Sunday Overview

07:30 - 17:00	Speaker Check-In — *Plaza Suite, Plaza Level*			
08:00 - 18:00	Conference Registration — *Grand Ballroom Foyer, Lower Level*			
08:45 - 12:00	**Sun.T1a** Domain Adaptation in Machine Learning and Speech Recognition	**Sun.T1b** Privacy-Preserving Speech Processing	**Sun.T1c** Uncertainty Handling for Environment-Robust Speech Recognition	**Sun.T1d** Voice, Speech, and Language Pathology: Biological Basis, Diagnosis, and Challenges
	10:00 - 10:30 Refreshment Break — *Grand Ballroom Foyer*			
	Grand Ballroom II	Galleria North	Galleria South	Grand Ballroom B-C
12:00 - 13:00	**Lunch Break**			
13:00 - 16:15	**Sun.T2a** Computer-assisted Language Learning (CALL) Systems	**Sun.T2b** Computational Paralinguistics: Emotion, Affect, and Personality in Speech and Language Processing	**Sun.T2c** From Stationary to Adaptive Sinusoidal Modeling of Speech with Applications to Speech Technologies and Voice Function Assessment	**Sun.T2d** Topic Models for Acoustic Processing of Speech
	14:30 - 15:00 Refreshment Break — *Grand Ballroom Foyer*			
	Grand Ballroom II	Galleria North	Galleria South	Grand Ballroom B-C

T - Tutorial

Sun.T1a Domain Adaptation in Machine Learning and Speech Recognition

08:45 – 12:00 Grand Ballroom II
Fei Sha and Brian Kingsbury

Most learning algorithms for pattern recognition assume that the training data and test data come from the same distribution. While this assumption enables convenient theoretical analysis and controlled testing of algorithms, it rarely holds in practice. Speech processing systems, for example, must deal with significant variability in speakers, background noise, channel characteristics and higher-level effects such as drift in genre and conversation topics over time. Inevitably, real-world data exhibits variability that has not been captured in training sets.

The problem of a mismatch between the training and test distributions has also received a lot of research attention in other application areas, including text processing, computer vision, and bioinformatics. A plethora of techniques has been proposed; in the machine learning community they are known under the names of domain adaptation, covariate shift, sample selection bias, and transfer learning. In speech processing, many classical techniques can be seen as instances of domain adaptation, e.g., robustness and speaker adaptation in speech recognition, compensation for intersession variability in speaker identification, and adaptation of language models to new domains. While some techniques from speech processing exploit specific characteristics of speech signals, many techniques from the two communities share similar intuitions about how to model and overcome shifts in data distributions.

This tutorial aims to bridge the gap between the speech and machine learning communities, share research results, and foster and inspire collaboration for addressing the challenge of domain adaptation. While the presenters will employ specific examples of techniques of domain adaptation from concrete application areas such as speech, language processing, and computer vision, they will also focus on general frameworks and theoretical underpinnings, thus providing essential tools and ideas for attendees to address the challenge of domain adaptation in the broad sense, and to make fundamental contributions to the field.

Sun.T1b Privacy-Preserving Speech Processing

08:45 – 12:00 Galleria North
Madhu Shashanka and Bhiksha Raj

Speech is one of the most private form of personal communication, yet, current speech processing techniques are not designed to preserve speaker privacy and require complete access to the speech data. In this tutorial we study privacy-preserving techniques for speech processing applications focusing on speaker verification and identification. A speaker verification system uses the speech input to authenticate the user. We discuss privacy-preserving speaker verification, where the system is able to perform authentication without observing the speech input provided by the user and the user does not observe the speech models used by the system. These privacy criteria are important in order to prevent an adversary having unauthorized access to the user's client device or the system data from impersonating the user in another system. We develop two privacy-preserving algorithms for speaker verification: firstly, we use Gaussian mixture models (GMMs) and create a homomorphic encryption based protocol to evaluate GMMs over private data. Secondly, we apply locality sensitive hashing (LSH) and one-way cryptographic functions to reduce the speaker verification problem to private string comparison.

Speaker identification is a related problem where we are interested in identifying the speaker among a given set of speakers best corresponding to a given speech sample. This task finds applications in surveillance applications, where a security agency such as the police has access to a speaker models for individuals, e.g., a set of criminals it is interested in monitoring and an independent party such as a phone company might have access to the phone conversations. The agency is interested in identifying the speaker participating in a given phone conversation among its set of speakers. The agency can demand the complete recording from the phone company if it has a warrant for that person. By using a privacy-preserving speaker identification system, the phone company can provide the privacy guarantee to its subscribers that the agency will not be able to obtain any phone conversation for the speakers that are not under surveillance. Similarly, the agency does not need to send the list of speakers under surveillance to the phone company. Speaker identification can be considered to be an extension of speaker verification to the multi-class setting. We extend the GMM-based and LSH-based approaches to create analogous privacy-preserving speaker identification frameworks.

As this is an emerging field that has not yet been exhaustively studied, we will have the opportunity to cover its theory from first principles. Due to that, the target audience of this tutorial can be very wide and will not be expected to have any prior experience in this area. The target participants of this tutorial are speech researchers without any background in privacy, We hope that this tutorial will enable these researchers to develop privacy-preserving variants of their speech processing algorithms, and help foster cross-pollination between these two research areas.

Sun.T1c Uncertainty Handling for Environment-Robust Speech Recognition

08:45 – 12:00 Galleria South
Ramon F. Astudillo, Emmanuel Vincent and Li Deng

In today's world, where mobile computing is more prevalent than any time in the history, automatic speech recognition (ASR) in environments with non-stationary noise remains a very challenging problem. The ubiquity of speech applications for hand-held devices, best exemplified by the recent success of personal assistant Siri on the iPhone 4S, requires ASR systems to deal with a wide variety of acoustic environments. Furthermore, the short interaction times left very little information for ASR systems to adapt.

While speech enhancement is typically carried out in the short-time Fourier (STFT) domain, where speech corruption is easier to model, ASR operates typically on nonlinearly transformed features such as MFCC, which result in more compact features and models. In recent years, a breed of robust ASR methods have arisen that exploit both the advantages of the STFT and the nonlinear feature domains by employing the notion of uncertainty propagation/decoding. These techniques estimate an uncertain description of speech in the feature domain which accounts for the effect of the distortions in the STFT domain or the residual noise after speech enhancement. This uncertain description of the features is then used to dynamically compensate the ASR model and thus attain robust ASR with lower computational loads than classical model-based compensation. Furthermore, uncertainty propagation/decoding provides a formal framework allowing the incorporation of expertise from the speech enhancement field into robust ASR. The field is also currently in expansion with promising directions including model training with uncertain data or integration with multichannel and multi-modal algorithms.

This tutorial will introduce the topic of uncertainty handling for robust ASR, review the latest trends and discuss future development directions. The tutorial will cover how an uncertain description of the speech features can be determined by exploiting STFT domain information and how uncertainty can be integrated into an ASR model. Both feature-domain and STFT-domain methods to determine uncertainty will be considered. Regarding feature domain, the latest developments around the ALGOQUIN model will be introduced. A general taxonomy of feature and model-domain approaches will also be provided. Regarding STFT domain, the STFT-Uncertainty Propagation approach, integrating STFT speech enhancement and robust ASR, will be presented. Along with uncertainty decoding and propagation approaches, recent progress in extending the use of uncertainty for robust recognition to training with uncertain data will also be presented. Other novel approaches like improved Bayesian estimation of STFT uncertainties will also be addressed. The tutorial will finish with an analysis of future perspectives.

Sun.T1d Voice, Speech, and Language Pathology: Biological Basis, Diagnosis, and Challenges

08:45 – 12:00 Grand Ballroom B-C
Maria Schuster, Shannon Kraft and Izhak Shafran

This is a tutorial presented mostly by two clinicians who have been invited to share their insights from their practices. Between them, their expertise covers a range of communication disorders. This tutorial does not assume medical background and is aimed at the typical InterSpeech attendees working in computational aspects of speech and language processing who want to learn about the pathologies in verbal communication. Communication disorders are among the most common disabilities from childhood to the elderly, which can be treated effectively by therapy in many cases. Verbal or oral communication involves several key components and disorders are often grouped according to the afflicted components — voice (lungs, vocal folds, respiration), speech (articulation, motor control), and language (memory, lexical access, semantic organization).

Topics of this tutorial include the brief biological background, etiology, phenomenology, and diagnostic methods of verbal communication disorders. Differences and similarities between different disorders are highlighted and the main characteristics for each disorder will be described. The topics will include language impairment, speech disorders to brain, nerve or morphologic alterations of the mouth and pharynx, and voice disorders. These topics are described both in children and adults and as congenital and acquired disorders.

This tutorial will examine and compare methods of clinical assessment. Main topics herein are methods commonly used in clinical practice, methods described in medical scientific literature, and internationally consented speech evaluation procedures. In practice, perceptual evaluation is an important tool for assessment and clinicians rely on their experience to efficiently sort through differentiate diagnosis and identify the malady. However, perceptual assessment is less reliable, often takes into account only a few characteristics and is hard to quantify. One of the aims of the tutorial is to explore the opportunities in diagnostics and monitoring of disorders where current methods are inadequate and there may be opportunities to help clinicians with technical innovations and tools.

Sun.T2a Computer-assisted Language Learning (CALL) Systems

13:00–16:15 Grand Ballroom II
Tatsuya Kawahara and Nobuaki Minematsu

Computer-assisted language learning (CALL) provides an effective learning environment so that students can practice in an interactive manner using multi-media content, either with the supervision of teachers or on their own pace in self-learning. The advancement of speech and language technologies has opened new perspectives on CALL systems, such as automatic pronunciation assessment and simulated conversational-style lessons. CALL is also regarded as one of new and promising applications of speech analysis, recognition and synthesis. CALL covers a variety of aspects including segmental, prosodic and lexical features. Modeling non-native speech to correctly segment/recognize utterances while detecting errors included in them poses a number of challenges in speech processing. Assessing intelligibility of non-native speech or proficiency of non-native speakers is also an important issue. In this tutorial, we will give an overview on these issues and current solutions. The tutorial is mainly targeted for speech researchers and engineers interested in CALL, but also for those engaged in language teaching or learning technology.

First we review speech recognition technologies for pronunciation learning, specifically pronunciation evaluation and error detection. Statistical approaches to these problems are formulated, and then acoustic and pronunciation modeling of non-native speech is described. Unlike the conventional non-native speech recognition, error detection capability is required in CALL, thus an effective error prediction scheme is vitally important. Next, we address prosodic modeling and evaluation, such as duration, stress and tones, and then the use of speech synthesis technologies including re-synthesis and morphing.

After the review of basic component technologies, we introduce a number of practical CALL systems which have been developed as commercial products or deployed in classrooms, including those in our universities. Majority of them focus on learning English as a second language (ESL), but some deal with other languages such as Japanese and Chinese. We also review databases of non-native speech, which are necessary to develop CALL systems.

Sun.T2b Computational Paralinguistics: Emotion, Affect, and Personality in Speech and Language Processing

13:00–16:15 Galleria North
Björn Schuller and Anton Batliner

When there is spoken or written language, there is paralinguistic information. Social awareness of this information on affect, personality, and other states and traits can be expected to be an integral factor in future multi-modal user interfaces and large scale retrieval systems on speech, text, and audiovisual databases. In this vein, the tutorial aims at covering the young fields of automatic recognition of human affect, emotion, personality, and speaker states and traits as reflected in one's speech or written text. It will first introduce the general topic covering a short history of the field as well as definitions of and examples for basic terms. Next, we will contrast the formal aspects of the linguistic code with the formal aspects of the "non-linguistic", i.e., paralinguistic, code. The part on functional aspects will cover the most important phenomena such as biological (age, gender) and cultural (regional/foreign accent) trait primitives, as well as the "big" topics personality, emotion, and pathology. Moreover, for these "big" topics, we will address theoretical foundations as well as fundamental aspects (e.g., categorical vs. dimensional modeling). This will be followed by corpus engineering including annotation and selection of units. Moreover, we will introduce important corpora and benchmarks and synthesized speech for training and semi-supervised learning. Next will be signal processing and machine learning aspects including pre-processing, feature extraction, and machine learning algorithms, followed by acoustic and linguistic analyses in isolation or combined within efficient and synergistic fusion. As for integration of the information in a system context, we will discuss standards for paralinguistic information encoding, error-prone prediction results and confidence measurement, real-time issues, application design, and real-life evaluation of systems. Finally, a practical "hands-on" part includes examples employing our open source "openEAR" toolkit for emotion and affect recognition and general speaker classification — the official feature and baseline toolkit as used in the series of Challenges the presenters co-organized at INTERSPEECH since 2009 — as well as data from these Challenges.

The objective of this tutorial is thus to give a comprehensive introduction and broad overview on recent algorithms and methodologies of "real-life" speech processing, focusing on paralinguistic aspects. We will present the facets and nuances that can be extracted from speaker states and states such as affect, emotion, personality, behavioral and social signals, including practical aspects as current datasets and research tools. While it will not be possible to discuss all aspects of "Computational Paralinguistics", a participant in this tutorial will gain all the skills needed to identify algorithms and tools for solving a particular problem from her/his field.

The main target audience is a broad group of scholars, practitioners, and experts in speech processing, natural language understanding or even human-computer-interaction: "real-life" speech touches any of these fields. The tutorial will assume very little knowledge of signal processing principles (so it will be suitable for the non-specialist), but it will cover many state-of-the-art subjects, so that also the specialist will find it interesting.

Sun.T2c From Stationary to Adaptive Sinusoidal Modeling of Speech with Applications to Speech Technologies and Voice Function Assessment

13:00–16:15 Galleria South
Yannis Stylianou

This tutorial will discuss a) the passage from the non-adaptive and stationary to the adaptive and non-stationary analysis of speech and b) the use of this speech analysis framework in the analysis of speech, focusing on the pathologic speech processing as well as its potential in speech technologies like speech synthesis and speech modifications. In the first part of the tutorial, novel algorithms for the adaptive speech analysis will be presented and how they are related to the well known sinusoidal representation as well as to non-linear frequency estimators like the Newton-Gauss. The second part will be dedicated to applications like tremor estimation, estimation of jitter and shimmer through a mathematical – sinusoidal based – description, objective evaluation of spasmodic dysphonia and vocal fatigue. Also, the potential that such a speech analysis framework may offer to well known speech technologies will be discussed. Care about the balance between theoretical and application aspects will be taken.

The main target audience of the suggested tutorial includes students, researchers, and engineers having specific interests in the recent developments of sinusoidal models, in frequency estimation, in non-linear speech signal processing, in speech synthesis and modifications, and in novel algorithms of signal processing applied in the domain of voice function assessment and pathologic voices.

Sun.T2d Topic Models for Acoustic Processing of Speech

13:00–16:15 Grand Ballroom B-C
Dr. Bhiksha Raj and Dr. Paris Smaragdisis

Topic models have recently been the center of a flurry of research activity and, in various forms, are the basis for many highly successful text and language tools (e.g. Google). Even though this technology has been initially developed for text applications and discrete data, it is not constrained to that domain and once expanded for use on time-series it can be quite a formidable tool for dealing with many of the major speech signal problems, especially these involving mixtures.

Sounds, particularly speech, are typically characterized through spectro-temporal representations such as short-time Fourier transforms and Mel-spectral representations. These representations naturally lend themselves to a histogram-based interpretation: the energy in any time-frequency bin for the signal is a scaled count of the number of quanta of energy in that frequency at that time. When abstracted, such a quanta-based representation instantly becomes indistinguishable from the histogram-based characterizations that un-derlie topic models and consequently much of the mathematical development that underlies topic models can also be employed to analyze and make highly useful inferences from the signals. By employing this model, various previously difficult-to-handle problems such as signal de-noising, bandwidth expansion, analysis of mixed signals, signal prediction, signal tracking, de-reverberation etc. now become easily tractable inference of additive components. This approach has increasingly become very visible in the signal processing field and, to date, has contributed to solutions which are very efficient and produce state-of-the-art results.

In this tutorial we will describe how topic models and their signal-specific extensions can be used to analyze and process speech. We will begin with the basics of latent variable multinomial decompositions, and work our way upwards through various higher-level models, their interpretations and extensions, and their relationship to other popular matrix decomposition techniques, computer vision methods, as well to the compressive sensing literature. We will show how this field combines elements from machine learning and signal processing to produce hybrid approaches to produce novel approaches (and solutions) to some of the hardest problems in speech processing. We will cover models that can be very effectively used for a large number of applications, ranging from signal separation, signal de-noising, speech recognition, pitch tracking, de-reverberation, audio/visual object extraction, user-assisted audio selection, echo cancellation, etc.

Because this is an emerging field that has not yet been exhaustively studied, we will have the opportunity to cover its theory from first principles. Due to that, the target audience of this tutorial can be very wide and will not be expected to have any prior experience in this area. Even though our target participant will be a signal-oriented researcher this tutorial will also help machine-learning and text/language-processing researchers see how their expertise can be used for many speech processing problems. We hope that this tutorial will help uncover some of the theoretical overlaps between these fields and help foster cross-pollination between these two types of participants.