INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	SYED SHAHNAWAZUDDIN
Roll Number	:	10610209
Programme of Study	:	Ph.D.

Thesis Title: IMPROVING CHILDREN'S MISMATCHED ASR THROUGH ADAPTIVE PITCH COMPENSATION

Name of Thesis Supervisor(s)	:	PROF. DR. ROHIT SINHA
Thesis Submitted to the Department/ Center	:	EEE
Date of completion of Thesis Viva-Voce Exam	:	08-08-2016
Key words for description of Thesis Work	:	Automatic speech recognition, acoustic mismatch, children's ASR, sparse representation, fast adaptation, pitch-adaptive MFCCs, adaptive-liftering, SGMM, DNN.

SHORT ABSTRACT

With the progress made in the speech processing over the last few decades, an increasing number of user applications employing automatic speech recognition (ASR) systems are being developed. In such human-machine interaction (HMI) applications, the ASR system is often accessed by both the adults and the children. It is well known that the ASR systems trained on the adults' speech exhibit a severely degraded recognition performance when used for transcribing the speech data from the child speakers and vice-versa. One of the ways to achieve good ASR performance for both the adults and the children is to pool a large amount of data from both the group of speakers in the training of the system. The scarcity of the children's speech corpus makes this approach infeasible. On the other hand, pooling a limited amount of children's data with the adults' training set is not found to be very effective. Consequently, this thesis explores the possibility of achieving improved recognition performance for the children's speech on adults' speech trained ASR systems. To enhance the performance of the children's mismatched ASR, the thesis begins with an exploration of some of the existing adaptation/normalization techniques. Despite the observed improvements with the application of the existing approaches, a large gap still remains between the adults' matched and children's mismatched testing cases. This gap in the performance is attributed to the severe mismatch in the acoustic and the linguistic correlates for the adults' and the children's speech. Among the various sources of mismatch identified in literature, the differences in the size of the vocal organs and the pitch (fundamental frequency) are known to be the most dominant ones. The frequency-warping-based vocal tract length normalization (VTLN) approach is already noted to be very effective in mitigating the ill effects of the differences in the vocal tract dimensions. Therefore, we have tried to analyze the cause and the extent of the pitch-induced mismatch between adults and children in this work. Based on our analysis, we have devised techniques that target the pitch variation across the speakers. One of the propositions in this thesis is the reduction of the pitch-induced variations through a structured projection of the front-end features and the parameters of the acoustic model to a lower dimensional subspace. Additionally, a spectral smoothing approach is proposed which address the pitch-induced distortions prior to computation of the acoustic features. Both these approaches are found to be highly effective in the context of the children's mismatched ASR. Furthermore, the proposed techniques are noted to result in additive improvements when combined with some of the existing feature-space normalization as well as the model-space adaptation techniques. In order to reduce the latency in the implementation of the model-space adaptation approaches,

we have also developed a few fast adaptation techniques suitable for those ASR tasks involving HMI. Most of the presented techniques were initially developed for the acoustic modelling employing the Gaussian-mixture-based hidden Markov models (GMM-HMM). But, the observed improvements are also found to hold largely for the recently introduced acoustic modelling techniques based on the subspace GMM (SGMM) and the deep neural network (DNN). In the case of the SGMM- and the DNN-based systems, we have also studied the relative effectiveness of the VTLN and the feature-space maximum likelihood linear regression (fMLLR). The fMLLR-based feature normalization is already reported to be very effective for the DNN-based system. On the other hand, the VTLN is observed to be largely ineffective in those cases where the number hidden layers is very large. On the contrary, our study finds that the VTLN is effective not only for the shallow networks but also for the deeper ones in the context of the children's mismatched ASR.