INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI
SHORT ABSTRACT OF THESIS

Name of the Student	:	Bhanu Priya
Roll Number	:	11610228
Programme of Study	:	Ph. D.
Thesis Title:	:	Speech Subspace Modelling With Speaker Adaptation for Stress Normalization
Name of Thesis Supervisor(s)	:	Prof. Samarendra Dandapat
Thesis Submitted to the Department/ Center	:	EEE
Date of completion of Thesis Viva-Voce Exam	:	23 April, 2018
Key words for description of Thesis Work	:	Stressed Speech, Stress Normalization, Orthogonal Projection, Singular Value Decomposition, Non-linear Transformation, Polynomial Function, Posteriorgram Representation, Sparse Representation, Speech Recognition, Visual Analysis, Error Analysis

---

SHORT ABSTRACT

This thesis work is an investigation on the normalization of stress information for the effective processing of stressed speech. Speakers change the speech production system to communicate the information about the adverse environmental factors and to retain the intelligibility of speech signals. Any diversifications in the environmental condition from the normal or neutral state lead to an adverse condition and it is referred as the stress condition. The speech signal produced under stress condition by any modification in the speech production system is called as the stressed speech. The speech produced under normal or neutral condition is generally referred to as the neutral speech. Stress induces a large acoustic mismatch between the different speech units of neutral and stressed speech. These mismatched properties severely affect various real life applications. Thus, there is an essential need of stress normalization, that can reduce the acoustic mismatch between the neutral and the stressed speech and help the users with a better robust practical application. The present thesis aims at developing robust and computationally efficient algorithms to normalize the stress information.

First, novel linear and non-linear subspace modelling approaches are proposed to reduce the acoustic mismatch between the neutral and the stressed speech signals. The linear characteristic of stressed speech has been studied on the linear subspace. The linear subspace is modelled by exploiting an orthogonal projection and linear transformation techniques. The non-linearity between the speech and the stress information has been investigated on the non-linear data space by exploring the subspace projection through the non-linear transformation using the polynomial function. The results show that, the non-linear subspace modelling using the polynomial function of

specific order is very effective for normalizing the stress information compared to the linear subspace modelling techniques.

Secondly, an effective stress normalization method has been developed by investigating the changes in the vocal-tract system under stress condition in the Gaussian-subspace. The acoustic mismatch between the vocal-tract system parameters of neutral and stressed speech is reduced by the subspace projection onto a common Gaussian-subspace, which consists of vocal-tract system parameters of neutral speech utterances. In this study, the proposed subspace projection is accomplished using the posterior probability information, which extracts the posteriorgram features. The synthesis of neutral and stressed speech signals using their estimated posteriorgram features corresponding to their vocal-tract system parameters has been observed very effective in reducing the acoustic mismatch between them. In the third approach, we have further investigated the deviation in the vocal-tract system under stress condition by exploring a novel subspace modelling technique in the sparse domain. In dictionary learning framework, two types of dictionaries have been proposed: the invariable size global dictionary and the utterance-specific adaptive dictionary. The invariable size global dictionary is learned using the well known K-SVD algorithm. The utterance-specific adaptive dictionary incorporates the information about the duration parameter of speech utterance, which is modeled using the K-nearest-neighbour (K-NN) algorithm. Both the neutral and the stressed speech are synthesized using their corresponding estimated vocal-tract system parameters and they are considered as the speech signals with the characteristics similar to the neutral speech. The experimental observations illustrate that, the sparse representation over the utterance-specific adaptive dictionary effectively reduces the acoustic variations between the vocal-tract system parameters of neutral and stressed speech signals with the significant improvement in stressed speech recognition performances compared to the conventional case and the case of using invariable size global dictionary.

Most of the methods reported in the literature to study the stressed speech are mainly based on the investigation of the changes in the speech production system under stress condition. It is observed that, variations in the anatomical and the physiological characteristics associated with different speakers under stress condition create a large acoustic mismatch between the neutral and the stressed speech. Hence, there is a need to develop robust methods which can normalize the speaker variabilities in the presence of stress. In order to mitigate the effect of such variabilities, at first, the heteroscedastic linear discriminant analysis (HLDA)- and the linear discriminant analysis (LDA)-based low-rank subspace projections have been explored in the maximum likelihood linear transformation (MLLT)-based semi-tied adaptation technique. This helps in reducing the dimension as well as the correlation parameter of the feature- and the model-space. After that, the feature-space maximum-likelihood (fMLLR) transformations are generated for the training and the test speech utterances in the speaker adaptive training (SAT) mode to reduce the speaker variability. The effectiveness of proposed stress normalization methods are evaluated on three distinct frameworks namely: the stressed speech recognition, the visual analysis and the error analysis, respectively. The speech recognition for the stressed speech has been accomplished over the speaker dependent (SD) automatic speech recognition (ASR) systems employing acoustic models based on Gaussian mixture model (GMM), subspace Gaussian mixture model (SGMM) and deep neural network (DNN). In the visual analysis, we have studied the variations in the air pressure that human auditory system are able to perceive as sound and the changes in the spectral distributions with respect to the time by interpreting the waveform and the spectrogram, respectively. The error analysis measures the relative entropy between the Gaussian-subspaces by exploiting the Kullback Leibler (KL) divergence metric.