Subspace Projection Based Analysis of Speech Under Stressed Condition

Sumitra Shukla, S.Dandapat and S. R. Mahadeva Prasanna
Dept. of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati
Guwahati, India
E-mail:{sumitra, samaren and prasanna}@iitg.ernet.in

***Abstract*—This paper proposes a novel subspace projection based approach for analysis of stressed speech signal. The projection of stressed speech vectors onto the neutral speech subspace can separate speech specific information from stress information. Orthogonality between speech and stress is assumed to separate these two information. The orthogonal relation between speech and stress subspaces is verified using speech and stress recognition techniques with a stressed speech database consisting of four stress conditions namely, neutral, angry, sad and Lombard from 30 words vocabulary. Studies show that the speech and stress specific information are present in their respective subspaces which proves orthogonality between these two subspaces.**

***Keywords*-Stressed speech recognition, Stress compensation techniques, Subspace projection based approach.**

I. Introduction

Recent developments in automatic speech processing are mainly focused on the mismatch condition that may be introduced due to background noise [1], reverberation [2] and when speakers are in stress [3]. Generally, the system is trained in controlled environment where it is assumed that the speaker is completely stress-free and other noises are negligible. The performance of this system may degrade if stressed speech is tested for recognition [4]. Under stressed condition, the speech information can be assumed as corrupted with additional stress information. To eliminate the effect of stress, stress compensation techniques are used. The stress components in the cepstral coefficients are assumed as additive and the adaptive cepstral mean compensation technique is proposed at the word level [3] and at the broad phoneme level [10]. Afify et.al assumed stress component as Gaussian additive component [12]. In addition, the speech and the stress components are assumed statistically independent. Maximum likelihood state based additive bias compensation technique for continuous density hidden Markov model is proposed [12]. In spectral domain, the stress effect is equalized from the formant locations and formant bandwidths of spectrum by using ratio of these features from the neutral and stressed conditions [5]. However, these techniques require explicit information of stress classes. This is a two stage recognition problem. Any failure of stress classifier can reduce the performance of system. Multi-style training is considered as an alternative approach to improve the performance [5]. This method does not require prior information of stress. This approach requires various stressed speech during the training phase. In view of the variability in stressed speech, this may be practically difficult [5].

In this paper, orthogonality is assumed between speech and stress components. Subspace projection based approach is proposed to decompose the stressed speech signal into speech and stress components. Section II presents the principle of subspace projection based approach. The orthogonal relation between speech and stress information is verified in Section III. Section IV describes the summary and conclusion of the work.

II. Subspace Projection Analysis of Speech under Stressed Condition

This section presents the proposed subspace projection based approach to separate speech and stress information. This approach assumes a linear model of neutral speech subspace of dimension M where the neutral speech vector ($\mathbf{s}$) can be represented as

$$\mathbf{s} = \sum_{m=1}^{M} w_m \mathbf{v}_m \tag{1}$$

where, $w = \{w_1, w_2, ...w_M\}$ are weights and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2.....\mathbf{v}_M]$ are M basis vectors [6]. Each basis is a K vector. The basis vectors can be damped real sinusoids taken from speech spectrum and they are assumed as linearly independent. The set of neutral speech vectors $\{\mathbf{s}\}$ can be assumed in a subspace of $\mathbf{R}^K$ spanned by the column of $\mathbf{V}$. This subspace is referred to as *Neutral Speech Subspace*. The vectors of this subspace are assumed to contain speech (semantic) information. The covariance matrix of the neutral speech vector $\mathbf{s}$ is given as

$$\mathbf{C}_s = E\{\mathbf{s}\mathbf{s}^T\} = \mathbf{V}\mathbf{C}_w\mathbf{V}^T \tag{2}$$

where, $\mathbf{C}_w$ denotes the covariance matrix of vector w. The number of basis vectors is smaller than dimension of vector, therefore, the covariance matrix of speech vector, $\mathbf{C}_s$ contains $K-M$ zero eigenvectors. Under stressed condition, the speech vector contains additional stress information. This vector can be termed as stressed speech vector. The stressed speech vector ($\mathbf{y}$) can be represented as function (f) of speech ($\mathbf{s}$) and stress ($\mathbf{x}$) information. Mathematically, this can be written as $\mathbf{y} = f(\mathbf{s}, \mathbf{x})$. The set of stressed speech vectors $\{\mathbf{y}\}$ spans the stressed speech subspace. The

978-1-4673-4805-8/12/$31.00

covariance matrix of **y** can be considered as summation of covariance matrix of speech and covariance matrix of stress, if stress is assumed additive and orthogonal with the speech component in stressed speech vector [6]. It can be represented as

$$\mathbf{C}_y = E\{\mathbf{y}\mathbf{y}^T\} = \mathbf{C}_s + \mathbf{C}_x \tag{3}$$

where, $\mathbf{C}_x$ is the covariance matrix of stress component **x**. The covariance matrix of stressed speech vector may also contains $K - M$ eigenvectors. These eigenvectors may contains stress information. Eigenvalue decomposition (EVD) can decompose the stress and speech components from $\mathbf{C}_y$ as given in

$$\mathbf{C}_y = \mathbf{U}\mathbf{\Lambda}_y\mathbf{U}^T \tag{4}$$

where $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_k]$ denotes the orthogonal matrix of eigenvectors of $\mathbf{C}_y$ and $\mathbf{\Lambda}_y = diag(\Lambda_y(1), ... \Lambda_y(k))$ denotes the diagonal matrix of eigenvalues of $\mathbf{C}_y$. The subspace based methods are based on a partitioning of the eigenvectors into a set belonging to the speech subspace spanned by the columns of **U** and an orthogonal complement known as the stress subspace. This type of approach is normally used for noise filtration [6]. EVD can decompose the speech and stress subspaces more reliably when the covariance matrix ($\mathbf{C}_x$) of stress **x** is known. However, the stress information is not known, therefore, the reliable decomposition using this approach is not possible. In this study, speech and stress components are assumed as orthogonal. To separate the stress and speech components from the stressed speech vector, subspace projection based approach is proposed [7]. The geometrical properties of matrix-valued statistic is exploited to estimate speech and stress components from stressed speech vector. All vectors of neutral speech subspace of $\mathbf{R}^K$ are represented by a set of representative vectors termed as codevectors. The codevectors $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, ...\mathbf{a}_N\}$ are the mean vectors. The codevectors of neutral speech subspace are linear combinations of speech vectors which is spanned by the column of **V**. Therefore, these vectors lie in the subspace of $\mathbf{R}^K$.

The subspace projection based analysis is shown in Figure 1. Figure shows that the stressed speech subspace deviates from the neutral speech subspace. In this study, the deviation is assumed due to stress present in the signal. Let, $\mathbf{P}_n$ be the projection matrix required to project a vector onto $\{\mathbf{a}_n\}$. $\mathbf{P}_n$ is given as

$$\mathbf{P}_n = \frac{\mathbf{a}_n\mathbf{a}_n^T}{\mathbf{a}_n^T\mathbf{a}_n} \qquad 1 \leq n \leq N \tag{5}$$

The multiplication of $\{\mathbf{P}_n\}$ with **y** produces projection of **y** onto $\{\mathbf{a}_n\}$ which can be represented as $\mathbf{p} = \{\mathbf{P}_n\mathbf{y}\}$ where, **p** is the component of **y** lies in the neutral speech subspace as shown in Figure 1. The set of projection of stressed speech vector, **y**, onto the set of codevectors $\{\mathbf{a}_n\}$, $1 \leq n \leq N$, can be seen as projection of **y** onto neutral speech subspace. These components contain speech specific information of stressed speech vector. The error between **y** and $\{\mathbf{P}_n\mathbf{y}\}$ is orthogonal to the neutral speech subspace as shown in Figure 1. In other word, the error between these two vectors represents the projection of **y** onto the orthogonal subspace of **A**. The orthogonal vector can be represented as $\mathbf{e} = \{(\mathbf{y} - \mathbf{P}_n\mathbf{y})\}$. Due to orthogonal assumption between speech and stress information, the orthogonal vectors may contain stress specific information of stressed speech vector.

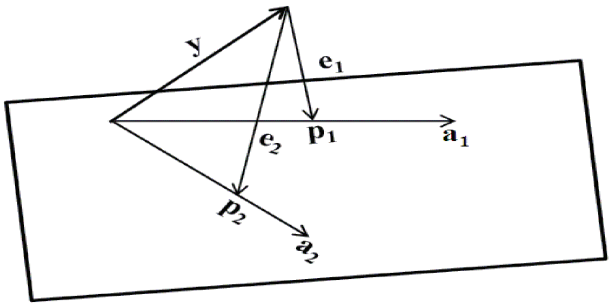

Figure 1. Subspace projection based analysis.

The projection of **y** onto the set of $\{\mathbf{a}_n\}$, $1 \leq n \leq N$ produces N projected vectors in neutral speech subspace **A** and corresponding N orthogonal vectors are in the orthogonal subspace. The stressed speech subspace can be decomposed into the speech (**S**) and the stress (**X**) subspaces. In this work, the stress vector **x** is decided the orthogonal vector which has minimum length as given in Eq. 6 and Eq. 7. The projected vector corresponding to that orthogonal vector is considered as speech vector as given in Eq.8.

$$\hat{n} = \arg\min_n [(\mathbf{y} - \mathbf{P}_n\mathbf{y})^T(\mathbf{y} - \mathbf{P}_n\mathbf{y})] \tag{6}$$

$$\mathbf{x} = (\mathbf{y} - \mathbf{P}_{\hat{n}}\mathbf{y}) \tag{7}$$

$$\mathbf{s} = \mathbf{P}_{\hat{n}}\mathbf{y} \tag{8}$$

III. Analysis of speech and stress subspaces under different stress conditions

The decomposition of stressed speech vector onto speech and stress components is performed under the assumption that the stress information is orthogonal to the speech information. According to this assumption, the speech information should be present in speech subspace and stress subspace should contains stress information. The speech information in stress and speech subspace should be negligible and similarly little stress information may be present in speech subspace. This section verifies this assumption experimentally by considering speech and stress recognition

techniques. The underlying principle of this study is to investigate how effectively the stress and the speech information are separated.

1) Description of speech database and experimental setup: A simulated stressed speech database is collected in Hindi, an Indian language, from fifteen non-professional speakers [8]. The dataset of 30 words vocabulary is considered under neutral, angry, sad and Lombard conditions [8]. The speech is recorded at 16 kHz sampling rate with 16 bits/sample resolution, in two separate sessions. The stressed speech vectors from different stress conditions are computed using $MFCC$ feature. The speech signal is segmented into number of frames with length of 160 samples and overlapping of 80 samples. $MFCC$ features of frames of a utterance is considered as vectors of that utterance. Vector quantization (VQ) and continuous density hidden Markov model (HMM) based classifiers are used to develop word and stress models. By varying VQ and HMM parameters, it is found that 32 size of codebook is suitable to cluster words in VQ. In HMM, the word model is developed using ten states and left to right transition with two mixture components per state. The neutral speech subspace of a word is developed by VQ techniques. The subspace is formed by 32 codevectors.

2) Analysis of speech information in speech and stress components: This subsection investigates speech information present in speech and stress subspaces. To evaluate speech information in these two subspaces, speaker dependent speech recognizer is considered. The projections of stressed speech vectors $\{\mathbf{y}\}$ of a utterance onto the neutral speech subspace ($\mathbf{A}$) corresponding to a word form speech subspace using Eq. 8. In speech recognition task, 30 words dataset produced by a speaker can be assumed to form neutral speech subspaces $\{\mathbf{A}_i\}$, $1 \leq i \leq 30$ which are subspaces of $\mathbf{R}^K$. Thus, the projection of stressed speech vectors $\{\mathbf{y}\}$ of that speaker onto the neutral speech subspaces $\{\mathbf{A}_i\}$ corresponding to 30 words form 30 speech subspaces $\{\mathbf{S}_i\}$ in $\mathbf{R}^K$. The minimum average orthogonal distance from $\{\mathbf{y}\}$ to different neutral speech subspaces $\{\mathbf{A}_i\}$ decide the speech component corresponding to that stressed speech vector.

The speech specific information in speech and stress subspaces from different stress conditions are investigated using speaker dependent HMM based speech recognizer. The speech and stress components under different stress conditions are investigated for two male and two female speakers. The vectors of neutral speech subspace of a speaker of 1^{st} session data set are used to develop model for different words. During recognition, stressed speech vectors $\{\mathbf{y}\}$ and corresponding speech vectors $\{\mathbf{s}\}$ and stress vectors $\{\mathbf{x}\}$ of 2^{nd} session of database are computed. The recognition performance of these vectors of four speakers are shown in Figure 2. Figure shows that the recognition performance of speech vectors $\{\mathbf{s}\}$ under stressed condition

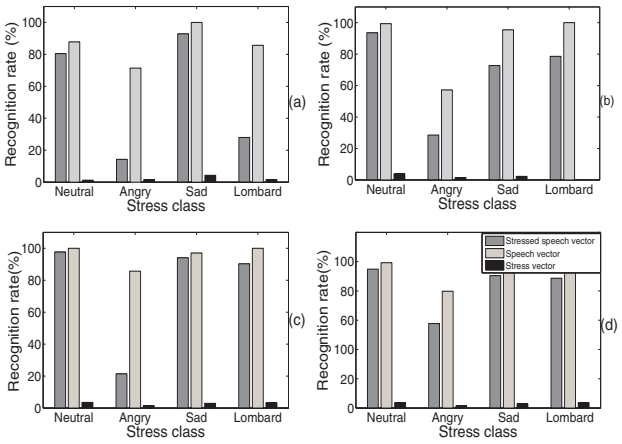

Figure 2. Recognition performance of speech recognition, (a)and (b) of two female speakers, and (c)and (d) of two male speakers.

increases compared to their stressed speech vectors $\{\mathbf{y}\}$. The speech vector lies in the neutral speech subspace, hence, it contains speech specific information. Alternatively, the recognition performance of stress vectors $\{\mathbf{x}\}$ under stressed condition decreases significantly from that of $\{\mathbf{y}\}$. These observations are same for all speakers. This result infers that the speech information in stress subspace is negligible.

A. Analysis of stress information in speech and stress components

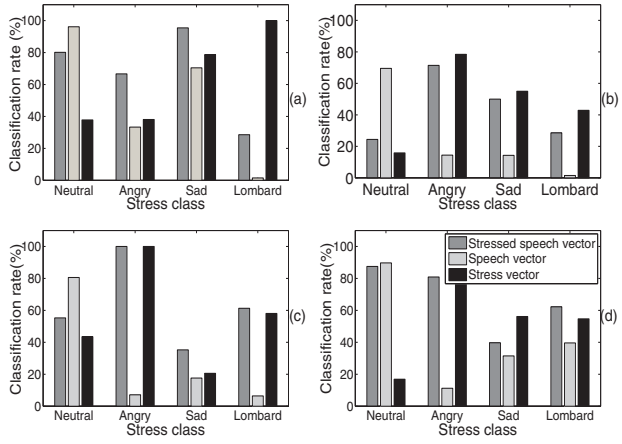

Figure 3. Classification rate of stress classifier, (a)and (b) of two female speakers, and (c)and (d) of two male speakers.

This subsection quantifies the level of stress present in speech and stress subspaces by using speaker dependent stress classifier. The stress classifiers are developed using HMM techniques. The stressed speech vectors $\{\mathbf{y}\}$ of all words produced by a speaker of 1^{st} session data set are considered to model stress. In HMM, the stress model is

developed using sixteen states and left to right transition with two mixture components per state. During testing, stressed speech vectors $\{\mathbf{y}\}$ of 2^{nd} session of database are tested with speech vectors $\{\mathbf{s}\}$ and stress vectors $\{\mathbf{x}\}$ from different stressed classes. The classification rates of these vectors of four speakers are shown in Figure 3. From this figure, it is observed that the classification rate of speech vector $\{\mathbf{s}\}$ under neutral condition increases whereas the performance of this vector for other stressed conditions decrease from that of $\{\mathbf{y}\}$. Speech vector lies in the neutral speech subspace after decomposition of stressed speech, hence are more neutral than stressed speech. On the other hand, the performance of $\{\mathbf{x}\}$ degrades for neutral speech and for other stress conditions, the performance of $\{\mathbf{x}\}$ more compared to speech vector $\{\mathbf{s}\}$. From these observations, it is inferred that $\{\mathbf{s}\}$ becomes neutral and $\{\mathbf{x}\}$ contains stress information.

IV. Conclusion

In this paper, a subspace projection based approach is proposed and tested for separation of speech and stress information present in stressed speech. The stress information is assumed as orthogonal to the speech information. Speech recognition experiment shows that the estimated speech component contains speech information, whereas, speech information is negligible in stress component. Similarly, stress classification experiment shows that stress component contains stress information of the signal. The speech and stress specific information are present in their respective subspaces which prove orthogonality assumption between these two subspaces. This speech and stress information can be used for various applications in the area of speech processing.

References

[1] Y. Gong, "Speech recognition in noisy environments: A survey", *Speech Commun., vol. 16, pp. 261-291, 1995.*

[2] P. Krishnamoorthy and S. R. M. Prasanna, "Reverberant speech enhancement by temporal and spectral processing", IEEE Trans. Speech, Audio and Lang. Process., vol. 17, pp. 53-266, 2009.

[3] Y. Chen, "Cepstral domain talker stress compensation for robust speech recognition", IEEE Trans. on Acoust., Speech and Signal Process., vol. 36, pp. 433-439, 1988.

[4] S. Ramamohan and S. Dandapat, "Sinusoidal model based analysis and classification of stressed speech", IEEE Trans. Audio, Speech and Lang. Process., vol. 14, pp. 737-746, 2006.

[5] J. H. L. Hansen and A. Clements, "Source generator equalization and enhancement of spectral properties for robust speech recognition in noise and stress", IEEE Trans. Speech Audio Process., vol. 3, pp. 407-415, 1995.

[6] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement", IEEE trans. on speech and audio process., vol. 3, pp. 251-266, 1995.

[7] G. Strang, "Linear algebra and its applications", fourth ed., Cengage Learing, 2006.

[8] S. Shukla and S. Dandapat and S. R. M. Prasanna, "Spectral slope based analysis and classification of stressed speech", Int J Speech Tech., vol. 14, pp. 245-258,2011.

[9] B. A. Carlson and M. A. Clements, "A projection-based likelihood measure for speech recognition in noise", IEEE Trans. Speech and Audio Process., vol. 2, pp. 97-102, 1994.

[10] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect", itIEEE Trans. Speech Audio Process., vol. 2, pp. 598-614, Oct. 1994.

[11] R. P. Lippmann, E. A. Mack and D . B. Paul, "Multi-style training for robust isolated -word speech recognition", Proc.1987 IEEE ICASSP, Apr. 1987, 705-708.

[12] M. Afify, Y. Gong and J. P. Haton, "A general additive and convolutive bias compnesation approach applied to noisy Lombard speech recognition", IEEE Trans. on peech and audio proces., vol. 6, pp. 524-538, 1998.