# Interspeech 2013 TUTORIAL

1) Title of the tutorial: Recent Advances in Large Vocabulary Continuous Speech Recognition

2) Biographies/web pages of presenter(s):

George Saon received his M.Sc. and PhD degrees in Computer Science from Henri Poincare University in Nancy, France in 1994 and 1997. In 1995, Dr. Saon obtained his engineer diploma from the Polytechnic University of Bucharest, Romania. From 1994 to 1998, he worked on two-dimensional stochastic models for off-line handwriting recognition at the Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA). Since 1998, Dr. Saon is with the IBM T.J. Watson Research Center where he worked on a variety of problems spanning several areas of large vocabulary continuous speech recognition such as discriminative feature processing, acoustic modeling, speaker adaptation and large vocabulary decoding algorithms. Some of the techniques that he co-invented are well known to the speech community like heteroscedastic discriminant analysis (HDA), implicit lattice discriminative training, lattice-MLLR, feature-space Gaussianization, fast FSM-based Viterbi decoding, etc. Since 2001, Dr. Saon has been a key member of IBM's speech recognition team which participated in several U.S, government-sponsored evaluations for the EARS, SPINE and GALE programs. He has published over 80 conference and journal papers and holds several patents in the field of ASR. He currently serves as an elected member of the IEEE Speech and Language Processing Technical Committee.

http://researcher.watson.ibm.com/researcher/view.php?person=us-gsaon

Jen-Tzung Chien received his Ph.D. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1997. During 1997-2012, he was with the National Cheng Kung University, Tainan, Taiwan. Since 2012, he has been with the Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, where he is currently a Distinguished Professor. He held the Visiting Researcher positions at the Panasonic Technologies Inc., Santa Barbara, CA, the Tokyo Institute of Technology, Tokyo, Japan, the Georgia Institute of Technology, Atlanta, GA, the Microsoft Research Asia, Beijing, China, and the IBM T. J. Watson Research Center, Yorktown Heights, NY. His research interests include machine learning, acoustic modeling, language modeling, speaker adaptation and discriminative training. Dr. Chien is a senior member of the IEEE Signal Processing Society. He served as the associate editor of the IEEE Signal Processing Letters, in 2008-2011, the guest editor of the IEEE Transactions on Audio, Speech and Language Processing in 2012, the organization committee member of the ICASSP 2009, the tutorial speaker of the ICASSP 2012, and the area coordinator of the Interspeech 2012. He is appointed as the APSIPA Distinguished Lecturer for 2012-2013. He was a co-recipient (with George Saon) of the Best Paper Award of the

IEEE Automatic Speech Recognition and Understanding Workshop in 2011. He received the Distinguished Research Award from the National Science Council, Taiwan, in 2006 and 2010.

http://chien.cm.nctu.edu.tw

3) Overview of tutorial: In this tutorial, we will present some state-of-the-art techniques for large vocabulary continuous speech recognition (LVCSR). This tutorial will cover different components in LVCSR including front-end processing, acoustic modeling, language modeling and hypothesis search and system combination. In the front-end processing module, we will introduce popular feature extraction methods and transformations and discuss adaptive and discriminative features. In the acoustic modeling section, we will present feature-space and model-space discriminative training and speaker adaptation. A series of discriminative estimation methods will also be compared. Additionally, we will report some recent progress on the use of deep neural network acoustic models. In language modeling (LM), we will address Bayesian learning to deal with the issues of insufficient training data, large-span modeling and model regularization. High-performance LMs will be presented. In hypothesis search, we will present state-of-the-art search algorithms and system combination methods. In addition, we will point out some new trends for LVCSR including structural state model, basis representation, model regularization and deep belief networks.

4) Rationale (importance/timeliness/target audience): LVCSR systems have broad applications ranging from early speaker-dependent dictation systems to speaker-independent broadcast news transcription and indexing, lectures and meeting transcription, conversational telephone speech transcription, open-domain voice search, medical and legal speech recognition, call center applications and many more. Considering the increasingly expanding applications, this tutorial is helpful to a large number of speech recognition researchers and practitioners attending Interspeech 2013.

5) Presentation outline:
   1. Introduction [George Saon]
   2. Front-end processing [George Saon]
      I. Feature extraction and transformation
      II. Speaker-adaptive features
      III. Discriminative features
   3. Acoustic modeling [George Saon and Jen-Tzung Chien]
      I. Hidden Markov models
      II. Discriminative training (feature & model space)
      III. Speaker adaptation (feature & model space)
      IV. Deep neural networks
   4. Language modeling [Jen-Tzung Chien]

    I. Smoothing algorithms
    II. Large-span modeling
    III. Exponential model
    IV. Model regularization
5. Hypothesis search and system combination [George Saon]
    I. WFST decoding and LM rescoring
    II. System combination
    III. Bagging and boosting
6. Future directions [Jen-Tzung Chien and George Saon]
    I. Structural state model
    II. Basis representation
    III. Robust modeling and deep learning

6) Description of outline:

The presentation of this tutorial is arranged into six parts. First of all, George will share the current status of research on LVCSR at IBM and other leading teams around the world and explain key components which considerably affect system performance. Secondly, he will address feature extraction and transformation and introduce front-end processing methods for speaker-adaptive features and discriminative features. In the third part, George and Jen-Tzung will present discriminative training and speaker adaptation methods in feature space and in model space. Different discriminative training and adaptation criteria will be illustrated and compared. Some LVCSR systems based on deep neural networks will also be described.

After the coffee break, Jen-Tzung will present the fourth part which focuses on some issues caused by traditional language models (LMs) based on statistical *n*-grams. He will introduce Bayesian approaches to dealing with the issues of insufficient training data, large-span modeling and model regularization. He will explain the motivations of considering compact modeling and uncertainty modeling by using Bayesian nonparametrics and variational Bayesian methods. Neural network LMs and structured LMs are also introduced. In the fifth part, George will introduce back-end processing components including hypothesis search and system combination. He will talk about LM rescoring and LVCSR decoding using weighted finite-state transducers (WFSTs). System combination based on bagging and boosting will also be described. In the final part, Jen-Tzung and George will point out some future directions for LVCSR. We will explain why machine learning methods can deal with the challenges of big data and model generalization. In particular, structural learning, sparse representation, robust modeling and deep learning will be emphasized.