management, Task Execution, Response Generation and Speech synthesis), underlying theoretical ideas, dialogue system architectures and interaction between modules.
3. *Toolkits and standards* (e.g. VoiceXML, SALT), annotation and data collection, evaluation methodologies. HCI: general principles and typical HCI design errors.

Presenters

Prof. Riccardi received his Laurea degree in Electrical Engineering and Master in Information Technology, in 1991, from the University of Padua and CEFRIEL Research Center, respectively. From 1990-1993 he collaborated with Alcatel-Telettra Research Laboratories (Milan, Italy). In 1995 he received his Ph.D. in Electrical Engineering from the University of Padua, Italy. From 1993-2005, he worked first at AT&T Bell Laboratories and then AT&T Labs-Research where he worked in the Speech and Language Processing Lab. In 2005 joined the faculty of Engineering at University of Trento (Italy). Prof. Riccardi's research on stochastic finite state machines for speech and language processing has been applied to a wide range of domains for task automation, including in the well-known "How May I Help You?" research program which led to a speech service breakthrough and for the creation of the first large scale finite state chain decoding for machine translation ("Anuvaadi").

Sebastian Varges is Senior Marie Curie Fellow in the ADAMACH project (ADaptive and meaning MACHines) at the DIT department of the University Trento. Before Trento, he conducted research on an in-car dialogue system at Stanford University and on natural language generation at the Universities of Brighton and Edinburgh (where he obtained his PhD on "Instance-based Natural Language Generation"). His research interests focus on dialogue systems and natural language generation, in particular in combining statistical and rule-based approaches.

Machine Learning for Text and Speech Processing

Antal van den Bosch (Tilburg University) and Walter Daelemans (Univ. of Antwerp)

Afternoon Tutorial (14:00 – 17:30)

Overview

The automatic synthesis of fluent speech requires robust and accurate natural language processing technology at the word level (word pronunciation) and the text level (prosody generation). Machine learning offers methods to learn these processing tasks and their intermediary prerequisites based on annotated lexicons and corpora.

At the word level we will review machine learning approaches for G2P, tokenization and POS tagging; at the sentence level we focus on phrase chunking and semantic information sources for accent and phrase break placement. We will review the main strands of existing algorithmic solutions (stressing the lazy-eager learning dimension), and describe state-of-the-art sequence processing methods that distinguish between local classification and global search.

Presenters

Antal van den Bosch is associate professor at the Dept. of Communication and Information at Tilburg University, The Netherlands, heading the ILK Research Group. His work focuses on memory-based machine learning methods applied to NLP (e.g. with Daelemans, "Memory-based language processing", Cambridge University Press, 2005). Van den Bosch currently works on memory-based language models and machine translation.

Walter Daelemans is professor of computational linguistics at the University of Antwerp and director of the CNTS Language Technology Group. He has published widely on Machine Learning applied to Natural Language Processing and NLP applications ranging from speech synthesis to text mining. He is currently associate editor of Research on Language and Computation.