NATIONAL UNIVERSITY OF SINGAPORE

School of Computing

PH.D DEFENCE - PUBLIC SEMINAR

Title: **Temporally Varying Weight Regression for Speech Recognition**

Speaker: Mr Liu Shilin

Date/Time: 21 July 2014, Monday, 10:00 AM to 11:30 AM

Venue: Executive Classroom, COM2-04-02

Supervisor : Dr Sim Khe Chai, Assistant Professor, School of Computing

Automatic Speech Recognition (ASR) has been one of the most popular research areas in computer science. Many state-of-the-art ASR systems still use the Hidden Markov Model (HMM) for acoustic modelling due to its efficient training and decoding. HMM state output probability of an observation is assumed to be independent of the other states and the surrounding observations. Since temporal correlation between observations exists due to the nature of speech, this assumption is poorly made for speech signal. Although the use of the dynamic parameters and the Gaussian mixture models (GMM) has greatly improved the system performance, implicitly or explicitly modelling the trajectory temporal correlation can potentially improve the ASR systems.

Firstly, an implicit trajectory model called Temporally Varying Weight Regression (TVWR) is proposed in this thesis. Motivated by the success of discriminative training of time-varying mean (fMPE) or variance (pMPE), TVWR aims of modelling the temporal correlation information using the temporally varying GMM weights. In this framework, the time-varying information is represented by the compact phone/state posterior features predicted from the long span acoustic features. The GMM weights are then temporally adjusted through a linear regression of the posterior features.
Both maximum likelihood and discriminative training criteria are formulated for parameter estimation. Next, the complexity control of the TVWR system is studied to achieve a good compromise between the decoding efficiency and recognition performance. To improve the decoding efficiency, parameter clustering of regression parameters is proposed. On the other hand, multi- stream TVWR is also proposed to boost the recognition performance by introducing temporal and spatial context expansions.

Secondly, TVWR is investigated as an approach to combine the GMM and the deep neural network (DNN). As reported by various research groups, DNN has been found to consistently outperform GMM and has become the new state-of-the-art for speech recognition. However, many advanced adaptation techniques have been developed for GMM based systems, while it is difficult to devise effective adaptation methods for DNNs. This thesis proposes a novel method of combining the DNN and the GMM using the TVWR

framework to take advantage of the superior performance of the DNNs and the robust adaptability of the GMMs. In particular, posterior grouping and sparse regression are proposed to address the issue of incorporating the high dimensional DNN posterior features.

Finally, adaptation and adaptive training of TVWR are investigated for robust speech recognition. In practice, many speech variabilities exist, which will lead to poor recognition performance for mismatched conditions.
TVWR has not been formulated to be robust against those speech variabilities, such as background noises, transmission channels, speakers, etc. The robustness of TVWR can be improved by applying the adaptation and adaptive training techniques, which have been developed for the GMMs.
Adaptation aims to change the model parameters to match the test condition using limited supervision data from either the reference or hypothesis.
Adaptive training estimates a canonical acoustic model by removing speech variabilities, such that adaptation can be more effective. Both techniques are investigated for the TVWR systems using either the GMM or the DNN-based posterior features. Benchmark tests on the Aurora 4 corpus for robust speech recognition showed that TVWR obtained 21.3% and 11.6% relative improvements over the DNN baseline system and the best system in currently reported literatures, respectively.