INTERSPEECH 2014

CELEBRATING THE DIVERSITY OF SPOKEN LANGUAGES

14-18 SEPTEMBER 2014
SINGAPORE
MAX ATRIA@SINGAPORE EXPO

Tutorials

HTTP://WWW.INTERSPEECH2014.ORG

Tutorials

The INTERSPEECH 2014 tutorial committee chaired by Professor Eliathamby Ambikairajah is pleased to announce the following eight tutorials at the conference and will be offered on Sunday, 14 September 2014. All tutorials will be of three (3) hours duration.

Morning Tutorials	
T1	**Non-Speech Acoustic Event Detection And Classification** *Tuomas Virtanen and Jort F. Gemmeke* ***Sunday, 14 September 2014, 09:30 - 12:30; Peridot 201, Level 2, MAX Atria***
T2	**Contribution of MRI to Exploring and Modeling Speech Production** *Kiyoshi Honda and Jianwu Dang* ***Sunday, 14 September 2014, 09:30 - 12:30; Peridot 202, Level 2, MAX Atria***
T3	**Computational Models for Audiovisual Emotion Perception** *Emily Mower Provost and Carlos Busso* ***Sunday, 14 September 2014, 09:30 - 12:30; Peridot 205, Level 2, MAX Atria***
T4	**The Art and Science of Speech Feature Engineering** *Samuel Thomas and Sriram Ganapathy* ***Sunday, 14 September 2014, 09:30 - 12:30; Peridot 206, Level 2, MAX Atria***

Afternoon Tutorials	
T5	**Recent Advances in Speaker Diarization** *Hagai Aronowitz* ***Sunday, 14 September 2014, 14:00 - 17:00; Peridot 201, Level 2, MAX Atria***
T6	**Multimodal Speech Recognition with the AusTalk 3D Audio-Visual Corpus** *Roberto Togneri, Mohammed Bennamoun and Chao Sui* ***Sunday, 14 September 2014, 14:00 - 17:00; Peridot 202, Level 2, MAX Atria***
T7	**Semantic Web and Linked Big Data Resources for Spoken Language Processing** *Dilek Hakkani-Tür and Larry Heck* ***Sunday, 14 September 2014, 14:00 - 17:00; Peridot 205, Level 2, MAX Atria***
T8	**Speech and Audio for Multimedia Semantics** *Florian Metze and Koichi Shinoda* ***Sunday, 14 September 2014, 14:00 - 17:00; Peridot 206, Level 2, MAX Atria***

T1: Non-Speech Acoustic Event Detection and Classification

Tuomas Virtanen (Tampere University of Technology, Finland) and Jort F. Gemmeke (KU Leuven, Belgium)

Sunday, 14 September 2014, 09:30 - 12:30; Peridot 201, Level 2, MAX Atria

Abstract: The research in audio signal processing has been dominated by speech research, but most of the sounds in our real-life environments are actually non-speech events such as cars passing by, wind, warning beeps, and animal sounds. These acoustic events contain much information about the environment and physical events that take place in it, enabling novel application areas such as safety, health monitoring and investigation of biodiversity. But while recent years have seen wide-spread adoption of applications such as speech recognition and song recognition, generic computer audition is still in its infancy. Non-speech acoustic events have several fundamental differences to speech, but many of the core algorithms used by speech researchers can be leveraged for generic audio analysis. The tutorial is a comprehensive review of the field of acoustic event detection as it currently stands. The goal of the tutorial is foster interest in the community, highlight the challenges and opportunities and provide a starting point for new researchers. We will discuss what acoustic event detection entails, the commonalities differences with speech processing, such as the large variation in sounds and the possible overlap with other sounds. We will then discuss basic experimental and algorithm design, including descriptions of available databases and machine learning methods. We will then discuss more advanced topics such as methods to deal with temporally overlapping sounds and modeling the relations between sounds. We will finish with a discussion of avenues for future research.

T2: Contribution of MRI to Exploring and Modeling Speech Production

Kiyoshi Honda (Tianjin University, China) and Jianwu Dang (JAIST, Japan)

Sunday, 14 September 2014, 09:30 - 12:30; Peridot 202, Level 2, MAX Atria

Abstract: Magnetic Resonance Imaging (MRI) provides us a magic vision to look into the human body in various ways not only with static imaging but also with motion imaging. MRI has been a powerful technique for speech research to study finer anatomy of the speech organs or to visualize true vocal tracts in three dimensions. Inherent problems of slow image acquisition for speech tasks or insufficient signal-to-noise ratio for microscopic observation have been the cost for researchers to search for task-specific imaging techniques. The recent advances of the 3-Tesla technology suggest more practical solutions to broader applications of MRI by overcoming previous technical limitations. In this joint tutorial in two parts, we summarize our previous effort to accumulate scientific knowledge with MRI and to advance speech modeling studies for future development. Part I, given by Kiyoshi Honda, introduces how to visualize the speech organs and vocal tracts by presenting techniques and data for finer static imaging, synchronized motion imaging, surface marker tracking, real-time imaging, and vocal-tract mechanical modeling. Part 2, presented by Jianwu Dang, focuses on applications of MRI for phonetics of Mandarin vowels, acoustics of the vocal tracts with side branches, analysis and simulation in search of talker characteristics, physiological modeling of the articulatory system, and motor control paradigm for speech articulation.

T3: Computational Models for Audiovisual Emotion Perception

Emily Mower Provost (University of Michigan, USA) and Carlos Busso (University of Texas, Dallas, USA)

Sunday, 14 September 2014, 09:30 - 12:30; Peridot 205, Level 2, MAX Atria

Abstract: In this tutorial we will explore engineering approaches to understanding human emotion perception, focusing both on modeling and application. We will highlight both current and historical trends in emotion perception modeling, focusing on both psychological and engineering-driven theories of perception (statistical analyses, data-driven computational modeling, and implicit sensing). The importance of this topic can be appreciated from both an engineering viewpoint, any system that either models human behavior or interacts with human partners must understand emotion perception as it fundamentally underlies and modulates our communication, or from a psychological perspective, emotion perception is also used in the diagnosis of many mental health conditions and is tracked in therapeutic interventions. Research in emotion perception seeks to identify models that describe the felt sense of 'typical' emotion expression – i.e., an observer/evaluator's attribution of the emotional state of the speaker. This felt sense is a function of the methods through which individuals integrate the presented multimodal emotional information. We will cover psychological theories of emotion, engineering models of emotion, and experimental approaches to measure emotion. We will demonstrate how these modeling strategies can be used as a component of emotion classification frameworks and how they can be used to inform the design of emotional behaviors.

T4: The Art and Science of Speech Feature Engineering

Sriram Ganapathy and Samuel Thomas (IBM T.J. Watson Research Center, USA)

Sunday, 14 September 2014, 09:30 - 12:30; Peridot 206, Level 2, MAX Atria

Abstract: With significant advances in mobile technology and audio sensing devices, there is a fundamental need to describe vast amounts of audio data in terms of well representative lower dimensional descriptors for efficient automatic processing. The extraction of these signal representations, also called features, constitutes the first step in processing a speech signal. The art and science of feature engineering relates to addressing the two inherent challenges - extracting sufficient information from the speech signal for the task at hand and suppressing the unwanted redundancies for computational efficiency and robustness. The area of speech feature extraction combines a wide variety of disciplines like signal processing, machine learning, psychophysics, information theory, linguistics and physiology. It has a rich history spanning more than five decades and has seen tremendous advances in the last few years. This has propelled the transition of the speech technology from controlled environments to millions of end user applications. In this tutorial, we review the evolution of speech feature processing methods, summarize the recent advances of the last two decades and provide insights into the future of feature engineering. This will include the discussions on the spectral representation methods developed in the past, human auditory motivated techniques for robust speech processing, data driven unsupervised features like iVectors and recent advances in deep neural network based techniques. With experimental results, we will also illustrate the impact of these features for various state-of-the-art speech processing systems. The future of speech signal processing will need to address various robustness issues in complex acoustic environments while being able to derive useful information from big data.

T5: Recent Advances in Speaker Diarization

Hagai Aronowitz (IBM Research, Haifa, Israel)

Sunday, 14 September 2014, 14:00 - 17:00; Peridot 201, Level 2, MAX Atria

Abstract: The tutorial will start with an introduction to speaker diarization giving a general overview of the subject. Afterwards, we will cover the basic background including feature extraction, and common modeling techniques such as GMMs and HMMs. Then, we will discuss the first processing step usually done in speaker diarization which is voice activity detection. We will consequently describe the classic approaches for speaker diarization which are widely used today. We will then introduce state-of-the-art techniques in speaker recognition required to understand modern speaker diarization techniques. Following, we will describe approaches for speaker diarization using advanced representation methods (supervectors, speaker factors, iVectors) and we will describe supervised and unsupervised learning techniques used for speaker diarization. We will also discuss issues such as coping with unknown number of speakers, detecting and dealing with overlapping speech, diarization confidence estimation, and online speaker diarization. Finally we will discuss two recent works: exploiting a-prioiri acoustic information (such as processing a meeting when some of the participants are known in advanced to the system, and training data is available for them), The second recent work is modeling speaker-turn dynamics. If time permits, we will also discuss concepts such as multi-modal diarization and using TDOA (time difference of arrival) for diarization of meetings.

T6: Multimodal Speech Recognition with the AusTalk 3D Audio-Visual Corpus

Roberto Togneri, Mohammed Bennamoun, and Chao (Luke) Sui (University of Western Australia, Australia)

Sunday, 14 September 2014, 14:00 - 17:00; Peridot 202, Level 2, MAX Atria

Abstract: This tutorial will provide attendees a brief overview of 3D based AVSR research. In this tutorial, attendees will learn how to use the newly developed 3D based audio visual data corpus we derived from the AusTalk corpus (https://austalk.edu.au/) for audio-visual speech/speaker recognition. In addition, we also plan to introduce some results using this newly developed 3D audio-visual data corpus, which show that there is a significant speech accuracy increase by integrating both depth-level and grey-level visual features. In the first part of the tutorial, we will review some recent works published in the last decade, so that attendees can obtain an overview of the fundamental concepts and challenges in this field. In the second part of the tutorial, we will briefly describe the recording protocol and contents of the 3D data corpus, and show attendees how to use this corpus for their own research. In the third part of this tutorial, we will present our results using the 3D data corpus. The experimental results show that, compared with the conventional AVSR based on the audio and grey-level visual features, the integration of grey and depth visual information can boost the AVSR accuracy significantly. Moreover, we will also experimentally explain why adding depth information can benefit the standard AVSR systems. Eventually, through our tutorial, we hope we can inspire more researchers in the community to contribute to this exciting research.

T7: Semantic Web and Linked Big Data Resources for Spoken Language Processing

Dilek Hakkani-Tür and Larry Heck (Microsoft Research, USA)

Sunday, 14 September 2014, 14:00 - 17:00; Peridot 205, Level 2, MAX Atria

Abstract: State-of-the-art statistical spoken language processing typically requires significant manual effort to construct domain-specific schemas (ontologies) as well as manual effort to annotate training data against these schemas. At the same time, a recent surge of activity and progress on semantic web-related concepts from the large search-engine companies represents a potential alternative to the manually intensive design of spoken language processing systems. Standards such as schema.org have been established for schemas (ontologies) that webmasters can use to semantically and uniformly markup their web pages. Search engines like Bing, Google, and Yandex have adopted these standards and are leveraging them to create semantic search engines at the scale of the web. As a result, the open linked data resources and semantic graphs covering various domains (such as Freebase [3]) have grown massively every year and contain far more information than any single resource anywhere on the Web. Furthermore, these resources contain links to text data (such as Wikipedia pages) related to the knowledge in the graph. Recently, several studies on speech language processing started exploiting these massive linked data resources for language modeling and spoken language understanding. This tutorial will include a brief introduction to the semantic web and the linked data structure, available resources, and querying languages. An overview of related work on information extraction and language processing will be presented, where the main focus will be on methods for learning spoken language understanding models from these resources.

T8: Speech and Audio for Multimedia Semantics

Florian Metze (Carnagie Mellon University, USA) and Koichi Shinoda (Tokyo Institute of Technology, Japan)

Sunday, 14 September 2014, 14:00 - 17:00; Peridot 206, Level 2, MAX Atria

Abstract: Internet media sharing sites and the one-click upload capability of smartphones are producing a deluge of multimedia content. While visual features are often dominant in such material, acoustic and speech information in particular often complements it. By facilitating access to large amounts of data, the text-based Internet gave a huge boost to the field of natural language processing. The vast amount of consumer-produced video becoming available now will do the same for video processing, eventually enabling semantic understanding of multimedia material, with implications for human computer interaction, robotics, etc. Large-scale multi-modal analysis of audio-visual material is now central to a number of multi-site research projects around the world. While each of these has slightly different targets, they are facing largely the same challenges: how to robustly and efficiently process large amounts of data, how to represent and then fuse information across modalities, how to train classifiers and segmenters on unlabeled data, how to include human feedback, etc. In this tutorial, we will present the state of the art in large-scale video, speech, and non-speech audio processing, and show how these approaches are being applied to tasks such as content-based video retrieval (CBVR) and multimedia event detection (MED). We will introduce the most important tools and techniques, and show how the combination of information across modalities can be used to induce semantics on multimedia material through ranking of information and fusion. Finally, we will discuss opportunities for research that the INTERSPEECH community specifically will find interesting and fertile.