

# 11. Interspeech 2008 Tutorials

Monday 22<sup>nd</sup> September

Rooms: BCEC PLAZA 1, 2, & 3

## Speech Recognition on Mobile Devices: Distributed and Embedded Solutions

### Tutorial AM1 (9:00-12:30) PLAZA 1

As mobile devices become pervasive and small, the design of efficient user interfaces is rapidly developing into a major issue. The expectation for speech-centric interfaces has stimulated a great interest in deploying automatic speech recognition (ASR) on devices like mobile phones and PDAs. The ability of state of the art large-vocabulary ASR system is in large part owed to recent increase in processing power of general purpose CPUs. One of the obstacles to implement ASR on mobile devices is the fact that embedded platforms lack in several dimensions behind the state of the art general CPUs, in particular: lower CPU clock, limited or missing floating point unit, smaller and slower memory and limited capability of development tools. To circumvent these restrictions, a great deal of effort has therefore been spent on enabling efficient ASR implementation on embedded platforms, e.g. through more efficient algorithms, more efficient implementation or approximation of algorithms, or reduction in the model complexity with an acceptable trade-off between accuracy and efficiency. The restrictions can also be largely bypassed from the architecture side: Distributed speech recognition (DSR) splits ASR processing into two parts, the client based front-end feature extraction and the server based back-end recognition. The relief of computational burden on mobile devices, however, comes at the cost of network deteriorations and additional components such as feature quantization and error concealment. Over the past decade, these areas have undergone substantial development. This tutorial will give the attendants a comprehensive view of the areas. The tutorial will consist of two major parts: distributed speech recognition covering such topics as feature extraction and quantization, error recovery and concealment, standards, voice only and multimodal systems, and applications; embedded speech recognition covering fixed-point arithmetic, algorithm optimisation for low computational complexity and low memory footprint, devices-specific issues, systems and applications. Distributed and embedded ASR systems are expected to co-exist in the future. This tutorial provides an up-to-date, unified introduction to these areas and working knowledge needed for research and application development. In this tutorial we will analyze main elements of embedded ASR and DSR systems and present several methods aimed to improve efficiency and robustness. This tutorial also discusses the pros and cons of different solutions. One needs to keep in mind that there is no silver-bullet solution; specific algorithmic improvements may best fit in specific applications only. The best solution can usually be achieved by algorithmic choices to maximize benefits for particular platform and task.

### Zheng-Hua Tan

[zt@es.aau.dk](mailto:zt@es.aau.dk)

Zheng-Hua is an Associate Professor in the Department of Electronic Systems at Aalborg University (AAU), Denmark. He received BS and MS degrees in electrical engineering from Hunan University, China, in 1990 and 1996, respectively, and a PhD degree in electronic engineering from Shanghai Jiao Tong University, China, in 1999. He was a postdoctoral fellow in the Department of Computer Science at Korea Advanced Institute of Science and Technology (KAIST), Korea. He was also an Associate Professor in the Department of Electronic Engineering at Shanghai Jiao Tong University, China. His primary research interests are speech recognition over communication networks, robust speech recognition, and machine learning. He has published three books and more than 50 technical articles in various journals, international conferences and books. His recently edited book is entitled *Automatic Speech Recognition on Mobile Devices and over Communication Networks*, Springer-Verlag, 2008 (<http://asr.es.aau.dk/>). He has a long-time experience in teaching with a focus on speech processing, signal processing and machine learning. He is a member of ISCA and a senior member of the IEEE.

**Miroslav Novak**

[miroslav@us.ibm.com](mailto:miroslav@us.ibm.com)

Miroslav received the MS degree in electrical engineering from Czech Technical University, Czech Republic in 1992 and a PhD in electrical and computer engineering from Rutgers, the State University of New Jersey in 2002. He is currently a Research Staff Member at IBM T.J. Watson Research Center, where he has been a member of the Speech group since 1993. His primary research interests include large vocabulary-speech recognition and efficient algorithms for speech recognition.

**History, Hardware and Sound Processing for the Bionic Ear****Tutorial AM2 (9:00-12:30) PLAZA 2**

The Bionic Ear (cochlear implant) has been implanted in tens of thousands of people throughout the world. However, the benefit received from the implant varies considerably. Some people can obtain near perfect perception of speech in quiet situations while others struggle. This tutorial takes a walk through the history of the development of the bionic ear, discussing the hardware and signal processing techniques along the way. The challenge now is to provide improved speech perception in noise and to improve music and pitch perception; current research directions in each of these areas will be discussed.

**David Grayden**

[grayden@unimelb.edu.au](mailto:grayden@unimelb.edu.au)

David obtained degrees in Electrical & Electronic Engineering and Computer Science followed by a PhD at the University of Melbourne. From 1997 to 2006 he worked as a Research Fellow and Senior Research Fellow at The Bionic Ear Institute, focussing on speech processing for cochlear implants and learning in neural systems. He then moved to the Department of Electrical and Electronic Engineering at the University of Melbourne where he undertakes research in Neuroengineering, including cochlear implants, retinal implants, epilepsy and neural modelling. He is currently secretary of the Australasian Speech Science and Technology Association (ASSTA).

**Modelling speech production : Insights from speech errors and neurogenic speech disorders.****Tutorial AM3 (9:00-12:30) PLAZA 3**

Speech production is a task-oriented process with multiple levels of organization transforming intent to communicate into coherent and perceptually identifiable meaningful sequences. Speaking is certainly one of the most complex skills that human beings perform. We have gained over the last 30 years a relatively good knowledge of the observable speech gestures at the periphery of the speech production system, but we have still a poor understanding of the way the brain initiates and controls these processes. Our knowledge of how concepts, thoughts and ideas are converted into movements of the speech apparatus remains largely based on inferences from speech errors and from production studies of other types of skilled movements. A crucial problem for speech production theories lies in the apparent dichotomy between the symbolic representation of an utterance and the physical aspects of speech production. Whereas the former is described as a system of abstract, invariant, discrete units, the latter is manifested as a sequence of context sensitive continuous overlapping patterns of articulatory movements resulting in a variable acoustic signal. The apparent translation process responsible for most of the observed variability is usually defined as coarticulation. This concept refers to the fact that at any given point during an utterance, the influences of gestures associated with several neighbouring segments can generally be observed in the acoustic and articulatory patterns of speech. There is however no consensus on what the nature of coarticulation is. The following questions remain largely unanswered by current speech production theories:

- What is the nature of the representation generated by the motor system?
- Do programming units and execution units differ and how?
- What signals to the motor system that the intended speech gesture has been achieved?

Errors in any system can have a tremendous explanatory value. A relatively unexploited source of information about normal speech production is speech disorders. Speech errors can be seen as unintentional

departures from what was meant to be said. In this tutorial, we will attempt to show how the investigation of speech disorders can give some insight into the processing stages of speech production. After a brief review of the main neuro-motor aspects of speech production, we will examine the rationale for the study of coarticulatory processes in neurogenic disorders such as apraxia and dysarthria. Methodological issues which are very crucial in particular when examining disorderd speech will be discussed. We will then indicate the benefit that could be gained from a cross-language approach to these processes for the development of a more comprehensive model of coarticulation.

### **Alain Marchal**

[Alain.Marchal@univ-provence.fr](mailto:Alain.Marchal@univ-provence.fr)

Alain studied linguistics and phonetics at the University of Nancy, France, before joining the University of Montréal, Canada, where he started his career as professor in the Linguistics department. He is now a senior researcher at the French National Research Centre, (CNRS) laboratory "Parole et langage" in Aix-en-Provence. He has developed a broad research program on a range of issues dealing with speech production, cognitive processing and speech quality assessment. Regarding speech production, he has contributed to a number of studies relating to respiration for speech, including large scale data collection of aerodynamic data such as intra-oral air pressure, oral and nasal airflow and subglottal pressure. His work has focussed for the last 20 years on coarticulation and he was coordinator of the large cross-language investigation "ACCOR". He is presently leading a European Research Council research project on coarticulatory processes in apraxic and dysarthric speech.

## **Speech Recognition by Mind and Machine**

### **Tutorial PM1 (13:30-17:00) PLAZA 1**

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 and language comprehension will be discussed alongside algorithms for automatic speech recognition, 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.

Topics to be covered include:

- speech recognition & interpretation
- discourse & dialogue
- plasticity and adaptation to speaker variability
- towards a unified view

Overall aims of the tutorial:

- to bridge the gap between human and machine spoken language recognition
- to provide a general overview of the current research issues in spoken language recognition by mind and by machine

Learning outcomes:

- a basic understanding of the similarities and differences between human and machine spoken language recognition
- an awareness of the key research issues

### **Anne Cutler**

[Anne.Cutler@mpi.nl](mailto:Anne.Cutler@mpi.nl)

Anne is a director of the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, Professor of Comparative Psycholinguistics at the Radboud University Nijmegen, and, since 2006, part-time Professor in MARCS Auditory Laboratories at the University of Western Sydney. Originally from Australia, Professor Cutler completed her undergraduate work and Masters of Arts (in German linguistics) at Melbourne University, and a PhD in Psycholinguistics at the University of Texas before taking up successive positions at MIT, Sussex, the Medical Research Council Applied Psychology Unit, Cambridge, UK, and, in 1993, as Director of the Max Planck Institute for Psycholinguistics in Nijmegen. 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. Professor Cutler is an internationally acclaimed researcher with over 100 refereed journal articles and more than 200 other publications, has supervised 26 PhDs to completion, and is the recipient of many prizes and awards, including the Spinoza Prize in 1999 from the Nederlandse Organisatie voor Wetenschappelijk Onderzoek.

### **Roger Moore**

[r.k.moore@dcs.shef.ac.uk](mailto:r.k.moore@dcs.shef.ac.uk)

Roger studied Computer and Communications Engineering at the University of Essex and was awarded the BA (Hons.) degree in 1973. He subsequently received the MSc and PhD degrees from the same university in 1975 and 1977 respectively, both theses being on the topic of automatic speech recognition. After a period of post-doctoral research in the Phonetics Department at University College London, he was head-hunted in 1980 to establish a speech recognition research team at the Royal Signals and Radar Establishment in Malvern. In 1985 he became head of the newly created Speech Research Unit and subsequently rose to the position of Senior Fellow (Deputy Chief Scientific Officer – Individual Merit) in the Defence and Evaluation Research Agency. Following the privatisation of the SRU in 1999, he continued to provide the technical lead as Chief Scientific Officer at 20/20 Speech Ltd. - a joint venture company between DERA (now QinetiQ) and NXT plc. Since 2004 he has been Professor of Spoken Language in the Speech and Hearing Research Group at Sheffield University. Prof. Moore has authored and co-authored over 100 scientific publications in the general area of speech technology applications, algorithms and assessment. He is a Fellow of the Institute of Acoustics and a Visiting Professor in the Department of Phonetics and Linguistics at University College London. He is also a member of the editorial/advisory boards for the scientific journals 'Computer Speech and Language' and 'Speech Communication'. He is past Chairman of the 'European Expert Advisory Group on Language Engineering Standards' (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".

## **Forensic Speaker Comparison – Likelihood Ratios – As Not Seen on TV**

### **Tutorial PM2 (13:30-17:00) PLAZA 2**

In the movies and on television a forensic scientist is given a recording from a wire tap and a recording of a suspect. The forensic scientist's computer monitor shows two waveforms. The waveforms move relative to each other and within a few seconds one matches the other perfectly. The forensic scientist then testifies in court that the voices on the two recordings are one and the same. Reality is quite different from movies and television both in terms of signal processing and acoustic analysis, and in terms of how forensic evidence can be presented in court (we make no claims as to whether real-life forensic scientists are better-looking than their fictional counterparts). This tutorial serves as an introduction as to how recordings of speakers can be compared so as to produce likelihood ratios, which are now widely recognised as the logically and legally

correct presentation of the strength of forensic evidence. A likelihood ratio can be used to express the probability of observing the acoustic difference between voice recordings under the hypothesis that they were produced by the same speaker versus under the hypothesis that they were produced by different speakers (probabilities of evidence given hypotheses). This is quite different from speaker recognition systems which identify the questioned speaker as one of a number of enrolled speakers or speaker verification systems which determine whether the questioned speaker is within some threshold of difference from one of the enrolled speaker (both probability of hypothesis given evidence, and both based on closed-set decisions).

Format: The first part of the tutorial explains the theory behind the likelihood-ratio approach, and the second part provides detailed examples of its application to speech data extracted via acoustic-phonetic and automatic procedures.

Learning goals: Participants will gain a basic understanding of likelihood ratios, and why the likelihood-ratio approach is the correct method for the presentation of forensic evidence. They will also gain sufficient insight into the application of likelihood-ratio-based forensic speaker comparison so as to be able to be applying it to their own speech data. They will also be made aware of some of the potential problems in dealing with forensic speech data.

### **Yuko Kinoshita**

[Yuko.Kinoshita@canberra.edu.au](mailto:Yuko.Kinoshita@canberra.edu.au)

Yuko is the Convener of the Japanese Program in the School of Languages and International Studies at the University of Canberra. She holds a doctorate in forensic speaker recognition from the Australian National University. Her 2001 Ph.D thesis *Testing realistic forensic speaker identification in Japanese: A likelihood-ratio based approach using formants* was the first study to successfully implement a forensically-motivated likelihood ratio-based approach to discrimination with formants. She has been a visiting fellow at the National Research Institute of Police Science in Japan, collaborating with Japanese forensic speaker identification experts. A member of the International Association of Forensic Phonetics and Acoustics, she continues to research forensic speaker recognition and has published several papers on it. Her research interests also include linguistic phonetics, and she also holds an M.A in phonetics from the Australian National University in which she conducted research on dialectal variation in Chinese.

### **Geoffrey Stewart Morrison**

[Yuko.Kinoshita@canberra.edu.au](mailto:Yuko.Kinoshita@canberra.edu.au)

Geoff obtained his PhD from the Department of Linguistics, University of Alberta, in 2006. He is currently Research Associate in Forensic Speaker Recognition at the School of Language Studies, Australian National University. He has recently given lectures and tutorials on likelihood-ratio-based forensic speaker comparison at universities in China, Spain, and the United Kingdom. He is the organiser of the Interspeech2008 Special Session: Forensic Speaker Recognition – Traditional and Automatic Approaches.

### **Daniel Ramos**

[daniel.ramos@uam.es](mailto:daniel.ramos@uam.es)

Daniel has been at the Universidad Autónoma de Madrid, Spain, since 2004 where his current position is Profesor Ayudante. He received a PhD in Telecommunication Engineering in 2007 from the Universidad Autónoma de Madrid and an MSc in Electrical Engineering in 2001 from the Universidad Politécnica de Madrid. Between these two degrees he worked for three years in industry. His research interests are focused on speech and signal processing, pattern recognition, biometrics, speaker and language recognition, statistics and evaluation of forensic evidence. He has participated in the development of the ATVS speaker and language recognition systems which have competed in several NIST evaluations since 2004. He has been a member of organizing and scientific committees for a number of conferences including the Odyssey 2004 and 2008 Speaker and Language Recognition Workshops. He was awarded the IBM Research Best Student Paper Award at Odyssey 2006.

## **Voice User Interface Design in Commercial Speech Applications**

### **Tutorial PM3 (13:30-17:00) PLAZA 3**

This tutorial originates from my experience in trying to teach undergraduate students what they might need to know in order to work in the speech recognition industry. In the six years I have taught that course, I have always been struck by the difference between the kinds of questions typically explored in dialog systems projects in research laboratories, and the challenges faced by those attempting to develop robust applications that can be used over the phone by real people aiming to complete some transaction or obtain some information. It can often seem that these are two completely unrelated worlds; and yet I'm convinced they shouldn't be. The intended audience of this tutorial is those who come at dialog systems from a research perspective, but may not be so aware of what is relevant when building real applications today. The aim of the tutorial is to provide a picture of the kinds of issues faced in building deployable applications, with the hope of bringing closer together the concerns of researchers and application developers.

#### **Robert Dale**

[rdale@ics.mq.edu.au](mailto:rdale@ics.mq.edu.au)

Robert Dale is a professor in the Department of Computing at Macquarie University, Director of Macquarie's Centre for Language Technology, and Convener of HCSNet, the ARC Research Network in Human Communication Science. He holds a PhD in Computational Linguistics (1989) from the University of Edinburgh. His research interests cover both speech and text processing, with a particular focus on natural language generation. He has frequently served as a consultant on the deployment of voice recognition applications in Australia, and he also teaches an undergraduate course in spoken language dialog systems.