



MAIN

Invitation

Programs

Information

Exhibition

Sponsors

Search

## Tutorial

Tutorial Main

### T1: Conversational and Multimodal Interaction

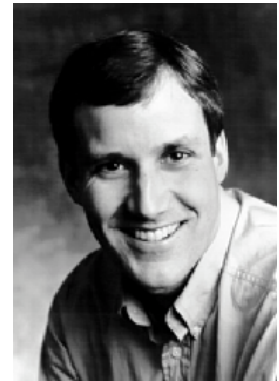
James Glass, MIT

#### *Overview*

Conversational interaction between humans and computers enables natural, flexible, and efficient communication for information access or problem solving. This tutorial will describe the technologies required to create a conversational interface, will survey the different approaches being explored, will show demonstrations of several existing conversational and multimodal interfaces, and discuss ongoing research challenges. In the latter part of the tutorial we will discuss the issue of portability and prototyping, and develop a prototype conversational interface using a web-based toolkit.

#### *Presenter*

**James Glass** is a Principal Research Scientist, Head of the Spoken Language Systems Group in the MIT Computer Science and Artificial Intelligence Laboratory, and a member of the Speech and Hearing Bioscience and Technology Program of the Harvard-MIT Division of Health, Sciences, and Technology. He received his S.M. and Ph.D. degrees from MIT in 1985 and 1988, respectively. His primary research interests are in the area of speech communication and human-computer interaction, centered on automatic speech recognition and spoken language understanding. He has lectured, taught courses, and has many publications in these areas. He has been a member of the IEEE Acoustics, Speech, and Signal Processing, Speech Technical Committee and an associate editor for the IEEE Transactions on Speech and Audio Processing.





MAIN

Invitation

Programs

Information

Exhibition

Sponsors

Search

## Tutorial

Tutorial Main

### T2: Building Spoken Language Systems: From Theory to Practice

Alex Acero, Microsoft

#### Overview

We will cover the theoretical foundation behind spoken language processing and give examples of how to design real systems. Time will be allowed for students to experiment with such tools. APIs and markup languages available today can reduce the burden of a user trying to build such systems but there's still significant work required to do so.

1. APIs and Markup languages
2. Voice recognition: using CFG and n-grams
3. Voice output: using TTS and prompts
4. Text and spoken language understanding: rule-based and machine learning based
5. Dialog: system initiative, user initiative and mixed initiative
6. Tools/examples of how to build real applications

#### Presenter

Before joining Microsoft in 1994, **Alex Acero** worked in the speech groups of Apple Computer and Telefonica Investigacion y Desarrollo. He received a Ph.D. from Carnegie Mellon University in 1990, a Master's from Rice University in 1987 and an engineering degree from the Universidad Politecnica de Madrid in 1985, all in Electrical Engineering. He is also an affiliate Professor of Electrical Engineering at University of Washington.



#### Professional activities

1. Fellow of IEEE.
2. Board of Governors of IEEE Signal Processing Society: Member-at-large (2004-2005)
3. Chair (2000-2002) and member (1996-2000) of the Speech Technical Committee of the IEEE Signal Processing Society.
4. General Co-Chair of the 2001 IEEE Workshop on Automatic Speech Recognition and Understanding.
5. Sponsorship Chair of the 1999 IEEE Workshop on Automatic Speech Recognition and Understanding.
6. Publications Chair of ICASSP98.
7. Associate Editor for IEEE Signal Processing Letters (2003-present).
8. Associate Editor for Computer, Speech and Language (1994-present).
9. Reviewer for numerous conferences and journals such as ICASSP, ICSLP, Eurospeech, IEEE Transactions on Speech and Audio Processing, IEEE Spectrum and Speech Communication.
10. Tutorial on Spoken Language Processing, with M. Rahim, at ICASSP 2002.



MAIN

Invitation

Programs

Information

Exhibition

Sponsors

Search

## Tutorial

Tutorial Main

### T3: Graphical Models in Speech and Language Research

Jeffrey A. Bilmes, University of Washington

#### Overview

Graphical models are a promising paradigm for studying both existing and novel techniques in speech and language processing. This tutorial will show how the assumptions behind many pattern recognition techniques commonly used for speech and language can be described by a graph, and how the algorithms associated with such models are often just instances of applying graphical inference algorithms. This includes (most famously) the Baum-Welch algorithm for HMMs. But graphs and their algorithms generalize many other techniques as well, including Gaussian densities (which can be seen either as Bayesian networks, or undirected models), mixture models, decision trees, factor analyzers, principle component analysis, linear discriminant analysis (and its generalizations), Kalman filters, and most language models. The same general inference algorithms, however, work on any graph. It will be demonstrated how, in this paradigm, one can rapidly apply a new technique to speech and language with relatively little effort.

The tutorial will show that many advanced models proposed for speech and language processing can be described by a graph. This includes segment models, HMM decomposition, speaker adaptation, IBM machine translation models, etc. Moreover, the tutorial will survey a number of speech recognition techniques that were born directly out of the graphical-model paradigm, such as Buried Markov models, structural discriminability, and explicit graphical structures for speech recognition. Many of the practical issues in using such models in speech and language will be discussed. The tutorial will also overview methods to learn graph structure from data. Finally, the tutorial will provide an overview of the graphical models toolkit (GMTK), a software system for rapid deployment of graphical modeling of speech, language, and time series.

In taking this tutorial, it will become apparent that the space of models describable by graphs is extremely large --- it will seem quite probable that a thorough exploration of the space of models easily describable by a graph could ultimately yield a technique that can perform far better than the HMM. Taking this tutorial will make it easier to begin such an exploration.

#### Presenter

**Jeff Bilmes** is an assistant professor at the Department of Electrical Engineering at the University of Washington, Seattle, and is also an adjunct assistant professor in Linguistics, and in Computer Science and Engineering. He cofounded the Signal, Speech, and Language Interpretation Laboratory at the university. He received a Masters degree from MIT, and a Ph.D. in Computer Science at the University of California in Berkeley in 1999. In the summer of 2001, he co-lead a Johns Hopkins University summer workshop on applying graphical models to speech recognition.



In 2002, he participated in another JHU workshop on Arabic speech recognition which lead to factored language models and generalized backoff (graphical generalizations of typical language models). While at JHU, he also gave a 6-week short course at JHU on graphical models theory. In the summer of 2003, he presented a tutorial [2] at the the world's premier graphical models conference (Uncertainty in Artificial Intelligence, UAI'03) on the application of graphical techniques to speech and language. He has taught many courses at the university of Washington on signal processing, information theory, and graphical models in pattern recognition. He was an associate editor of the IEEE Transactions on Multimedia, is a 2001 recipient of the NSF CAREER award, and a 2001 CRA digital government fellow.



MAIN

Invitation

Programs

Information

Exhibition

Sponsors

Search

## Tutorial

Tutorial Main

### T4: Adaptive Acoustic Modeling for Speech Recognition

Olivier Siohan, IBM and Frank Soong, ATR

#### Overview

Despite significant progress in the past several years, automatic speech recognition (ASR) systems remain sensitive to a wide range of variabilities between training and operating conditions that can significantly upset recognition performance. Such variabilities typically include speaker-related phenomena (eg. speaking style, speaking rate, accented speech), channel-related variations (eg. transducer and transmission channels), environment-related distortions (background noise, reverberation), and task-related variabilities. Adaptation techniques attempt to increase ASR systems robustness by limiting their sensitivity to such sources of mismatch, and are therefore of utmost interest for both theoretical and practical purposes.

In this tutorial, we review acoustic model adaptation techniques. We distinguish two broad families of techniques: first, direct adaptation, which adapts model parameters typically through Bayesian estimation; second, indirect adaptation, which adapts model parameters through shared transformations. We also describe how adaptation can be reformulated under robust decision rules such as Minimax and Bayesian predictive classification. Finally, we present some special acoustic modeling techniques allowing for fast adaptation, such as eigenvoices and cluster-adaptive training. We describe the theoretical rationale underlying these various techniques and illustrate their applications to several classes of problems such as speaker and noise adaptation on various tasks.

#### Presenter

**Olivier Siohan** received the M.S. degree in Electrical Engineering in 1991 from Supelec, an Electrical Engineering Institute in France. He earned the M.S. degree and the Ph.D. degree in Computer Science both from the University of Nancy, France, in 1991 and 1995, respectively. In 1995, he joined AT&T Bell Laboratories on a post-doctoral position to carry out research on robust speech recognition. From 1996 to 1998, he was with AT&T Labs working on speaker recognition. From 1998 to 2002, he was a member of technical staff at Lucent Technologies - Bell Labs, where he led a broadcast news speech recognition project. In 2003, he moved to IBM T.J. Watson Research Center as a research staff member. His major research interests are speech and speaker recognition, acoustic modeling, speaker adaptation, noise robustness. He is the author of over 50 articles in these areas



**Frank K. Soong** has been an invited researcher at the Advanced Telecommunication Research Institute (ATR), Kyoto, Japan since 2002, after his retirement from Bell Labs Research in 2001. He joined the Acoustics and Speech Research at Bell Labs in 1982 and retired as a Distinguished Member of Technical Staff. In his 20 years of research at Bell Labs, he has worked on a wide variety of speech and acoustics topics, including: speech coding, speech and speaker recognition, fast N-best search for LVCSR, discriminative acoustic model training, echo cancellation, room de-reverberation and robust hands-free, human-machine interface. His current interests are on modeling stochastic processes, assessing speech recognition output for improving statistical machine translation performance. He was a co-recipient of the Lucent President Golden Award for developing the Bell Labs Automatic Speech Recognition (BLASR) software package. .

His invention, the tree-trellis based fast search, forms the core algorithm of the most popular, large vocabulary, speech recognition software, JULIUS, in Japan. He served the IEEE Speech Technical Committee both as a committee member and an associate editor of the Transactions on Speech and Audio Processing. He co-chaired the IEEE International Arden House ASR Workshop in 1991. He visited NTT Human Interface Labs in Japan in 1987 for one year where he worked on parsimonious speech analysis/synthesis for a very low bit rate speech coding applications. He is also currently a visiting Professor at the Electronic Engineering Department of the Chinese University of HongKong, Satin, Hong Kong. He holds BS, MS and Ph. D degrees, all in EE, from the National Taiwan University, the University of Rhode Island and Stanford University, respectively. He published extensively and has more than 100 papers in the speech processing field.



MAIN

Invitation

Programs

Information

Exhibition

Sponsors

Search

## Tutorial

Tutorial Main

### T5: Auditory Scene Analysis

Shihab A. Shamma, University of Maryland

#### Overview

1. Introduction: What is ASR and what function it serves in sound perception and speech intelligibility?
2. Basic Perceptual attributes of complex sounds: Loudness, Pitch, Timbre, Location
  - a. What are the main physical correlates of the different percepts?
  - b. How are these percepts quantitatively measured and explored psychoacoustically?
3. Anatomy and functional organization of the auditory nervous system
  - a. Basic processing stages in the early and central auditory pathway
  - b. Cell types and the concept of parallel pathways
4. Auditory processing of sound along the auditory pathway
  - a. Where are the physical cues extracted?
  - b. What basic transformations known to occur at each auditory stage?
5. The basic findings of Auditory Scene Analysis (ASR): Grouping and Streaming
  - a. Spectral and temporal grouping: Harmonicity, regularity, and temporal onsets
  - b. Temporal streaming: Role of timbre and location
6. ASR in other animals: Fish, birds, and mammals
7. Physiological correlates of ASR
  - a. Responses in single cells
  - b. ASR in recordings with fMRI, MEG, and the MMN
8. Quantitative models of ASR phenomena
9. ASR in audio speech recognition

Tutorial will include numerous audio demonstrations and a hands-on exercise with a MATLAB auditory toolbox mimicking processing at various auditory stages.

#### Presenter

**Shihab A. Shamma** obtained his Ph.D. degree in electrical engineering from Stanford University in 1980. He joined the Department of Electrical Engineering at the University of Maryland, in 1984, where his research has dealt with issues in computational neuroscience and the development of microsensor systems for experimental research and neural prostheses. Primary focus has been on uncovering the computational principles underlying the processing and recognition of complex sounds (speech and music) in the auditory system, and the relationship between auditory and visual processing. Other research include the development of photolithographic microelectrode arrays for recording and stimulation of neural signals, VLSI implementations of auditory processing algorithms, and development of algorithms for the detection, classification, and analysis of neural activity from multiple simultaneous sources.





MAIN

Invitation

Programs

Information

Exhibition

Sponsors

Search

### T6: Signal Separation and Speech Enhancement

Chang D. Yoo, KAIST and

Te-Won Lee, Univ. of California, San Diego

#### *Overview*

The problems of signal separation and speech enhancement arise in variety of different scenarios and can be approached in a number of different ways with different objectives. In the tutorial, signal separation based on independent component analysis (ICA) will be covered including multichannel blind deconvolution for speech separation in acoustic environment. The tutorial will also cover various single channel enhancement methods for improving the human perception of speech that has been degraded by additive noise. The tutorial will cover various enhancement algorithms by classifying them into either transform based approach and model based approach. The tutorial will also cover both psycho-acoustical models and various methods for estimating the noise statistics relevant for speech enhancement.

#### *Presenter*

**Chang D. Yoo** is an associate professor in the EECS department at the Korea Advanced Institute of Science and Technology. He received his BS from CALTECH in 1986, MS from Cornell University in 1988 and Ph D from MIT in 1996. His primary interests are in the applications of digital signal processing theory in speech, image and audio. He is the author of over 40 articles in these areas. He is a member of Tau Beta Pi and Sigma Xi.



**Te-Won Lee** is an associate research professor at the Institute for Neural Computation at the University of California San Diego and a collaborating professor in the Biosystems Department at the Korea Advanced Institute of Science and Technology (KAIST). Dr. Lee received his diploma degree in March 1995 and his Ph.D. degree in October 1997 with summa cum laude in electrical engineering from the University of Technology Berlin. He was a Max-Planck Institute fellow (1995-97) and a visiting professor KAIST in 1999. During his studies he received the Erwin-Stephan prize for excellent studies from the University of Technology Berlin and the Carl-Ramhauser prize for excellent dissertations from the Daimler-Chrysler Corporation. Dr. Lee's research interests include machine learning algorithms with applications in signal and image processing and inference. Recently, he has worked on variation Bayesian methods for independent component analysis, algorithms for speech enhancement and recognition, models for computational vision, and classification algorithms for medical informatics.