**ISCA DL Travel Report by Sadaoki Furui**

From December 9, 2013, to December 18, I travelled to India as an ISCA Distinguished Lecturer. I visited Kottayyam, Guwahati and Hyderabad, and gave five lectures/talks, including a keynote talk at WISP 2013 (Workshop on Image and Speech Processing), held at Indian Institute of Technology (IIT) Guwahati. The cost for international flights was covered by ISCA, and domestic flights, accommodation and food were covered by local hosts. My visit to India was very fruitful.

My lectures were well attended, and led to useful and inspiring discussions. I could not only encourage researchers and students to pursue research on automatic speech recognition, but also get new ideas for future research from various useful and positive feedback on my lectures. I was invited to participate in the opening ceremony of Advanced Digital Signal Processing Research Lab at Rajiv Gandhi Institute of Technology (RIT) Kottayyam. I could also visit the RIT Kottayyam and IIIT Hyderabad to have various technical discussions with faculty, researchers and students. I am planning to keep in touch with them, as it would be nice to set up some form of collaboration in the near future.

Details of the lectures/talks are as follows:

1. Keynote talk at WISP 2013

Talk title: “Automatic speech recognition: trials, tribulations and triumphs”

Abstract:

Although many important scientific advances have taken place in automatic speech recognition (ASR) research, we have also encountered a number of practical limitations which hinder a widespread deployment of applications and services. In most speech recognition tasks, human subjects produce one to two orders of magnitude fewer errors than machines. One of the most significant differences exists in that human subjects are far more flexible and adaptive than machines against various variations of speech, including individuality, speaking style, additive noise, and channel distortions. How to train and adapt statistical models for speech recognition using a limited amount of data is one of the most important research issues.

What we know about human speech processing and the natural variation of speech is very limited. It is important to spend more effort to clarify especially the mechanism underlying speaker-to-speaker variability, and devise a method for simultaneously modeling multiple sources of variations based on statistical analysis using large-scale databases. Future systems need to have an efficient way of representing, storing, and retrieving various knowledge resources.

Data-intensive science is rapidly emerging in scientific and computing research communities. The size of speech databases/corpora used in ASR research and development is typically 100 to 1,000 hours of utterances, which is too small considering the variety of sources of variations. We need to focus on solving various problems before efficiently constructing and utilizing huge speech databases, which will be essential to next-generation ASR systems.

Attendance: around 250 people from academia and industry.

Host: Professor S. R. Mahadeva Prasanna, Department of Electronics and Electrical Eng., IIT Guwahati, India, prasanna@iitg.ernet.in

1. Invited talk at WISP 2013

Talk title: “Selected topics from 40 years of my research on speech processing

Abstract:

This talk introduces major topics from my research on speech processing which I have conducted for more than 40 years, always bearing innovation in mind, at NTT (Nippon Telegraph and Telephone Corporation) Research Labs and Tokyo Institute of Technology (Tokyo Tech). I have conducted research mainly on automatic speech recognition (ASR), including automatic speaker recognition. ASR researchers have made significant progress since 1970 when I started my research, and now the techniques are successfully used in various application systems. I have been fortunate to have witnessed and participated in a period of significant technological progress.

In the 1970’s, I conducted research on speaker recognition using statistical features and showed that speech variation over time is one of the most serious problems in speaker recognition. At the end of 1970’s, I conducted research at Bell Labs and proposed a speaker recognition method using dynamic features of the speech spectrum. In the 1980’s, I verified that dynamic features of the speech spectrum play an important role in phoneme perception by human beings, and applied the method using the combination of instantaneous and dynamic cepstrum, which I originally proposed for speaker recognition, to ASR. This method became a standard and is still widely used. For speaker recognition, we conducted pioneering research using GMMs. In the 1990’s, we conducted various research on model adaptation for ASR, and proposed text-prompted speaker recognition. In the 2000’s, I supervised a project on spontaneous speech recognition based on a large speech corpus, which significantly contributed to the progress of spontaneous speech recognition. In this project, we conducted pioneering research on automatic speech summarization. I also supervised a COE (Center of Excellence) project for large-scale knowledge resources at our university, and a national project for developing practical ASR techniques in cooperation with major Japanese companies. In the latter project, we successfully constructed a very high-performance WFST-based ASR decoder.

I believe a major direction of future ASR research is modeling of various speech variations based on advanced machine-learning techniques and efficient construction and utilization of large-scale speech corpora.

This talk also includes my recent activities at Tokyo Tech and Toyota Technological Institute at Chicago.

Attendance: around 250 people from academia and industry.

1. 1st lecture at RIT Kottayyam

Lecture title: “Introduction to automatic speech and speaker recognition

Abstract:

This talk introduces principles of automatic speech and speaker recognition technology. The automatic speech recognition part describes major applications, history of speech recognition research, and statistical modeling (feature extraction, HMM. language modeling, and decoding). This part also describes sources of acoustic variation in speech, such as individuality, speaking style, additive noise, and channel distortions, which are the major reasons of speech recognition errors. The automatic speaker recognition part describes principles of speaker verification and identification methods by text-dependent/-independent frameworks, major applications, and basic structure of speaker recognition systems. This talk also includes unsolved problems and topics for future research.

Attendance: around 150 people consisting of faculty, researchers and students.

Host: Professor Leena Mary, Dept of Electronics & Communication, RIT Kottayyam, [leena.mary@rit.ac.in](mailto:leena.mary@rit.ac.in)

1. 2nd lecture at RIT Kottayyam

Lecture title: “Automatic speech recognition: trials, tribulations and triumphs”

Abstract: (same as above)

Attendance: around 150 people consisting of faculty, researchers and students.

1. Lecture at IIIT Hyderabad

Lecture title: “Data-intensive ASR based on machine learning”

Abstract:

Since speech is highly variable, even if we have a fairly large-scale database, we cannot avoid the data sparseness problem in constructing automatic speech recognition (ASR) systems. How to train and adapt statistical models using limited amounts of data is one of the most important research issues in ASR. This talk summarizes major techniques that have been proposed to solve the generalization problem in acoustic model training and adaptation, that is, how to achieve high recognition accuracy for new utterances. One of the common approaches is controlling the degree of freedom in model training and adaptation. The techniques can be classified by whether a priori knowledge of speech obtained from a speech database such as those recorded using many speakers is used or not. Another approach is maximizing “margins” between training samples and the decision boundaries. Many of these techniques have also been combined and extended to further improve performance.

Although many useful techniques have been developed, we still do not have a golden standard that can be applied to any kind of speech variation and any condition of the speech data available for training and adaptation. We need to focus on collecting rich and effective speech databases covering a wide range of variations, active learning for automatically selecting data for annotation, cheap, fast and good-enough transcription, and unsupervised, semi-supervised or lightly-supervised training/adaptation, based on advanced machine learning techniques. We also need to extend current efforts to understand more about human speech processing and the mechanism of natural speech variation.

Attendance: around 50 people consisting of faculty and graduate students

Host: Professor B. Yegnanarayana, IIIT Gachibowli, Hyderabad, India, yegna@iiit.ac.in