Australian Journal of Basic and Applied Sciences, 9(16) Special 2015, Pages: 146-154

ISSN:1991-8178

Australian Journal of Basic and Applied Sciences

Journal home page: www.ajbasweb.com

Spectral Transformation of Lombard Speech to Normal Speech for Speaker Recognition Systems

[1]S. Uma Maheswari, [2]J. Divya, [1]A. Shahina, [3]A. Nayeemulla Khan

[1]*Department of Information Technology, SSN College of Engineering, Chennai, India*
[2]*Infosys Limited, Chennai, India*
[3]*School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India*

ARTICLE INFO

Article history:
Received 20 January 2015
Accepted 02 April 2015
Published 20 May 2015

Keywords:
Lombard speech spectral mapping
MLFFNN speaker recognitionGMM

ABSTRACT

In a noisy environment, the speaker tends to increase his/her vocal effort due to the hindrance in the auditory self-feedback, in order to ensure effective communication. This is called Lombard effect. Lombard effect degrades the performance of speech systems that are built using normal speech due to the mismatch between the test (Lombard speech) data and training (normal speech) data. This study proposes a spectral transformation technique that maps the weighted Linear Prediction Cepstral Coefficient (wLPCC) features of the Lombard speech to that of normal speech using Multi-Layer Feed Forward Neural Network (MLFFNN). The efficiency of mapping is objectively tested using the Itakura distance metric. A text independent speaker recognition system is built in the Gaussian Mixture Model (GMM) framework using the normal speech as training data, to test the effectiveness of mapping technique. While the performance of the system when tested with Lombard speech drops to 71%, it improves significantly to 99% when the estimated features are used. Also, a text independent speaker identification system is built using Lombard speech in order to remove the mismatch in training and testing data, shows an improvement in performance from 71% to 89%.

To Cite This Article: S. Uma Maheswari, A. Shahina, A. Nayeemulla Khan and J. Divya., Spectral Transformation of Lombard Speech to Normal Speech for Speaker Recognition Systems. **Adv. in Nat. Appl. Sci.,** *9(16): 146-154, 2015*

INTRODUCTION

A speech system developed for a noise free environment degrades in performance when deployed in a noisy environment. This performance degradation is due to the changes in the acoustic characteristics of speech produced in noisy environment from that of noise free environment. In a noise free environment, the speaker is able to hear his/her own voice level, which is referred as self-feedback. In other words, there exists a feedback between the speech production and the auditory perception of the speaker. In a noisy environment, there is a loss in the level of self-feedback. This is compensated by the speaker by increasing his or her vocal effort to ensure effective communication. This is termed as Lombard effect, which was first described by Etienne Lombard in 1911 (Lombard, 1911). The increase in vocal effort alters the speech production mechanism, thereby causing changes in the acoustic characteristics of the speech produced in noise. The speech produced in a noise free environment is henceforth termed as normal speech and that produced by increasing vocal effort in a noisy environment is termed as Lombard speech. The Lombard speech is different from loud speech, where the speaker deliberately increases his/her vocal effort to increase the audible range (for example, in a large but quiet classroom). However the changes in the acoustic characteristics of loud speech are similar to that of Lombard speech (Bond and Moore, 1990).

Several speech systems are employed in different kinds of noisy environments like battlefield, cockpit, factory, and crowded places to name a few. In such environments, the speaker is influenced by Lombard effect and thereby Lombard speech is fed to the system instead of normal speech of a speaker. If the existing system is not designed to process the Lombard speech, the system performance decreases due to the mismatch in speech characteristics. Consider a scenario where a speaker in a noisy environment communicates with a listener in a noise free environment, which indirectly creates stress on the listeners. In order to improve the listener's comfort and to maintain the naturalness of the speech, it becomes essential to compensate the Lombard speech. This paper aims to address these two issues, namely, compensation of Lombard

Corresponding Author: S. Uma Maheswari, Assistant Professor, Department of Information Technology, SSN College of Engineering, Chennai, India.

speech and performance degradation due to mismatched conditions.

Many researchers have studied the changes in acoustic characteristics of Lombard speech (Bapineedu, 2010; Junqua and Anglade, 1990; Rajasekaran et al., 1986) the influence of Lombard effect on recognition system (Goldenberg et al., 2006; Varadarajan and Hansen, 2006; Varadarajan and Hansen, 2009; Wakao et al., 1996) and adopted various compensation approaches to reduce the performance degradation and can be categorized into (a) Developing features robust to Lombard effect or transforming Lombard speech to normal speech, (b) training or front-end processing and (c) recognizing or back-end processing.

Acoustic analyses have been carried out on source parameters and system parameters like fundamental frequency (Bapineedu, 2010; Junqua and Anglade, 1990), strength of excitation (Bapineedu, 2010), loudness due to glottal excitation (Bapineedu, 2010), formants (Junqua and Anglade, 1990) and spectral tilt (Junqua and Anglade, 1990) among others. Duration of phoneme, word and sentence (Bapineedu, 2010; Junqua and Anglade, 1990; Varadarajan and Hansen, 2006; Varadarajan and Hansen, 2009) energy across various frequency bands (Junqua and Anglade, 1990), zero-crossings (Junqua and Anglade, 1990) and several other parameters have also been considered. Almost all the results from the various acoustic analysis of Lombard speech have shown similar changes for each of the acoustic parameters considered. Some of the changes in the acoustic characteristics of Lombard speech are increase in fundamental frequency ((Bapineedu, 2010; Junqua and Anglade, 1990) energy (Bapineedu, 2010; Junqua and Anglade, 1990) loudness (Bapineedu, 2010), duration of vowels (Bapineedu, 2010; Junqua and Anglade, 1990; Varadarajan and Hansen, 2009) and decrease in spectral tilt (Varadarajan and Hansen, 2006; Varadarajan and Hansen, 2009) duration of consonants (Bapineedu, 2010; Junqua and Anglade, 1990; Varadarajan and Hansen, 2009) and silence (Varadarajan and Hansen, 2006; Varadarajan and Hansen, 2009) and strength of excitation (Bapineedu, 2010) as compared to normal speech.

The performance degradation caused by Lombard effect on a recognition system is more than that caused by additive noise (Rajasekaran et al., 1986). An Automatic Speech Recognition System (ASR) (Boril and Hansen, 2010) showed a word error rate of 8.7% and 37.7% for female normal and Lombard speech respectively, and of 8.7% and 32.8% for male normal and Lombard speech, respectively.

A few compensation techniques developed in the past to improve the recognition rate are summarized as follows: The preprocessing steps that includes successive speech enhancement and stress compensation based on formant location, bandwidth and intensity produce speech features that are resistant to stress and noise, gave an improvement of 42% in recognition rate for Lombard condition (Hansen and Clements, 1989). A GMM based speaker verification system (Goldenberg et al., 2006) showed performance degradation by 10.1% due to Lombard effect. Two types of compensation techniques was adopted based on robust speech features (ExpoLog) and transformation, where Lombard speech is transformed back to normal speech. By applying these techniques, an improvement in recognition error rate of 13.9% (from 22.3% to 8.4%) and 5.4% (from 13.5% to 8.5%) respectively had been achieved. In another study (Lippmann et al., 1987), a multi-style training approach based on different talking styles inclusive of Lombard speech had been carried out. Collecting data representing all different types of noise is a major challenge for this approach. Study on family of distortion measures such as first order norm equalization, frame-optimal adaptive equalization, Euclidean distance between normalized vectors and cepstral norm weighting are used to project either the test cepstral vectors closer to that of Lombard cepstral vectors or vice versa (Carlson and M. A. Clements, 1992). This similarity measure outperformed the traditional Euclidean measure when tested under noisy condition and trained with clean speech. In another study, where a degradation model representing the spectral changes (variation of formant location, formant bandwidth, spectral tilt and energy) of Lombard speech in each frequency band had been developed using non-linear warping and amplitude scaling function, and multiple linear transformation had been used to estimate the clean speech from that of Lombard speech (Chi and Oh, 1996).

In this paper we analyze some of the acoustic characteristics of Lombard speech as compared to normal speech and the influence of Lombard effect on speaker recognition system. To improve the performance of a speaker recognition system a spectral transformation technique is employed to map the Lombard speech features to that of normal speech features. The effectiveness of mapping is measured by Itakura distance metric and a text independent speaker identification system is built with normal speech data to measure the performance. Also, a text independent speaker recognition system is built with Lombard speech in order to remove the mismatch in training and test conditions. The weighted LPCC parameter of order 19 is used in this study, as it shows higher accuracy than LPCC (Zhu and Yang, 2012).

This paper is organized as follows: Section 2 describes some of the acoustic characteristics of Lombard speech and normal speech. Section 3 deals with the database used in this study and feature extraction. The mapping technique is also discussed. Section 4 describes the speaker recognition system.

The performance of the speaker recognition system is discussed in Section 5. Section 6 summarizes the work.

1. Acousic Analysis of Lombard speech and normal speech:

The influence of the Lombard effect on a speaker degrades the performance of speech systems. It is necessary to study the variation in the acoustic characteristics of Lombard speech and normal speech so that an appropriate compensation technique can be incorporated to improve the efficiency of the speech system. The acoustic features such as fundamental frequency, duration and normalized energy, are compared for Lombard speech and normal speech for the TIMIT sentence *she had your dark suit in greasy wash water all year* collected from 6 (3 male and 3 female) speakers.

1.1 Fundamental Frequency:

The fundamental frequency (f0) is shown to increase for Lombard speech (Bapineedu, 2010). The increase in fundamental frequency for male speakers is high compared to female speakers (Junqua and Anglade, 1990). Fig.1 shows the f0 contours for normal speech and Lombard speech, respectively, for the above-mentioned sentence. It clearly shows that there is an increase in fundamental frequency for Lombard speech. The average fundamental frequency is about 134 Hz and 160 Hz for normal and Lombard speech, respectively. The results show that the increase in average fundamental frequency for Lombard speech is about 13% for male speakers and 9% for female speakers.

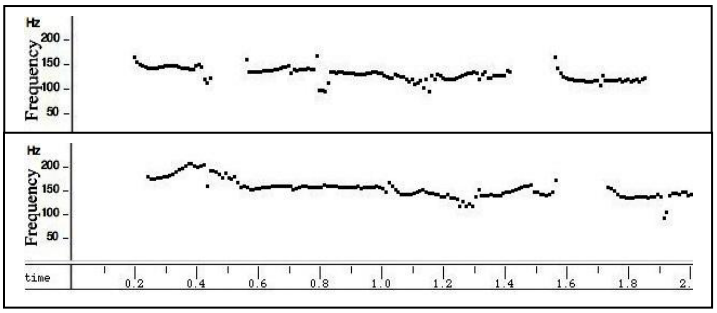

Fig. 1: f0 contours of (a) normal speech (b) Lombard speech for a TIIMIT sentence *she had your dark suit in greasy wash water all year*

1.2 Duration:

The duration of vowels increases and that of consonants decreases for Lombard speech as compared to normal speech. This causes increase in sentence duration for Lombard speech (Bapineedu, 2010). Contrary to this, it is also noted that the decrease in silence region decreases the sentence duration (Varadarajan and Hansen, 2006; Varadarajan and Hansen, 2009). In this study, both phoneme and sentence duration are analyzed. The observation of phoneme duration follows the results of other studies. Both increase and decrease in sentence duration is observed. The decrease in duration is due to the decrease in silence duration, i.e., the speaking rate increases significantly in the presence of noise. Thus the duration of the sentence depends upon the sound units in the speech and the psychological effect on speakers in the presence of noise.

1.3 Normalized Energy:

Normalized energy increases for vowels in Lombard speech (Bapineedu, 2010) and decreases for fricatives, nasals, plosives and affricatives (Junqua and Anglade, 1990). In this study, the normalized speech signal is segmented into frames of 20ms duration with frame shift of 5ms duration. The energy is calculated for each frame. Fig.2 shows the normalized energy for normal speech and Lombard speech for the above-mentioned sentence respectively. It clearly shows that there is a significant increase in energy for Lombard speech as compared to normal speech.

2. Mapping Lombard speech to normal speech:

This section deals with the mapping of spectral features of Lombard speech to normal speech and the objective evaluation.

2.1 Database for this study:

The Database used in this study consists of speech data collected from 20 speakers (10 male and 10 female speakers). The recordings are done in laboratory environment using a sampling frequency of 16 KHz. Lombard speech is recorded with the speaker wearing a close-ear headphone through which white noise (85dB) is played.

Four minutes of training data is collected from each speaker for both normal speech and Lombard speech. The training data (both for normal speech and Lombard speech) is obtained in two sessions, each of 2 minutes duration, as the speaking rate varies tremendously for prolonged speech. Five test cases each of 20ms duration are obtained for both normal and Lombard speech.

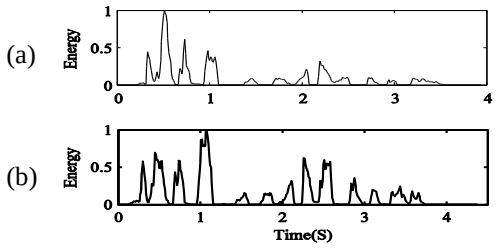

Fig. 2: Normalized energy for (a) normal speech (b) Lombard speech

2.2 Preprocessing:

Since the speaking rate varies and hence the duration varies for the Lombard speech and normal speech, the two speech signals need to be time aligned to enable mapping. Dynamic Time Warping (DTW), a time-alignment technique is used to warp the Lombard speech according to normal speech (Turetsky and Ellis, 2003). The speech is then segmented into frames of 20ms duration with a frame-shift of 5ms duration. The segmented frames are categorized as voiced frames or unvoiced frames using Voice Activity Detection (VAD) algorithm (Sohn et al., 1999). For the voiced frames, a 19 dimensional feature vector represented by weighted Linear Prediction Cepstral Coefficient (wLPCC) is extracted for the entire speech data, as explained in the next section.

2.3 Feature Extraction:

Linear prediction analysis of speech is used to extract the speech features. An 8th order LP analysis is performed on speech frames of 20ms duration with an overlap of 5ms duration.

In LP analysis, the nth speech sample is predicted by the past p samples of speech and is given by (Rabiner and Juang, 1993),

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

where s(n) is the speech sample at time n, and $\{a_k,\}$, k = 1,2,.....p, is the set of predictor coefficients.

Linear prediction cepstral coefficients (LPCC) were obtained from the linear prediction coefficients (a_k) directly as (Rabiner and Juang, 1993),

$$c_0 = \ln s^2 \tag{2}$$

$$c_m = a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad 1 \le m \le p \tag{3}$$

$$c_m = \sum_{k=1}^{m-1} \left(\frac{k}{m}\right) c_k a_{m-k}, \quad m > p \tag{4}$$

where s^2 is the gain term in the LPC model.

The weighted Linear Prediction Cepstral Coefficient (wLPCC) are obtained from LPCC and is given by,

$$w_i = i c_i, \qquad 1 \le i \le m \tag{5}$$

The 14th order LP spectrum for a voiced frame for normal speech and Lombard speech is shown in Fig.3 and Fig.4.

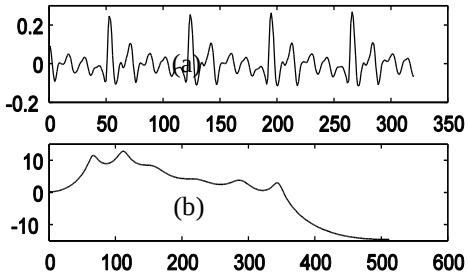

Fig. 3: (a) A normal speech frame, (b) LP spectrum of (a)

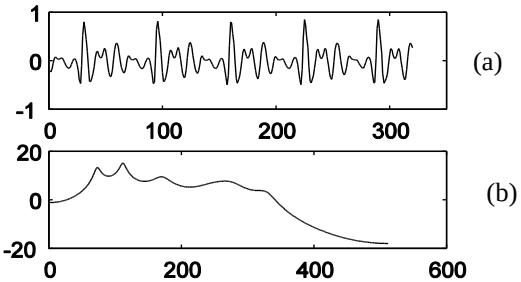

Fig. 4: (a) The Lombard speech frame and (b) LP spectrum of (a)

2.4 Training and Testing using MLFNN:

The mapping of Lombard speech features to normal speech features has two stages. In the training stage, the wLPCCs extracted from Lombard speech are mapped to the wLPCCs extracted from the corresponding normal speech. That is, the wLPCCs derived from the Lombard speech are used as input to the mapping network, while the wLPCCs derived from the normal speech form the desired output. Speaker dependent models are built for each speaker in this stage. The Multi Layered Feed Forward Neural Network (MLFFNN) is used to learn the mapping (Shahina and Yegnanarayana, 2007). In the testing stage, the wLPCCs derived from a test Lombard speech utterance are given as input to the trained network model.The output from the network is the estimated wLPCCs of the normal speech corresponding to the test input. The block diagram of the training and testing phases are shown in Fig. 5.

Multi-Layered Feed-Forward Neural Network:

Given a set of input and output pattern pairs, the objective is to capture the unknown system behaviour from the samples of the input-output data pair. Once the system behaviour is captured by the network, it would produce a possible output pattern for the new given input pattern (test data) which is not used in the training set. The output patterns would be an interpolated version of the output pattern corresponding to the input training patterns which are closest to the test input pattern.

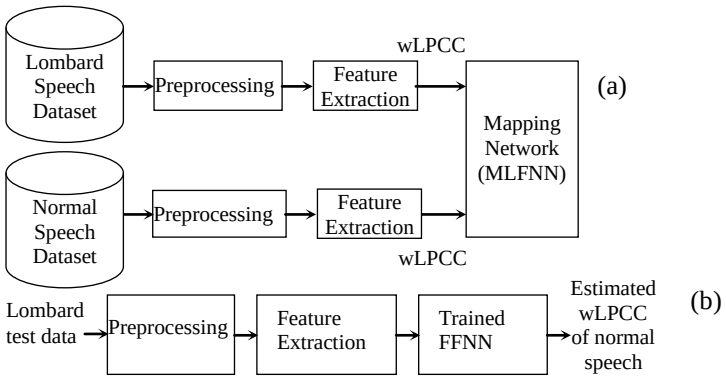

Fig. 5: (a) The block diagram of training stage, x_1: training input, x_2: training output (b) testing stage

A multi layered feed forward neural network with two intermediate layers in addition to the input and output layers can perform a pattern mapping task (Yegnanarayana, 1999). The additional intermediate layers are called as the hidden layers. The number of units in the hidden layers depends on the complexity of the mapping problem. The neurons in the input and output layers are linear units whereas the neurons in the hidden layers are nonlinear units.

The pattern mapping problem involves determining the interconnection weights, given a training set consisting of input-output pattern pairs. The weights are updated for each input-output pattern pair using back propagation algorithm. A 19L 38N 38 N 19L network structure where L and N refer to linear unit and non-linear unit, respectively is shown in the Fig.6. Each MLFFNN is trained using 200 epochs, for which the mean square error between the actual output and the training output is minimum.

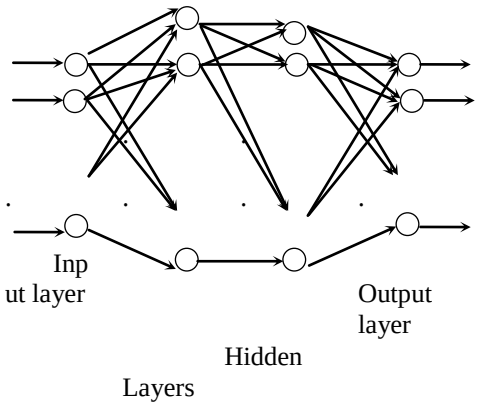

Fig. 6: A 4-layer mapping neural network

Objective evaluation of mapping:

The effectiveness of mapping is studied by observing the Log magnitude spectra. The Log magnitude spectra of the test input Lombard speech, and the corresponding normal speech (desired) and the estimated Log magnitude spectra for two speech frames are shown in Fig.7. The estimated spectrum is observed to be similar to the normal speech spectrum.

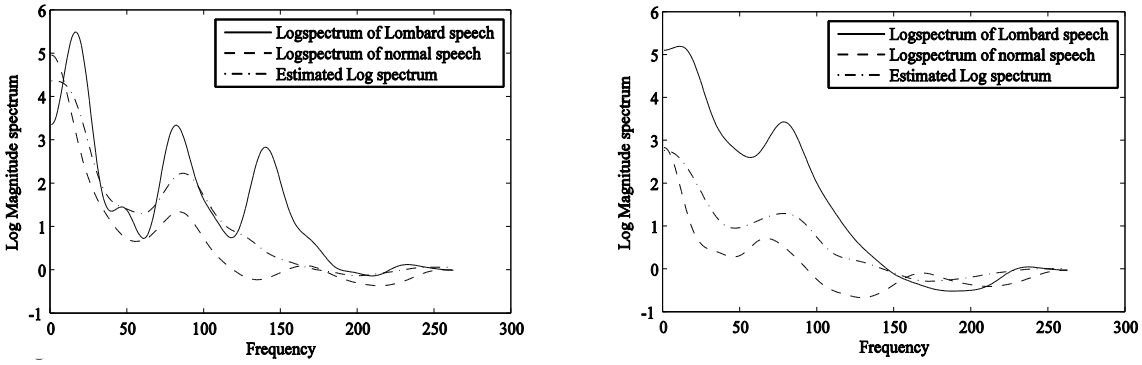

Fig. 7: Log magnitude spectra for two speech frames of the Lombard, normal and estimated features. The Log magnitude spectra of the normal &estimated features are seen to be similar.

The performance of the mapping technique is evaluated using the Itakura distance measure. The Itakura distance measures the distance between two Log spectra. The Itakura distance between two LP vectors, say a_k and b_k is given by (Deller et al., 1993)

$$d_{ab}\left[a_k, b_k\right] = \frac{b_k^T \tilde{R}_{s_a} b_k}{a_k^T \tilde{R}_{s_a} a_k} \tag{6}$$

$$d_{ba}\left[a_k, b_k\right] = \frac{a_k^T \tilde{R}_{s_b} a_k}{b_k^T \tilde{R}_{s_b} b_k} \tag{7}$$

where d_{ab} and d_{ba} are the asymmetric distances from a_k to b_k and vice versa., respectively. $\tilde{R}_{s_a} = \{r_{s_a}\}$ and $\tilde{R}_{s_b} = \{r_{s_b}\}$, where $\{r_{s_a}\}$ and $\{r_{s_b}\}$ are the signal auto correlation coefficients of the speech frames corresponding to a_k and b_k respectively. The symmetric Itakura distance between the two vectors is given by, $d = 0.5\left(d_{ab} + d_{ba}\right)$.

The Itakura distance between the spectrum of the normal speech and the estimated spectrum, and the spectrum of the normal speech and that of the Lombard speech are computed for each frame. The Itakura distance plot for two utterances is shown in Fig.8. It is observed from the plots that the distance between the normal speech spectrum and the estimated spectrum is small when compared to the distance between the normal speech spectrum and Lombard speech spectrum. This shows that the estimated spectrum is close to the normal spectrum. Thus the mapping network is able to capture the spectral correlation between the normal and the Lombard speech.

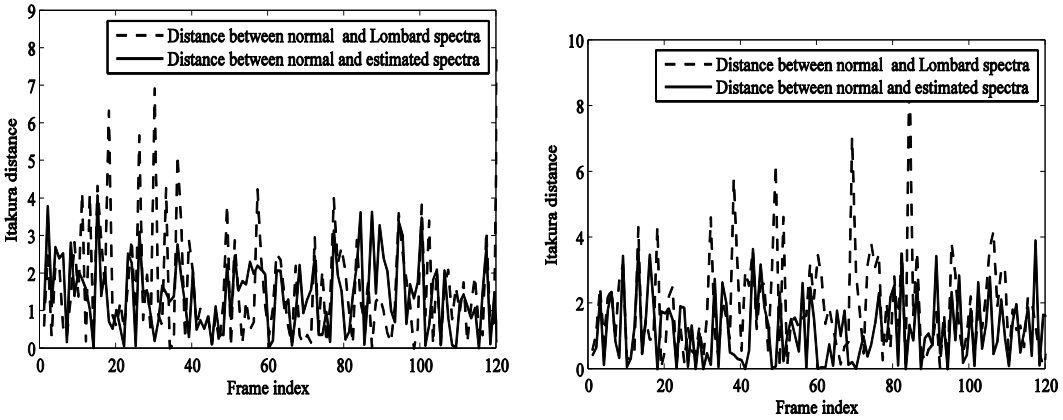

Fig. 8: Itakura distance between normal and Lombard spectra (dotted lines) and Itakura distance between normal and estimated spectra (solid lines) for two different utterances.

3. Speaker Recognition System:

The effectiveness of the spectral mapping is measured by building a speaker identification system using the spectral features of the normal speech and tested using the estimated spectral features. A speaker identification system identifies a target speaker from a group of known speakers. The speaker identification system consists of two phases namely enrollment or training phase and identification or testing phase. In the enrollment phase, a user enrolls by providing his/her voice samples to the system and the system extracts the speaker-specific features i.e. wLPCC (derived from LP coefficients obtained using 12^{th} and 14^{th} order LP analysis, as discussed in Section 2.3). For each enrolling speaker, a speaker model is built. In the identification phase, a user provides a test sample, the system extracts the features and then a similarity measure based on the distance metrics or probability is measured between the test sample and to each of all the stored speaker models. The speaker associated with the model that has smallest distance or highest probability is referred to as the target speaker.

Gaussian Mixture Model (GMM) based speaker models are built using normal speech. The enrollment process and identification process for the GMM based speaker identification system is depicted in Fig.9. In the testing phase, three test cases are used; normal speech, Lombard speech and estimated features (obtained from mapping). A separate speaker recognition system is built using the Lombard speech training data to overcome the mismatch condition. The results of this study are discussed in Section 5.

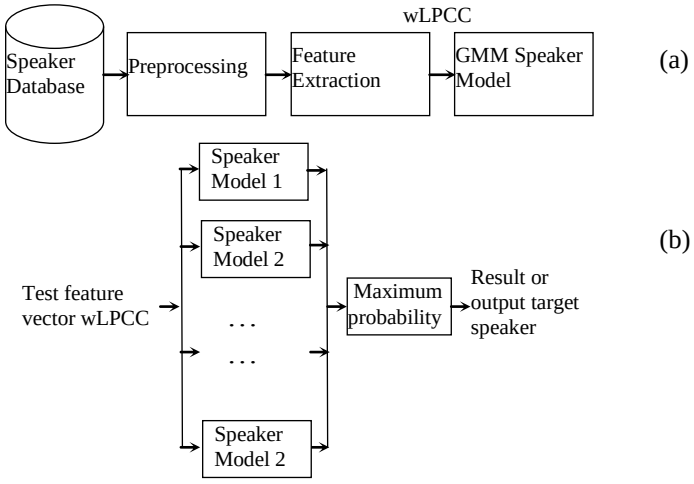

Fig. 9: (a) Enrollment phase or training phase (b) Identification phase or testing phase

3.1 GMM for speaker models:

Gaussian Mixture Model (GMM) is a parametric model of the probability distribution. The Gaussian probability density function (pdf) of the feature vector, $\underline{x}$ is given by (Quatieri, 2006).

$$b_i(\underline{x}) = \frac{1}{(2\pi)^{\frac{R}{2}} \left|\Sigma_i\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1} (\underline{x} - \underline{\mu}_i)} \quad (8)$$

where $\underline{\mu}_i$ is the state mean vector, Σ_i is the covariance matrix and R is the dimension of the feature vector.

The probability of the feature vector being in any of *I* speaker models denoted by λ, is

$$p(\underline{x} \mid \lambda) = \sum_{i=1}^{I} p_i \, b_i(\underline{x}) \quad (9)$$

where p_i are the mixture weights and $\sum_{i=1}^{I} p_i = 1$.

For each speaker a GMM model is represented by GMM mean, covariance and a weight parameter given by,

$$\lambda = \{p_i, \underline{\mu}_i, \Sigma_i\} \quad (10)$$

For a large set of training feature vectors, $x = \{\underline{x}_0, \underline{x}_1, \ldots \underline{x}_n\}$ of a particular speaker, the GMM is the union of Gaussian pdfs, which is estimated by maximum likelihood estimation. Expectation-Maximization (EM) algorithm improves the GMM parameter estimates.

In speaker identification, the probability of each speaker model given the feature vector $\underline{x}_n$ is, computed as,

$$P(\lambda_j \mid \underline{x}_n) = \frac{p(\underline{x}_n \mid \lambda_j)\, P(\lambda_j)}{P(\underline{x}_n)} \quad (11)$$

If the test utterance has M number of feature vectors, the likelihood of an utterance belonging to a particular model is the product of likelihoods for each frame, given by,

$$P(\{\underline{x}_0, \underline{x}_1, \ldots \underline{x}_{M-1}\} \mid \lambda_j) = \prod_{m=0}^{M-1} p(\underline{x}_m \mid \lambda_j) \quad (12)$$

The target speaker is the speaker associated with speaker model with highest probability.

4. Performance Measure:

The performance of the proposed method is

measured in terms of percentage of the number of test utterances identified correctly out of the total test utterances. Table 1 shows the performance measure for the system trained with normal speech and tested using normal, Lombard and estimated features. The performance is low for the Lombard test case due to the mismatch in the training and testing conditions. The estimated features significantly improve the performance of the speaker identification system for both the 12^{th} order and 14^{th} order LP analysis as seen in Table 1. The performance of the identification system, where the speaker models are trained using normal speech is shown in detail in Table 2. It shows a better identification performance for normal test case as compared to the Lombard test case. The performance of the identification system for both the LP orders (12 and 14) clearly shows the performance degradation due to Lombard effect. In order to remove the mismatch between the training and testing conditions a separate GMM based speaker identification system is built using Lombard speech. The performance of this identification system is shown in Table 3. It shows a significant improvement in the identification rate.

A third speaker identification system is built using 38 dimensional feature vectors, obtained by combining the normal feature vectors and Lombard feature vectors. The performance of the identification system trained using combined feature vectors and tested against combined test utterances are shown in Table 4. This performance result shows a marginal improvement over that of the system trained and tested using normal speech. The graph in Fig.10 shows an improvement in the performance achieved using combined features of normal and Lombard speech over other systems. This shows the presence of some speaker-specific complementary information in the spectral features of normal and Lombard speech.

Table 1: Performance of the speaker identification system for speaker models trained using normal speech and tested using normal, Lombard and estimated features

Test Case	Identification rate (%)	
	LP order 12	*LP order 14*
Normal	92	94
Lombard	70	71
Estimated Parameter	99	99

Table 2: Performance (in %) of normal and lombard test utterances tested against models trained using normal speech

Test Case		Identification rate (%)	
		LP order 12	*LP order 14*
Normal	Male	94	99
	Female	90	88
	Average	92	94
Lombard	Male	76	84
	Female	64	58
	Average	70	71

Table 3: Performance (in %) of lombard test utterances against models trained using lombard speech

Lombard Test Case	Identification rate (%)	
	LP order 12	*LP order 14*
Male	99	99
Female	74	78
Average	87	89

Table 4: Performance (in %) of combined models trained and tested against combined test features against models trained using combined feature vector

	Identification rate (%)	
	LP order 12	*LP order 14*
Male	99	99
Female	86	92
Average	93	96

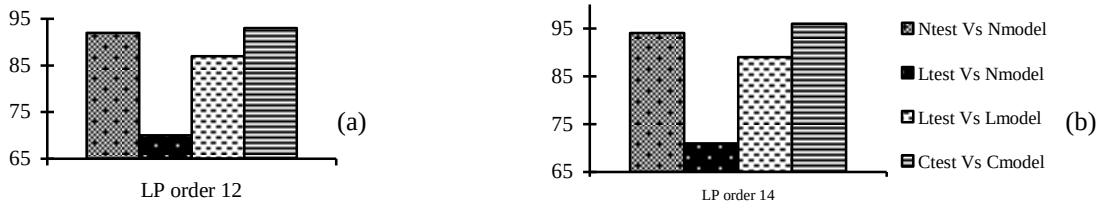

Fig. 10: Performance (in %) of four speaker recognition systems for (a) LP order 12 and (b) LP order 14. System using combined feature vectors of normal and Lombard speech performs marginaally better for both LP order 12 and 14. N, L and C indicates normal, Lombard and combined features of normal and Lombard speech respectively.

Conclusion:

In this paper, we have analyzed the changes in acoustic characteristics such as fundamental frequency, duration and normalized energy of Lombard speech from normal speech. The impact of Lombard effect on the performance of the speaker identification system is studied and a spectral transformation technique to map the Lombard speech to normal speech is performed. The efficiency of mapping technique is analyzed using Itakura distance metric and the plots shows that the estimated parameters are close to that of normal parameters. A text independent speaker identification system is built using GMM with the normal speech data to measure the effectiveness of mapping technique. Also a text independent speaker identification system with Lombard speech data is built using GMM to remove the mismatch in training and testing conditions. Both the approaches show significant improvement in the recognition rate.

REFERENCES

Bapineedu, G., 2010. Analysis of Lombard effect speech and its application in speaker verification for imposter detection, M.S. Thesis, International Institute of Information Technology, Hyderabad, India.

Bond, Z.S. and T.J. Moore., 1990. A note on loud and Lombard speech, ICSLP'90, pp: 969-972.

Boril, H and J.H.L. Hansen, 2010. Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environment, IEEE Transactions on Audio, Speech, and Language Processing, 18(6): 1379-1393.

Carlson, B.A. and M.A. Clements, 1992. Speech recognition in noise using a projection based likelihood measure for mixture density HMM's, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'92), 1: 237-240.

Chi, S.M. and Y.H. Oh, 1996. Lombard effect compensation and noise suppression for noisy Lombard speech recognition, in Proc. Fourth International Conference on Spoken Language (ICSLP'96), 4: 2013-2016.

Deller, J.R., J.G. Proakis and J.H.L. Hansen, 1993. Discrete-Time processing of speech signals, Macmillan, NewYork, NY, USA.

Goldenberg, R., A. Cohen and I. Shallom, 2006. The Lombard effect's influence on automatic speaker verification systems and methods for its compensation, International Conference on Information Technology: Research and Education, pp: 233-237.

Hansen, J.H.L. and M.A. Clements, 1989. Stress compensation and noise reduction algorithms for robust speech recognition, International Conference on Acoustics, Speech and Signal Processing (ICASSP-89), 1: 266-269.

Junqua, J.C. and Y. Anglade, 1990. Acoustics and perceptual studies of Lombard speech: Application to isolated words automatic speech recognition, International Conference on Acoustics, Speech and Signal Processing (ICASSP-90), 2: 841-844.

Lippmann, R.P., E.A. Martin and D.B. Paul, 1987. Multi-style training for robust isolated-word speech recognition, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'87), 12: 705-708.

Lombard, E., 1911. Le signe de l'elevation de la voix, annals maladies oreille, Larynx, Nez, Pharynx, 37: 101-119.

Quatieri, T.F., 2006. Discrete-Time speech signal processing, Pearson Education, ch 14, sec 14.3, pp: 719-725.

Rabiner, L.R. and B.H. Juang, 1993. Fundamentals of speech recognition, Prentice- Hall, Englewood Cliffs and N.J.

Rajasekaran, P., G. Doddington and J. Picone, 1986. Recognition of speech under stress and in noise, IEEE International conference on Acoustics, Speech and Signal Processing (ICASSP'86), 11: 733-736.

Shahina, A and B. Yegnanarayana, 2007. Mapping speech spectra from throat microphone to close-speaking microphone: A Neural Network Approach, EURASIP Journal on Advances in Signal Processing.

Sohn, J., N.S. Kim and W. Sung, 1999. A statistical model-based voice activity detection, IEEE Signal Processing Letters, 6(1): 1-3.

Turetsky, R and D. Ellis, 2003. Ground-Truth transcriptions of real music from force- aligned MIDI syntheses, 4th International Symposium on Music Information Retrieval ISMIR-03, pp: 135-141, Baltimore.

Varadarajan, V.S and J.H.L. Hansen, 2006. Analysis of Lombard effect under different noise types and level of noise with application to in-set speaker ID systems, in Proc. INTERSPEECH, ISCA, pp: 937-940.

Varadarajan, V.S and J.H.L. Hansen, 2009. Analysis and compensation of lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition, IEEE Transactions on Audio,Speech and Language Processing, 17(2): 366-378.

Wakao, A., K. Takeda and F. Itakura, 1996. Variability of Lombard effects under different noise conditions, in Proc. International Conference on Spoken Language Processing (ICSLP-96), 4: 2009-2012.

Yegnanarayana, B., 1999. Artificial neural networks, Prentice-Hall, New Delhi, India.

Zhu, L and Q. Yang, 2012. Speaker recognition system based on weighted feature parameter, International Conference on Solid State Devices and Materials Science, 25: 1515-1522.