ANALYSIS OF LOMBARD EFFECT SPEECH AND ITS APPLICATION IN SPEAKER VERIFICATION FOR IMPOSTER DETECTION

by

G. BAPINEEDU

200402013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Master of Science (by Research)
in
Computer Science and Engineering

Speech and Vision Lab.

Language Technologies Research Centre

International Institute of Information Technology

Hyderabad, India

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Analysis of Lombard effect speech and its application in speaker verification for imposter detection**" by **G. Bapineedu (200402013)**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. B. Yegnanarayana

Abstract

Speaking in the presence of noise changes the characteristics of the speech produced which is known as the Lombard effect. This effect is perceptually felt with an increase in intensity of speaking. These changes in the characteristics of speech production is to ensure an intelligible communication in noisy environment. These changes also result in the performance degradation of speech systems like speaker recognition, speech recognition, etc. Human speech production mechanism is affected due to the Lombard effect, and is reflected mainly in the excitation source. Previous studies have focussed mainly on the changes in system level features. In our work, we examine the excitation source to study its changes due to Lombard effect.

The excitation source features studied in this work are the instantaneous fundamental frequency F_0 (i.e., pitch), the strength of excitation at the instants of significant excitation and a loudness measure reflecting the sharpness of the impulse-like excitation around epochs. Another feature, called the normalized energy is used to study the articulation variability of speech due to Lombard effect using the high variations of energy within an utterance. Analysis is performed using the distributions of the features which are seen to distinguish between normal speech and Lombard effect speech, and also to explain the speaker-specific nature of Lombard effect. The extent of Lombard effect on speech depends on the nature and intensity of the noise that causes the Lombard effect. Characteristics of speech produced without a feedback and difference between loud speech and Lombard effect speech are also studied using the excitation source features. Duration is used to study the phonetic changes due to Lombard effect. Intelligibility due to Lombard effect at a sentence level is studied and the mechanism of the Lombard effect is described based on the analysis performed. This analysis is extended into an application where the imposters during speaker verification can be detected with the theory that speech produced under noise is different from that spoken under normal conditions. Further, the performance of text-dependent speaker verification system due to Lombard effect is also described.

Keywords: *Lombard effect, fundamental frequency, strength of excitation, zero-frequency resonator, loudness, intelligibility, speaker verification, imposter, playback speech*

Contents

List of Tables

List of Figures

Abbreviations

DET	- Detection error trade-off
DTW	- Dynamic time warping
EER	- Equal error rate
EGG	- Electroglottograph
FAR	- False acceptance rate
FIR	- Finite impulse response
HMM	- Hidden Markov model
IFT	- Inverse Fourier transform
KL	- Kullback-Leibler
LP	- Linear prediction
LPCC	- Linear prediction cepstral coefficients
MDR	- Missed detection rate
MFCC	- Mel-frequency cepstral coefficients
NF	- No feedback
NE	- No effect
NPZC	- Negative to positive zero crossing
PLP	- Perceptual linear prediction
PNZC	- Positive to negative zero crossing
RFCC	- Repartitioned frequency cepstral coefficients
SEF	- Single ear feedback
SNR	- Signal-to-noise ratio

Chapter 1

Introduction to Lombard Effect

Humans are the most powerful communicators. The best way of communication among humans is through speech. The speech produced by a person depends on several factors, which include the environment he/she is speaking and the auditory self-feedback of the speech of his/her own voice. Adverse environment not only corrupts the speech signals by additive noise, but they also affects the self-feedback of the speech of the person. Lack of self-feedback also affects the articulatory movement in the speech production process, resulting in speech which the listener perceives as not normal. The speaker tries to adjust the articulatory and acoustic parameters to produce speech as intelligible as possible to the listeners. This psychological effect on speaker for producing speech in the presence of noise is termed as Lombard effect, which was first discovered by Etienne Lombard in 1911 [1]. The Lombard effect not only affects the intelligibility in speech communication, but it also affects the performance of automatic speech and speaker recognition systems.

The Lombard effect on speech depends on the environment, speaker and the context of speech communication. Lombard effect is caused due to hampering of self feedback and not just speaking in the presence of noise. The self-feedback can be hampered by various types of noises ranging from a low intensity air flow noise to a high intensity fighter-cockpit noise. The resulting instability is compensated by modification of the speech produced, which is termed as the Lombard effect speech. Analysis of Lombard effect speech signal is based on time domain properties such as duration of voiced and unvoiced segments, and spectral domain properties such as spectral tilt and formants. The only source parameter used extensively is the variation of the fundamental frequency (F_0). On the other hand, perceptually several factors are noticed like loudness, stress and intensity. But very few attempts have been made in reporting the changes in the excitation source information due to Lombard effect.

With increasing use of speech systems for several applications, it is essential to make speech synthesis systems as natural as possible, and to incorporate robustness into speech recognition and speaker recognition systems. It is also required to enhance the speech and make it intelligible, independent of the environment. Since features extracted from the Lombard effect speech are different from those obtained from the normal speech, the affected features need to be compensated when using the speech systems designed for normal speech. For this, modifications at the signal or parameter or feature levels have to be performed by determining the level of compensation required. The first step in developing the process of modification is the analysis of features of the Lombard effect speech in relation to the normal speech. The noise signals are presented to the speaker through earphones which do not allow any external sound to pass through them, and also do not allow the presented noise signals to leak out. The Lombard effect speech is recorded using a close speaking microphone. Hereafter, the noise which causes Lombard effect is termed as external feedback,as speaking under the influence of speech of another person also causes Lombard effect.

1.1 Issues addressed in this thesis

The main objective of the present study is to analyze the Lombard effect speech in terms of the features of excitation source in speech production, when the speech is produced under different types and levels of degradation. We use 3 types of noise at 3 different intensities, which are pink noise at 70, 60, 50 dB, babble noise at 65, 55, 45 dB and factory noise at 65, 55, 45 dB. We also examine several cases relating to Lombard effect, which include external feedback through a single ear, no self-feedback and no effect case where the speaker pretends that he is not under the influence of an external feedback. The difference between loud speech and Lombard effect speech is examined. Changes in duration and energy due to Lombard effect are also examined. Intelligibility of speech in noise is also studied. Further the analysis of Lombard effect will be handful to describe the mechanism of Lombard effect. Another objective of the study is to improve the performance of the speaker verification system by detecting imposters using the Lombard effect. In the process, the performance of speaker verification system due to Lombard effect is also observed.

1.2 Organization of the thesis

The evolution of ideas presented in this thesis is listed in Table 1.1. This thesis is organized as follows:

In Chapter 2, previous studies on Lombard effect are reviewed. These studies are categorized into several parts which include mechanism of Lombard effect, characteristics of Lombard effect speech, intelligibility of Lombard effect speech and performance of speech systems due to Lombard effect.

In Chapter 3, extraction of the features which are used for the analysis of Lombard effect speech are described. The features used are fundamental frequency, strength of excitation , loudness measure which form the excitation source features apart from another feature which is the normalized energy.

In Chapter 4, different cases of speech produced under different conditions are described which are speaking under different types of noise at varying intensities, speaking with a noise fed to a single ear while normal hearing with the other ear, speaking without a self feedback, speaking in the presence of noise by pretending we are not affected by Lombard effect. Finally difference between loud speech and Lombard effect speech are also studied. Their characteristics are also described using the proposed excitation source features.

In Chapter 5, analysis of Lombard effect speech is performed based on the proposed features and also an additional feature which is the change in duration is used. Intelligi-

Table 1.1: Evolution of ideas presented in the thesis

- Lombard effect affects the excitation source which is reflected in the excitation source features.
- The extent of Lombard effect depends on the type and intensity of the external feedback. This can be analyzed based on the change in the excitation source features.
- Lombard effect also results in phonetic changes which are reflected in duration and energy.
- Lombard effect speech in noise is more intelligible than normal speech in noise.
- Lombard effect speech as reference and test speech gives better results compared to normal speech as both reference and test speech.
- An impersonator cannot mimic the Lombard effect speech of another speaker. Thus Lombard effect can be used to avoid imposters.

bility and perception of Lombard effect speech is also studied. Finally, the mechanism of the Lombard effect speech is discussed based on the analysis performed.

In Chapter 6, a text dependent speaker verification system is described and a method is proposed using Lombard effect to detect imposters. The performance of Lombard effect speech on speaker verification is also seen.

Chapter 7 presents the summary of the work, major contributions of the work and outlines the directions for further research.

Chapter 2

Lombard Effect - A Review

This chapter reviews some of the previous studies on Lombard effect. The major studies are related to the analysis of Lombard effect at acoustic, phonetic and perception level. A number of studies were also made on the compensation techniques on Lombard effect speech for improving the performance on speech systems. Section 2.1, reviews the mechanism of Lombard effect as described by several researchers. In Section 2.2, characteristics of Lombard effect speech are described . In Section 2.3, previous studies on intelligibility of Lombard effect speech are discussed. In Section 2.4, few studies on psychological effects caused due to the Lombard effect are described. In Section 2.5, we review the performance degradation of speech systems due to Lombard effect and several compensation methods used to improve the performance. In Section 2.6, importance for analyzing the Lombard effect are addressed.

2.1 Mechanism of the Lombard effect

As a first step it is essential to study the cause of Lombard effect. The phenomenon is claimed to be due to an automatic auditory regulating device [2][3][4]. It is also claimed to be an emphasis on the speakers response to the listeners [5][6]. For few others, the mechanism of the Lombard effect is viewed as a combination of both [7][8].

Auditory feedback is considered to be an element of the human speech production system [2]. It was suggested that the mechanism responsible for the Lombard effect is a result of noise being introduced into the auditory feedback channel [3]. This is based on the model proposed in [4] where the human communication mechanism is viewed as a servosystem, a kind of self-regulatory feedback system. But this proposal is claimed to be a misinterpretation of the Lombard effect. It was claimed to be dependent only on the communication factor [6]. In adverse conditions, the speaker suffers from reduced intelligibility. Thus the reason for exhibiting Lombard effect in such conditions is to ensure intelligible communication by compensating for the reduced signal to noise ratio (snr). Also the extent of compensatory strategies depends on how demanding the communicative situation is [6]. The magnitude of responses may also be dependent on how much the speaker feels responsible for the communication [9].

Thus the notion of auditory feedback control is claimed to be insufficient to describe the phenomenon. It is widely accepted that a cause for Lombard effect is to ensure intelligible communication, but it has always been unclear whether to assume the human auditory system as a servosystem.

2.2 Characteristics of Lombard effect speech

Several studies have been reported on the analysis of Lombard effect speech. Several acoustic-phonetic analysis were performed in [10]. Some of the changes observed between normal speech and Lombard effect speech are:

- Increase in duration of vowels and decrease in duration of unvoiced sounds [10].
- Decrease in the spectral tilt with relatively more energy in the high frequency region of the spectrum [10].
- Increase in pitch or the fundamental frequency (F_0) and the first formant in some vowels [8][11].
- Great lung volumes are used [12].

- Migration of energy from low and high frequency to middle range for vowels, and from low to high frequency for unvoiced stops and fricatives [13].
- Increase of the speech energy in the frequency bands with high noise energy was observed [14].
- Deletion of certain phonemes like /t/, /p/ and /f/ occurring at the end of a word, and aspiration after /m/ and /n/ increases. Certain vocabularies are more affected than others by the increase of the vocal effort [7].
- It is accompanied by larger facial movements but these do not aid as much as its sound changes [15].

The dependence of Lombard effect on gender and language were also reported [7]. Lombard effect speech of female speakers seem to be more intelligible than that of male speakers, and it was the opposite for normal speech. The acoustic-phonetic characteristics of Lombard effect speech in different languages were also studied [16][17]. Speech produced by wearing oxygen mask was also studied [18]. An increase in the vowel duration, fundamental frequency and total energy were reported along with the change in formant center frequency. It was suggested that these results may be attributed to a combination of the effective lengthening of the vocal tract and the restriction on freedom of jaw movement provided by the oxygen mask.

2.3 Intelligibility of Lombard effect speech

Acoustic-phonetic differences between Lombard effect speech and normal speech seem to have an effect on speech intelligibility. Intelligibility is commonly evaluated by presenting words masked with noise to listeners for identification [19]. Similar studies on word identification in noise have been performed [7][20][21][22][23]. The intelligibility of speech was found to increase due to increase in the vocal effort due to Lombard effect, and it was found to be nearly constant with increase in the intensity of feedback [21]. Beyond a certain level of the vocal effort, the intelligibility was found to decrease [24], especially when the speech becomes shouted speech. In the case of shouted speech, it was found that the increase in the vocal effort increases the energy and decreases the phonetic information [25]. It was also reported that the type of masking noise and the gender of the speakers are also crucial to the difference in intelligibility of speech produced in noise-free and in noisy conditions [14]. Multi-talker noise is found to degrade the intelligibility of English digit vocabulary more than white noise. Also female Lombard effect speech is more intelligible than male Lombard effect speech [7]. It seems that breathiness decreases the intelligibility of speech and when producing speech in a noisy background, female speakers tend to decrease the breathiness in their productions more than male speakers.

Intelligibility of speech was also associated with speaking rate [26]. In [27] the effect of four levels of aircraft noise (about 100 dB, 106 dB, 114 dB, 122 dB) on speaking rate in a passage read by 48 male American English speakers was reported. It was found that the higher the noise level was, the slower was the speaking rate. The mean speaking rate dropped successively from 183.2 to 165.4 (words per minute). Study on intelligibility of foreign accented speech produced in noise was also reported [28]. The effects of cafeteria noise on the perception of English sentences produced by groups of native speakers of Mandarin and of English were examined. The findings suggested that the cafeteria noise did degrade the intelligibility of the sentences overall, but the sentences spoken by Mandarin speakers were less intelligible than those of the English speakers in both noise-free and noisy conditions.

2.4 Psychological effects due to Lombard effect

Lombard effect results in psychological effect on several people under the influence of external feedback under various conditions. It was reported that both vocal rate and intensity were affected by the rooms in which the recording was made [29]. Phrases were read slowly in large rooms than in small ones, and among the large rooms, the rate was slower in live rooms than in dead rooms due to effects of the delayed feedback and reverberation. It was observed that a speaker speaks with high intensity in open air conditions than in close rooms. Lombard effect in case of choral singers is also studied [30]. Choral singers experience reduced feedback due to the sound of other singers upon their own voice. Thus the people in choruses sing at a louder level. Trained soloists can control this effect but it has been suggested that after a concert they might speaker more louder in noisy surrounding as in after-concert parties. Lombard effect on people playing instruments such as guitar is also studied [31].

Lombard effect is also studied to affect the vocalizations of animals. Beluga whales in the St. Lawrence River estuary adjust their whale song so it can be heard against shipping noise [32]. Great tits in Leiden sing with a higher frequency than do those in quieter area to overcome the masking effect of the low frequency background noise pollution of cities [33]. Other animals those vocalizations were effected due to Lombard effect are Budgerigars [34], Cats [35], Chickens [36], Common marmosets [37], Cottontop tamarins [38], Japanese quail [39], Nightingales [40], Rhesus Macaques [41], Squirrel monkey [42] and Zebra finches [43]

2.5 Lombard effect on speech systems

Degradation in the performance of speech recognition system and speaker recognition system due to Lombard effect are reported in [44][45]. The speaker recognition system performance with Lombard effect was 48% with mismatched training and 99% with matched training. The performance of the system was good when both the training and testing data was of the same condition. The performance degradation caused by the Lombard effect on speech recognition systems was seen to be more than that caused by noise [46]. The degradation was caused due to significant change in the acoustic features due to Lombard effect. Several compensation methods have been proposed to improve the robustness of speech systems. In [47], the Mel-frequency cepstral coefficient (MFCC) features for Lombard effect speech were compensated by modifying the mel scale to improve the performance of a verification system. In [48], the performance of perceptual speaker recognition using Lombard effect speech was reported. It was concluded that speaker recognition was better in the case of Lombard effect speech compared to normal speech.

Isolated word recognition experiments in a car environment was performed at three different engine speeds of 0, 90, 130 kph. An improvement in the performance was obtained with a combination of speech enhancement and spectral slope compensation [49]. In [50], robust front-end filter banks were used to improve the performance of recognition of the Lombard effect speech. A linear transformation of the linear prediction (LP) cepstral features was suggested with applications to dynamic time warping (DTW) based speech recognition in [51]. The variations due to Lombard effect were estimated and compensated using multiple linear transformations in [52]. A morphological constrained feature enhancement with adaptive cepstral compensation was used for speech recognition in noisy conditions [53]. In [54], time derivatives of cepstral coefficients have been used to improve noisy Lombard speech recognition. A 2-stage recognition system is proposed in [55], with one stage as a style classifier, and the other stage as a recognizer. The recognition uses perceptual linear prediction (PLP) features for normal speech and repartitioned frequency cepstral coefficients (RFCC) - linear prediction coefficient (LPC) features in case of Lombard effect speech. In [56], a method was proposed for speech recognition by integrating audio and visual information by training using hidden Markov model (HMM) and using Viterbi algorithm for decoding. The recognizer was tested using Lombard effect speech. Visual information also seem to detect the Lombard effect.

2.6 Importance of analysis of Lombard effect

Analyzing Lombard effect is helpful in several ways:

- Since it is not always possible to have a silent environment, speech is spoken under noisy conditions resulting in Lombard effect. Lombard effect speech degrades the performance of speech systems and therefore it is necessary to analyze the characteristics of Lombard effect to develop a compensation for better performance of speech systems.

- Since Lombard effect is more intelligible than normal speech in the presence of noise, adapting the characteristics of Lombard effect to normal speech will increase the intelligibility of normal speech. This has applications in public places which are noisy like a railway station where the announcements are make by a speaker who is generally in a silent environment and the listeners are in noisy environment.

- Analysis of Lombard effect can also be used in several forensic cases like estimating the environment condition under which a given speech is spoken by a person, etc.
- Since Lombard effect changes the characteristics of speech and is speaker dependent, it can be used to detect imposters in speaker verification systems as proposed in this thesis.

Chapter 3

Features for Analysis of Lombard Effect Speech

Speech is caused due to the time varying excitation of the time varying vocal tract system. These excitations initiate the production of speech which is manipulated by the vocal tract system. These excitation source features are obtained by removing the effects of the vocal-tract system on the speech signal. This time varying excitation changes by a greater extent under the influence of an external feedback. In this chapter, the excitation features used in the analysis of the Lombard effect speech are discussed along with another feature called normalized energy. Three excitation source features are considered, namely, instantaneous F_0 (pitch), strength of excitation at the epochs and a measure of loudness.

3.1 Fundamental frequency

The fundamental frequency (F_0) of speech is extracted by a recently proposed method [57], [58] using the zero-frequency filter, which was first reported in [59]. During the production of voiced speech, the excitation to the vocal-tract system can be approximated by a sequence of impulses of varying strengths. These impulse-like excitations result in discontinuity which is spread uniformly across the frequency range including zero-frequency. Filtering the speech signal using a zero-frequency resonator emphasizes the

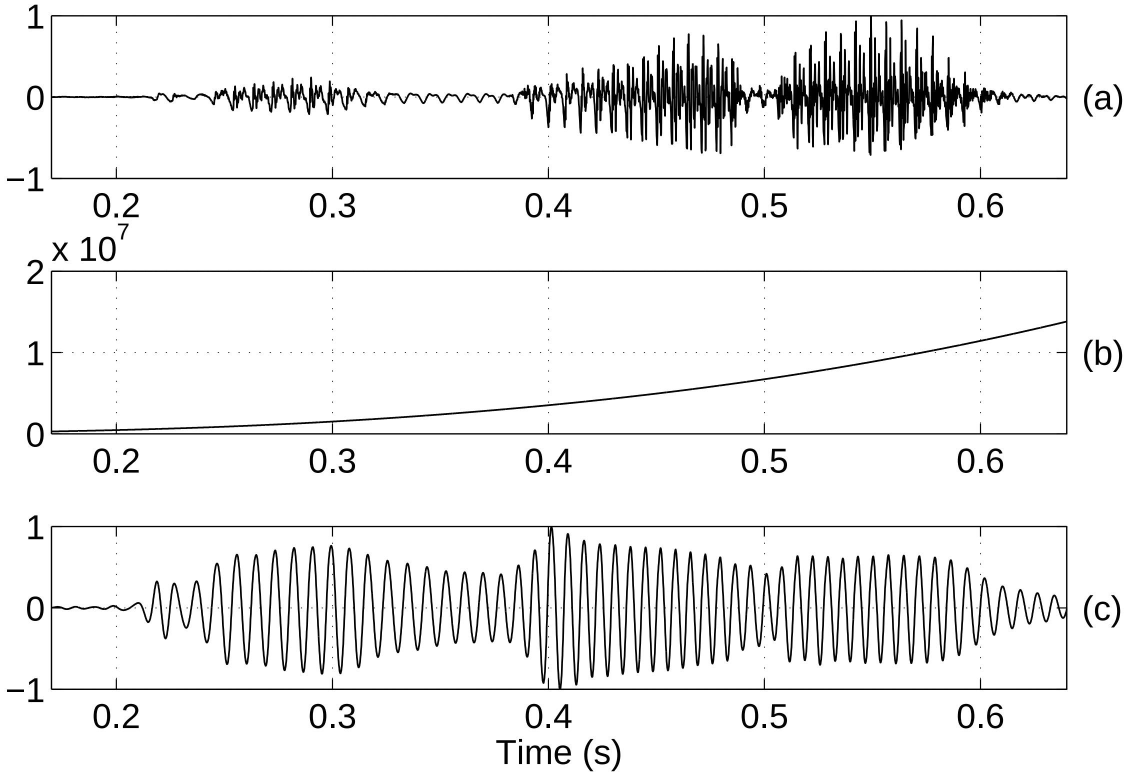

Fig. 3.1: Illustration of deriving filtered signal from speech signal. (a) A segment of speech signal taken from continuous speech. (b) output of cascade of two 0 Hz resonators. (c) Filtered signal obtained after mean subtraction.

characteristics of excitation. The output of the zero-frequency resonator is not affected by the characteristics of the vocal-tract system since it has resonances at much higher frequencies.

A zero-frequency resonator is an all-pole system with two poles on the positive real axis in the z-plane. We use a cascade of two ideal zero-frequency resonators to characterize the discontinuities due to impulse-like excitation in voiced speech. A cascade of two zero-frequency resonators provides sharper cut-off to reduce the effect of resonances of the vocal-tract system. Filtering a speech signal twice through a zero-frequency resonator results in an output that grows/decays as a polynomial function of time. Fig. 3.1(b) shows the output of filtering process for a segment of speech signal shown in Fig. 3.1(a). The characteristics of discontinuities can be highlighted by subtracting the local mean computed over a small window. A window size of about one to two times the average pitch period is adequate for local mean subtraction. The resulting mean subtracted signal is shown in Fig. 3.1(c) for the filtered output shown in Fig. 3.1(b). This mean subtracted signal is termed as the *zero-frequency filter signal* or merely the *filtered signal*. The following steps are involved in processing the speech signal to derive the filtered signal:

1. The speech signal $s[n]$ is differenced to remove any slowly varying component introduced by the recording device.

$$x[n] = s[n] - s[n-1] \tag{3.1}$$

2. Pass the differenced speech signal $x[n]$ through a cascade of two ideal zero-frequency (digital) resonators. That is

$$y_0[n] = -\sum_{k=1}^{4} a_k y_0[n-k] + x[n], \tag{3.2}$$

   where $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$. The resulting signal $y_0[n]$ grows approximately as a polynomial function of time.

3. The average pitch period is computed using the autocorrelation function of 30 ms segments of $x[n]$.

4. Remove the trend in $y_0[n]$ by subtracting the local mean computed over the average pitch period, at each sample. The resulting signal

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m] \tag{3.3}$$

is the zero-frequency filtered signal. Here $2N+1$ corresponds to the number of samples in the window used for mean subtraction. The choice of the window size is not critical as long as it is in the range of one to two pitch periods.

The filtered signal clearly shows sharper zero crossings around the epoch locations. The sharper zero crossings can either be positive-to-negative zero crossings (PNZC) or negative-to-positive zero crossings (NPZC) depending on the polarity of the signal (typically introduced by recording devices). In Fig. 3.1(c), the NPZCs are sharper than the

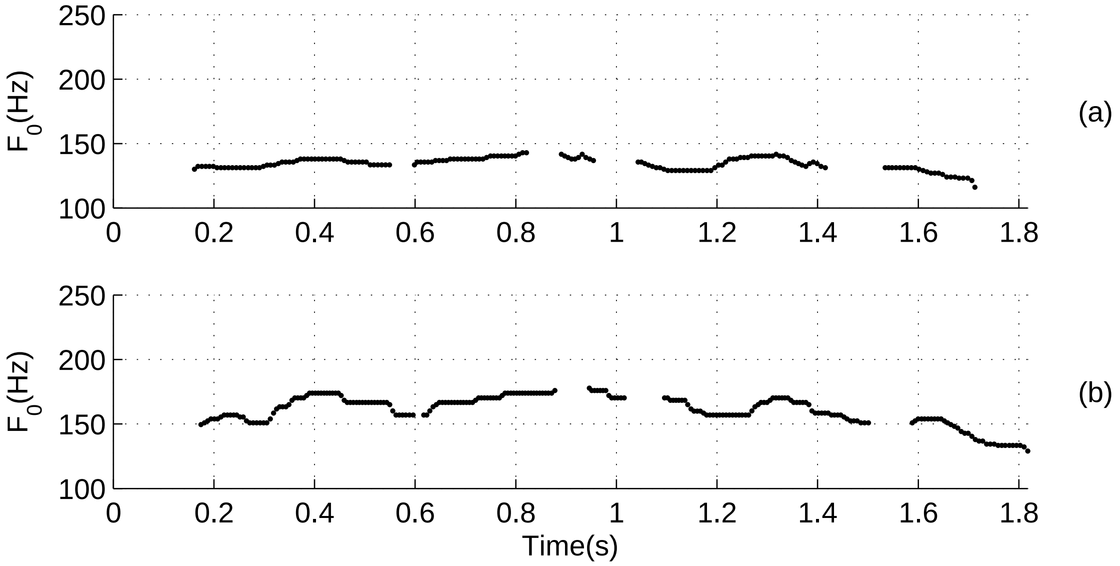

Fig. 3.2: F_0 contours for (a) normal speech, and (b) Lombard effect speech, for an utterance of the sentence, "Regular attendance is seldom required".

PNZCs, and hence indicate the epoch locations. The polarity of the sharper zero crossings can be automatically determined by comparing the slopes of the filtered signal around the PNZCs and the NPZCs over the entire duration of the utterance.

The interval between the two adjacent epochs is the pitch period. The reciprocal of the pitch period gives the fundamental frequency (F_0). Fig. 3.2 shows the F_0 contour for normal speech and the Lombard effect speech. We can see an increase in F_0 for the Lombard effect speech compared to the normal speech. The average fundamental frequency for these normal and Lombard effect speech utterances are 135 Hz and 169 Hz, respectively.

3.2 Strength of excitation

Strength of excitation (α) is measured as the slope at the positive zero-crossings at epoch locations in the zero-frequency filtered signal. It gives an idea of the amplitude of the equivalent impulse-like excitation [60]. It was also shown that the strength of excitation is proportional to the actual strength of excitation observed from EGG signal. But the strength at an epoch may not give an indication of the sharpness of the impulse, as

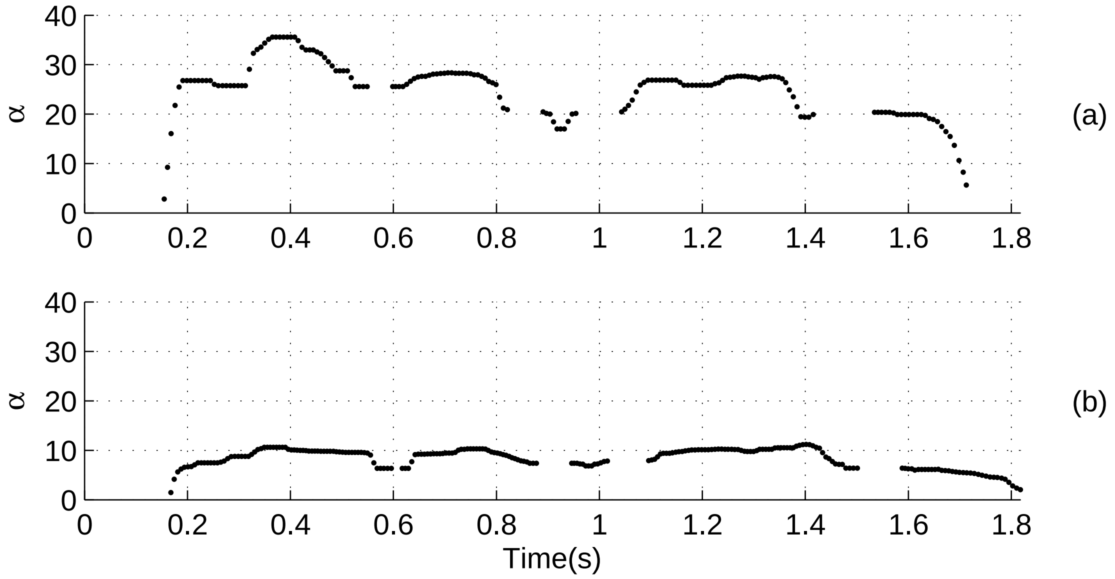

Fig. 3.3: α contours for (a) normal speech, and (b) Lombard effect speech, for an utterance of the sentence, “Regular attendance is seldom required”.

the sharpness of the impulse depends on the relative amplitudes of the excitation signal samples around the impulse. Fig. 3.3 shows the strength of excitation contour for normal speech and Lombard effect speech. We can see that the strength of excitation, as measured by the slope of the zero-frequency filtered signal at each epoch, was found to decrease for Lombard effect speech compared to normal speech. Though the vocal tract system is known to release large amount of acoustic energy in the case of Lombard effect speech compared to normal speech, the strength of excitation is found to decrease. Thus the strength of excitation need not depend on the acoustic energy released by the vocal tract.

3.3 Measures of Loudness

Here we consider two measures of loudness: loudness due to glottal excitation and perceived loudness

3.3.1 Loudness due to glottal excitation

A measure (η) of loudness is derived from the Hilbert envelope of the linear prediction (LP) residual as proposed in [61]. It indicates the sharpness of the impulse, which contributes to the perception of loudness. The LP residual $e[n]$ is obtained using a 10^{th} order LP analysis on each 30 ms frame of speech signal with a frame shift of 10 ms. The Hilbert Envelope $r[n]$ of the LP residual is given by

$$r[n] = \sqrt{e^2[n] + e_H^2[n]}, \tag{3.4}$$

where $e_H[n]$ denotes the Hilbert transform of $e[n]$. The Hilbert transform $e_H[n]$ is given by

$$e_H[n] = \text{IFT}(E_H(\omega)), \tag{3.5}$$

where IFT denotes the inverse Fourier transform, and $E_H(\omega)$ is given by [62]

$$E_H(\omega) = \begin{cases} +jE(\omega), & \omega \leq 0 \\ -jE(\omega), & \omega > 0. \end{cases} \tag{3.6}$$

Here $E(\omega)$ denotes the Fourier transform of the signal $e[n]$. The η is measured as the ratio of the standard deviation (σ) and the mean (μ) of the Hilbert Envelope in the 3 ms region around each epoch. Fig. 3.4(c) shows the Hilbert envelope of the LP residual (Fig. 3.4(b)) of the speech signal shown in Fig. 3.4(a) whose η contour is illustrated in Fig. 3.4(d). The peaks in the Hilbert envelope around the epoch locations are sharper in the case of loud speech. Figs. 3.5 (a), (b) and (c) show segments of the Hilbert Envelope around the peaks at the instants of significant excitation (epochs), which are superimposed for the cases of normal, Lombard effect and loud speech, respectively. Peaks are sharper in case of loud speech compared to Lombard effect speech. Fig. 3.6 (b) shows the η contour of the speech signal shown in the Fig. 3.6 (a). The regions corresponding to the high values of η are the loud regions in the speech signal. It was also seen that those regions correspond to vowel regions. The measure (η) of loudness represents the loudness due to glottal excitation (actual loudness) and not the perceived loudness. Loudness is perceived

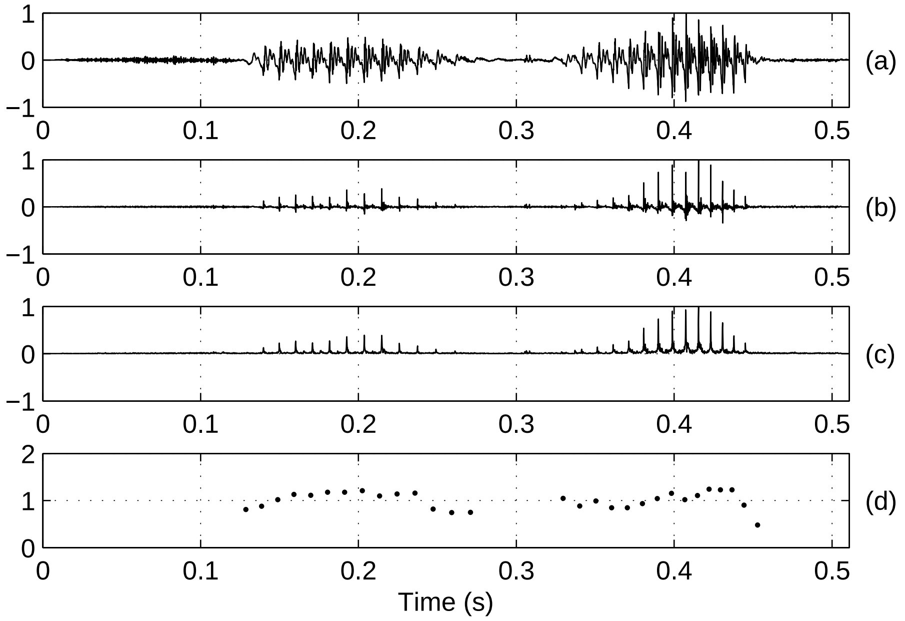

Fig. 3.4: (a) Segment of speech signal, (b) 10^{th} order LP residual, (c) Hilbert envelope of the LP residual, and (d) η contour extracted from the Hilbert envelope.

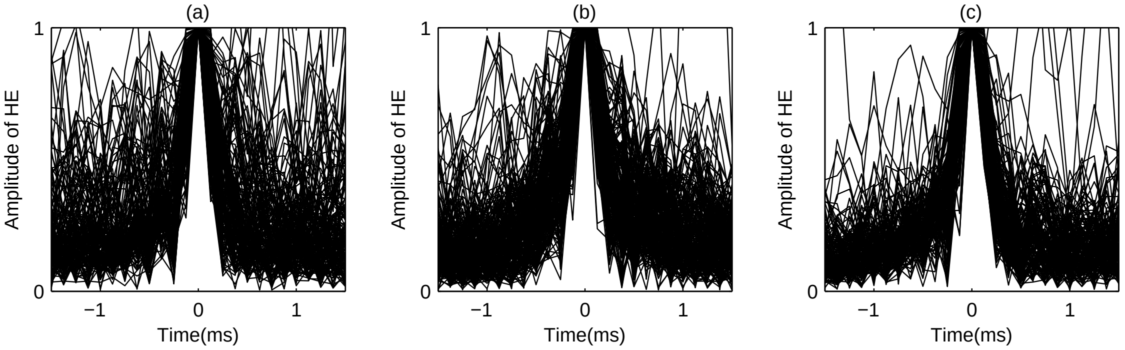

Fig. 3.5: Superimposed segments of Hilbert envelope of the LP residual around the epochs for (a) normal speech, (b) Lombard effect speech, and (c) loud speech.

higher in the case of Lombard effect speech, but this is not reflected well in the loudness due to glottal excitation. This measure of loudness is different for different sound units in a speech signal.

3.3.2 Perceived loudness

Since the loudness is perceived higher in case of Lombard effect speech compared to normal speech, we propose a measure of perceived loudness (β) to capture this effect.

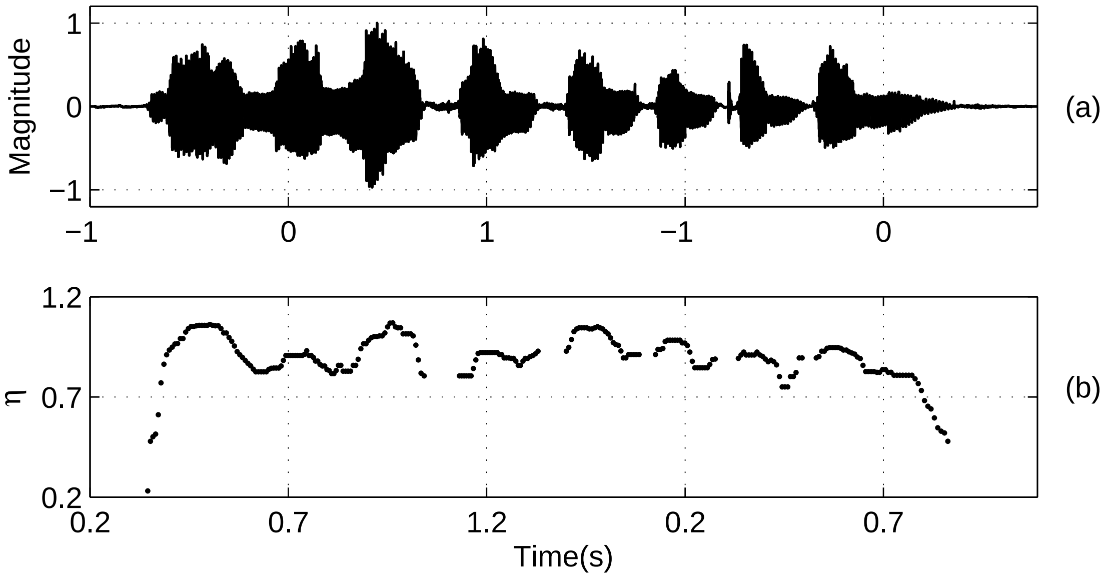

Fig. 3.6: (a) Speech signal, (b) η contour of the speech signal.

Perceptual loudness depends not only on the sharpness of the peaks in the Hilbert Envelope around the epoch, but also on the fundamental frequency (F_0). Since Lombard effect increase the fundamental frequency, Lombard effect speech is perceived to be louder. The β measure is a product of intrinsic loudness (η) and F_0.

$$\beta = \eta \times F_0. \tag{3.7}$$

Fig. 3.7 shows the β contour of normal speech and for Lombard effect speech. An increase in the perceived loudness is seen for the Lombard effect speech. The increase in the perceived loudness seems to be different for different sound units. Stressed vowels show a greater increase in the loudness. The increase in the loudness for Lombard effect speech is also different for different speakers. Thus it might be a speaker-dependent property. The increase in loudness is found to be significant in the vowel regions.

3.4 Normalized energy

Given a normalized speech signal, framewise energy is calculated with a framesize of 10 ms and a frameshift of 3 ms. Selection of framesize and frameshift is not critical. Energy

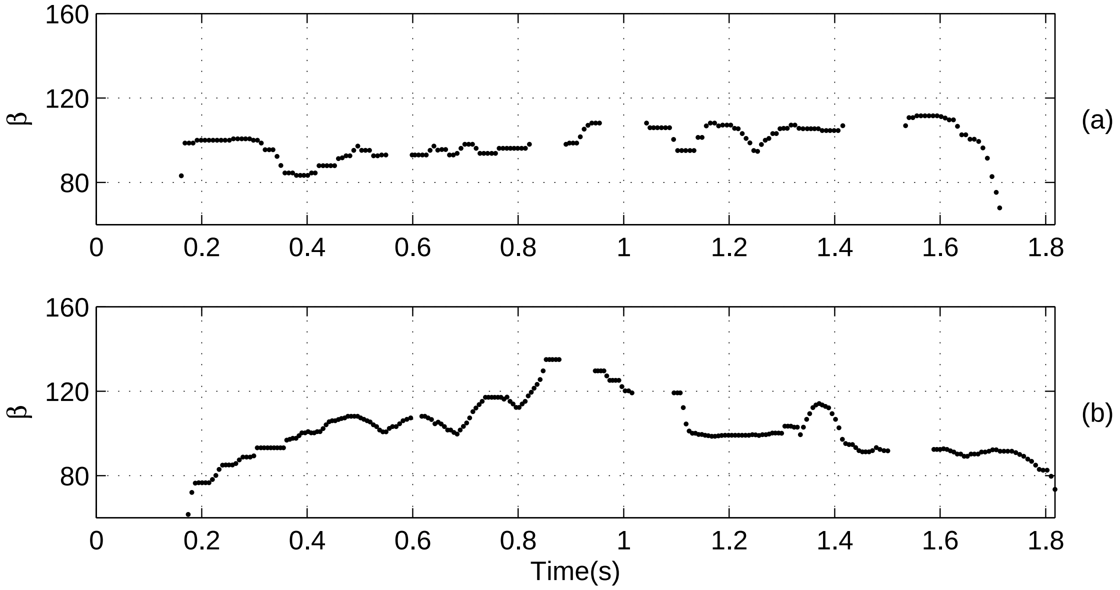

Fig. 3.7: β contours for (a) normal speech, and (b) Lombard effect speech, for an utterance of the sentence, “Regular attendance is seldom required”.

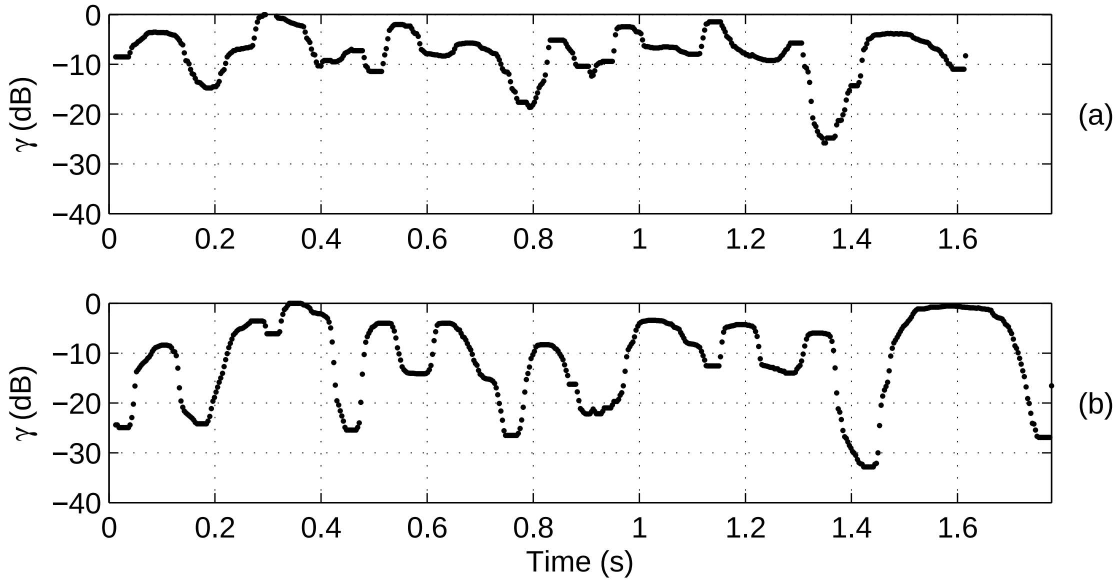

Fig. 3.8: γ contours for (a) normal speech, and (b) Lombard effect speech, for the utterance of the sentence, “Regular attendance is seldom required”.

is then normalized so that the maximum energy maps to 0 dB, to obtain the normalized energy (γ). Figs. 3.8(a) and 3.8(b) show the γ contours of normal speech and the Lombard effect speech, respectively. Lombard effect results in the articulation variability in speech which is reflected in the variations of the normalized energy. It can be seen that there are large variations in normalized energy in the case of Lombard effect speech compared to normal speech which is due to the decrease in the consonant to vowel energy ratio due to Lombard effect which was reported in [10]. There is a significant increase in energy in vowel regions compared to other regions. The mean value is not considered here as it is normalized and thus we only consider the variations for discriminating Lombard effect speech from normal speech.

3.5 Summary

In this chapter, we have described three excitation source features which are fundamental frequency (F_0), strength of excitation (α) and a perceived loudness measure (β) along with another feature which is the normalized frequency (γ). We have described the extraction of each of the features. F_0, α are extracted using the zero-frequency filter, β is extracted from the peaks in the Hilbert envelope and γ is directly extracted from the speech signal. We have also shown the change in the features due to Lombard effect.

Chapter 4

Studies on Different Cases of Lombard Effect

In this chapter we study the behavior of speech produced by affecting the self feedback in several ways. Self feedback of our speech determines the characteristics of our speech. These characteristics are affected by hampering our self feedback. The Lombard effect due to external feedback as measured by the three features and their distributions can vary over a wide range depending on the type and level of the external feedback. It was reported that the Lombard effect speech depends on the type and intensity of feedback [63]. Here we study the extent of Lombard effect for various cases as follows:

1. Different types of feedback.
2. Different intensity levels of feedback.
3. External feedback only through a single ear (SEF).
4. Ears closed (no feedback case (NF)).
5. Lombard effect when a speaker is asked to speak normally pretending he is not under the influence of external feedback (no-effect case (NE)).
6. Lombard effect speech vs loud speech

Speech of a person can be modified by three methods:

1. Voluntary modification
2. Hampering feedback
3. Emotion

We do not study speech due to different emotions. Loud speech and no-effect case stated above are voluntary. The other cases deal with the modification of speech due to hampering of feedback. Before we study each of the above cases in terms of the features of excitation, we describe the method of data collection. Section 4.1 describes the procedure for data collection of Lombard effect speech. Section 4.2 describes the Lombard effect for different intensities of external feedback. Section 4.3 describes the Lombard effect for different types of external feedback. In Section 4.4, Lombard effect due to external feedback through single ear is described. In Section 4.5, speech produced without a self-feedback is described. Section 4.6 describes the Lombard effect when a speaker is asked to control his speech in presence of an external feedback. In Section 4.7, difference between Lombard effect speech and loud speech are described. Summary of this chapter is presented in Section 4.8. Change in normalized energy is not considered here as its distributions do not show significant evidence.

4.1 Data Collection

Speech was collected from 18 male speakers in the age group of 21-24 years. Recording was done in a single session, within a time interval of a few minutes for each speaker. Five sentences were chosen from the TIMIT database [64], and the speakers were asked to speak them three times each, under normal conditions. The noise signals are presented to the speaker through earphones, which do not allow any external sound to pass through them, and also do not allow the presented noise signals to leak out. The Lombard effect speech is recorded using a close speaking microphone. External feedback of 3 noise types and 3 intensity levels are considered: (a) pink noise (PN) at 70, 60, 50 dB, (b) babble noise

(BN) at 65, 55, 45 dB, and (c) factory noise (FN) at 65, 55, 45 dB. The noises were taken from the NOISEX-92 database [65]. The noises were amplified to obtain the required intensity levels, and the speaker was asked to speak under the following conditions:

1. 3 noise types at 3 intensity levels played through the earphones.
2. External feedback (PN-70 dB) through a single ear.
3. Ears closed.
4. Speaker was instructed to speak normally under the influence of external feedback (PN-70 dB).
5. Loud speech

The speaker was asked to repeat each of the five sentences three times in the case of PN-70, BN-65, FN-65, and in case of their loud speech, as these cases are emphasized in our work. For the remaining cases each of the five sentences are spoken only once. Speakers were not informed about the Lombard effect phenomenon, to avoid any bias in their anticipation of speech. The durations of the utterances ranged from 1 second to 4 seconds. The speech signals were sampled at 8 kHz.

Now we describe the procedure used to amplify the noise to the required level. Since the magnitude of a signal is estimated with respect to some reference noise, we use the noise in the environment under which data recording is done, as the reference noise. We consider the speech produced under the reference noise as the normal speech. The noise signal which are to be amplified to the required level are taken from the the NOISEX database which is specified above. Let E_0 be the energy of the reference noise, E_1 be the initial energy of the required noise. The intensity (I) of this required noise is given by

$$I = 10\log_{10}\left(\frac{E_1}{E_0}\right). \tag{4.1}$$

If the desired intensity of the noise is I', then the noise need to be amplified by k times. If E_1' denotes the energy of the amplified noise, then

$$E_1' = k^2 E_1. \tag{4.2}$$

Thus we know the value of I, I', E_1 and E_0 and we need to find k. By solving the equations we get

$$k = 10^{\frac{1}{2}\left(\frac{I'}{I}-1\right)\left(\log_{10}\frac{E_1}{E_0}\right)}. \tag{4.3}$$

Thus the noise signals are amplified by the factor k. Precautions need to be taken such that the signal doesn't get clipped. Thus the maximum amplification factor possible is the level of the noise above which it gets clipped.

4.2 Lombard effect for different intensities of external feedback

In general the F_0 and β increase, and the α decreases, with increase in the intensity of the feedback. The extent of Lombard effect is seen to decrease with decrease in the intensity of the external feedback. Fig. 4.1 shows the distribution of F_0, α and β for Lombard effect speech produced under an external feedback of pink noise at intensities 70, 60 and 50 dB for 3 different speakers. Figure shows the speaker specific nature of Lombard effect. Another observation is that the Lombard effect speech at 60 dB is more closer to the Lombard effect speech at 50 dB than the Lombard effect speech at 70 dB for the same type of noise. This can also be seen from the mean standard deviation values of the features as shown in Table 4.1. Thus Lombard effect decreases rapidly with decrease in the intensity of the feedback.

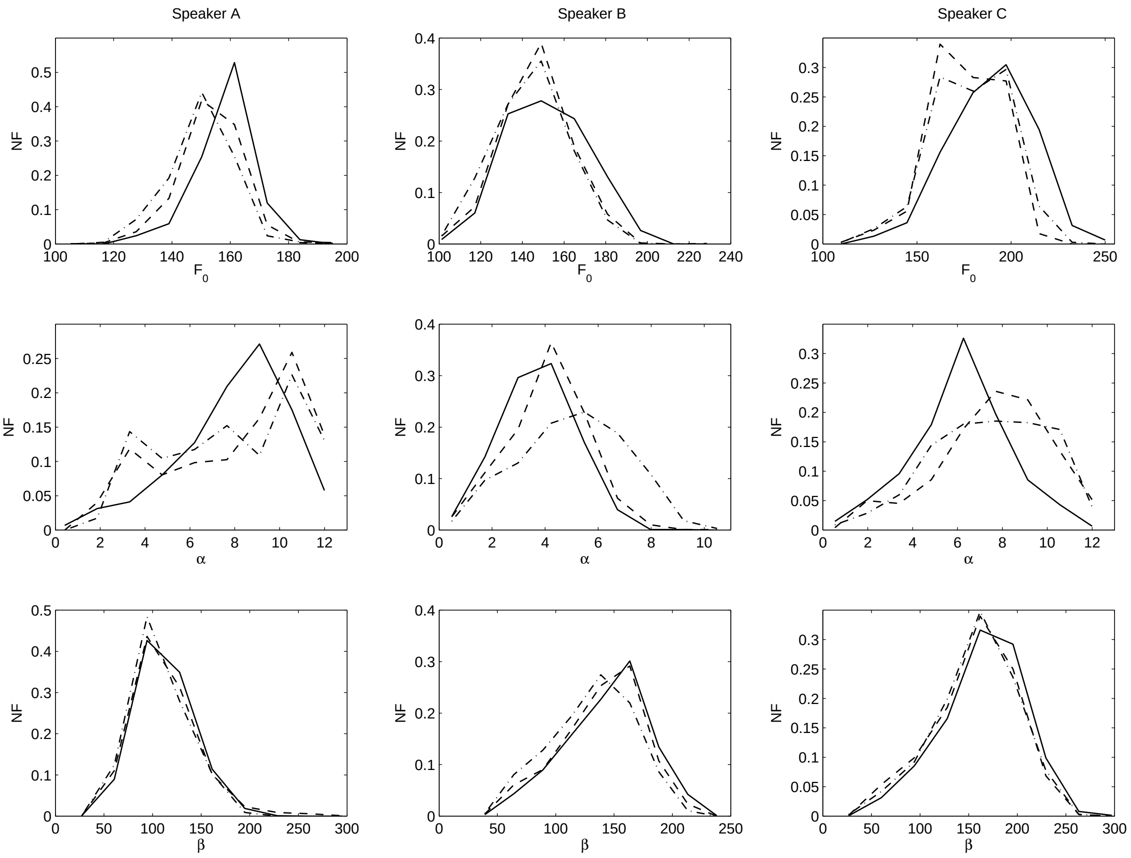

Fig. 4.1: Distribution of F_0, α and β for Lombard effect speech with pink noise as external feedback of intensities 70 dB (solid lines, 60 dB (dashed lines) and 50 dB (dash-dotted lines) for 3 speakers.

Table 4.1: Mean (μ) and standard deviation (σ) of the excitation features and standard deviation for Lombard effect speech under different external feedbacks at different intensities

Feature		PN-70	PN-60	PN-50	BN-65	BN-55	BN-45	FN-65	FN-55	FN-45
F_0	μ	165.52	160.79	156.45	164.94	159.78	155.83	168.44	160.71	156.88
	σ	20.13	19.87	19.78	20.91	19.91	19.88	21.11	20.36	20.09
α	μ	8.41	9.37	11.42	8.34	10.66	12.48	8.03	9.77	12.99
	σ	3.02	3.78	4.34	3.27	4.28	4.86	3.13	3.79	4.94
F_0	μ	120.21	116.43	113.86	120.74	115.38	113.12	121.21	115.69	113.03
	σ	31.61	31.24	30.66	32.11	30.76	30.74	32.39	31.03	30.79
γ	σ	9.40	9.14	8.99	9.33	8.84	8.67	9.15	8.87	8.79

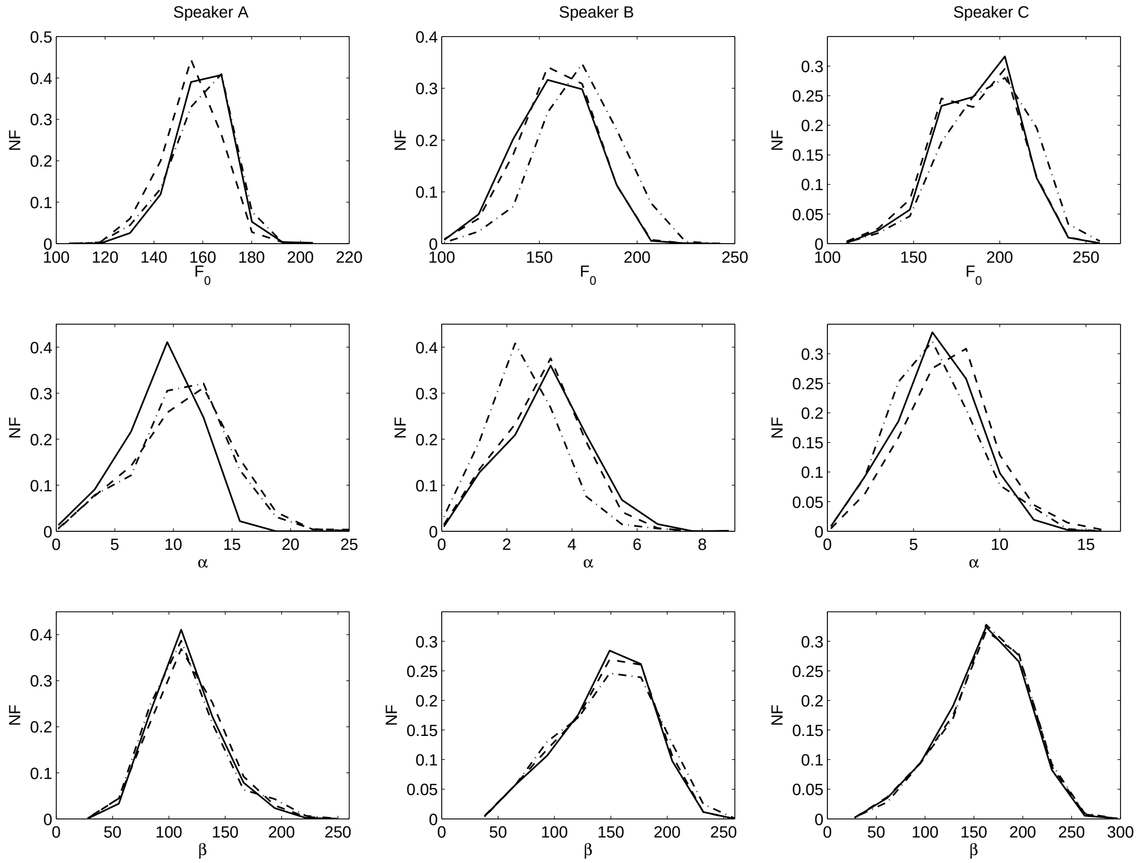

Fig. 4.2: Distribution of F_0, α and β for Lombard effect speech with external feedback as pink noise-70 dB (solid lines), babble noise-65 dB (dashed lines) and factory noise-65 dB (dash-dotted lines) for 3 speakers.

4.3 Lombard effect for different types of external feedback

The influence of self-feedback due to different types of external feedback is different. It was stated in [14] that the frequency distribution of the noise affects the Lombard effect and shape the acoustic changes observed in the speech signal. But this observation is not reflected perceptually. The frequency distribution of the noise affects the characteristics of the speech signal, although it is not felt perceptually. The changes in the features due to different types of feedback are different, and are dependent on the speaker. Fig. 4.2 shows the distribution of F_0, α and β for Lombard effect speech produced under an external feedback of pink noise at 70 dB, babble noise at 65 dB and factory noise at 65 dB. The distributions seen in the figure are close, but from the small deviations, we can

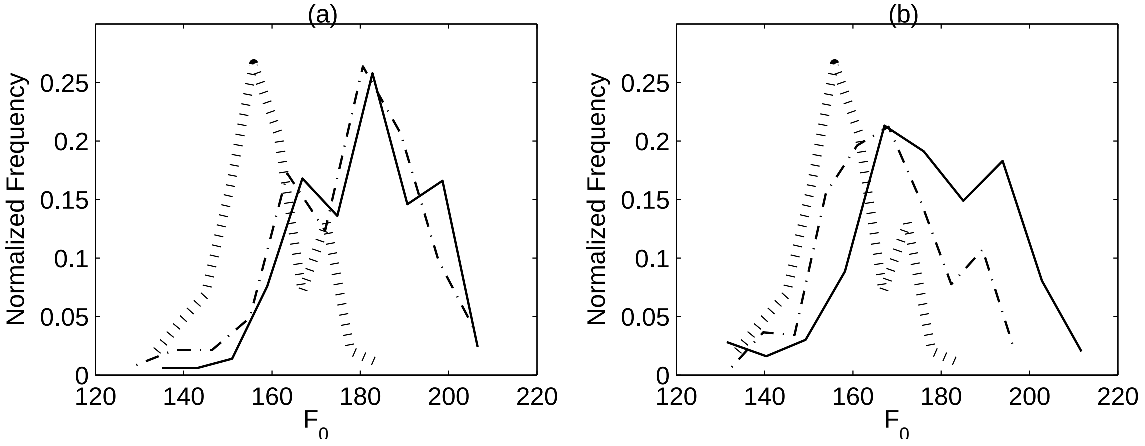

Fig. 4.3: Distribution of F_0 for normal speech (dotted lines), Lombard effect speech with low intensity feedback (dash-dotted lines), Lombard effect speech with high intensity feedback (solid lines) for 2 cases of feedback: (a) noise, and (b) normal speech.

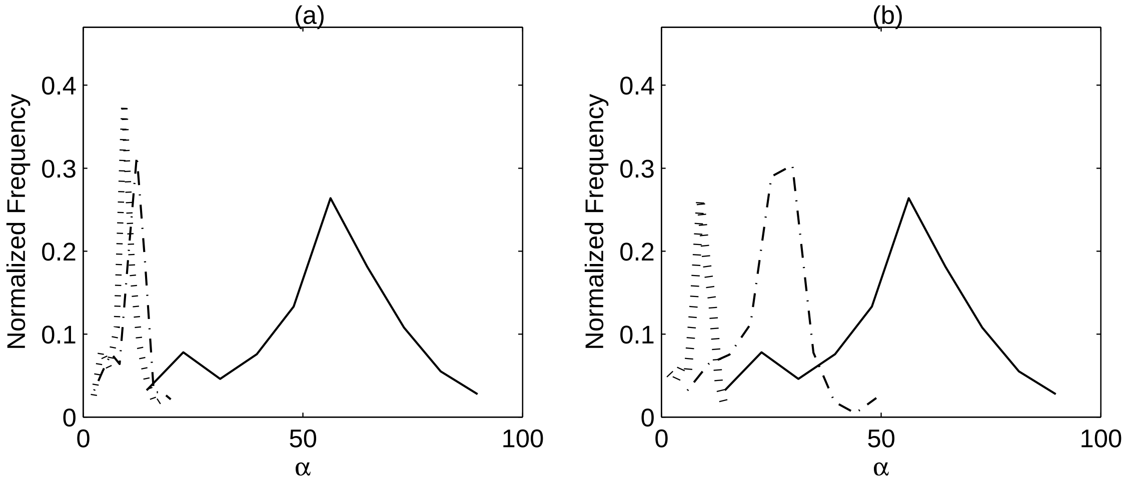

Fig. 4.4: Distribution of α for normal speech (dotted lines), Lombard effect speech with low intensity feedback (dash-dotted lines), Lombard effect speech with high intensity feedback (solid lines) for 2 cases of feedback: (a) Noise, and (b) Normal speech.

see that factory noise has affected the speech production more than pink noise and babble noise. This can also be seen from the mean and standard deviation values of the excitation features for the speech produced due to different types of noise in Table 4.1. It is difficult to differentiate the effect caused by pink noise and babble noise, though babble noise has shown to affect slightly more. Previous studies have reported that multi-speaker noise degrades the intelligibility more than the white Gaussian noise does for digit vocabulary [7].

In another experiment, speech of a person is recorded under the influence of (a) white noise, (b) normal speech spoken by another person [66]. The task of this experiment is to study Lombard effect in presence of speech of another speaker instead of noise. Figs. 4.3 and 4.4 show distributions of F_0 and α, respectively, for two types of feedback: (a) white noise and (b) normal speech, for 3 cases: (1) Speech under silent conditions. (2) Speech with low intensity of feedback. (3) Speech with high intensity of feedback. We find an increase in F_0 and a decrease in α with the increase in intensity of the external feedback signal. Another observation is that the distribution of the α (i.e., width of the spread) decreases with increase in the intensity level of the external feedback signal. The distribution of the α is also more for the case of normal speech as external feedback, when compared with the same intensity white noise as external feedback. This shows that Lombard effect under noisy conditions is more than that under the influence of another speakers voice.

4.4 Lombard effect due to external feedback through a single ear

Another case is the study of Lombard effect when the feedback is received by a person only through a single ear. Since binaural hearing is required for a person to speak normally, it is interesting to study the Lombard effect when the feedback is received by a person only through a single ear, and with normal hearing from the other ear. Fig. 4.5 shows the distribution of F_0, α and β for normal speech, speech with single ear feedback, Lombard effect speech with pink noise-70 dB for 3 speakers. Lombard effect is seen to reduce to a large extent by restricting the external feedback only through a single ear. Lombard effect due to single ear feedback was close to normal speech in case of speaker C compared to speakers A and B. The Lombard effect seems to be not significant in this case, as can be seen through the mean and standard deviation values of the features in comparison with the mean values for normal speech in Table 4.2.

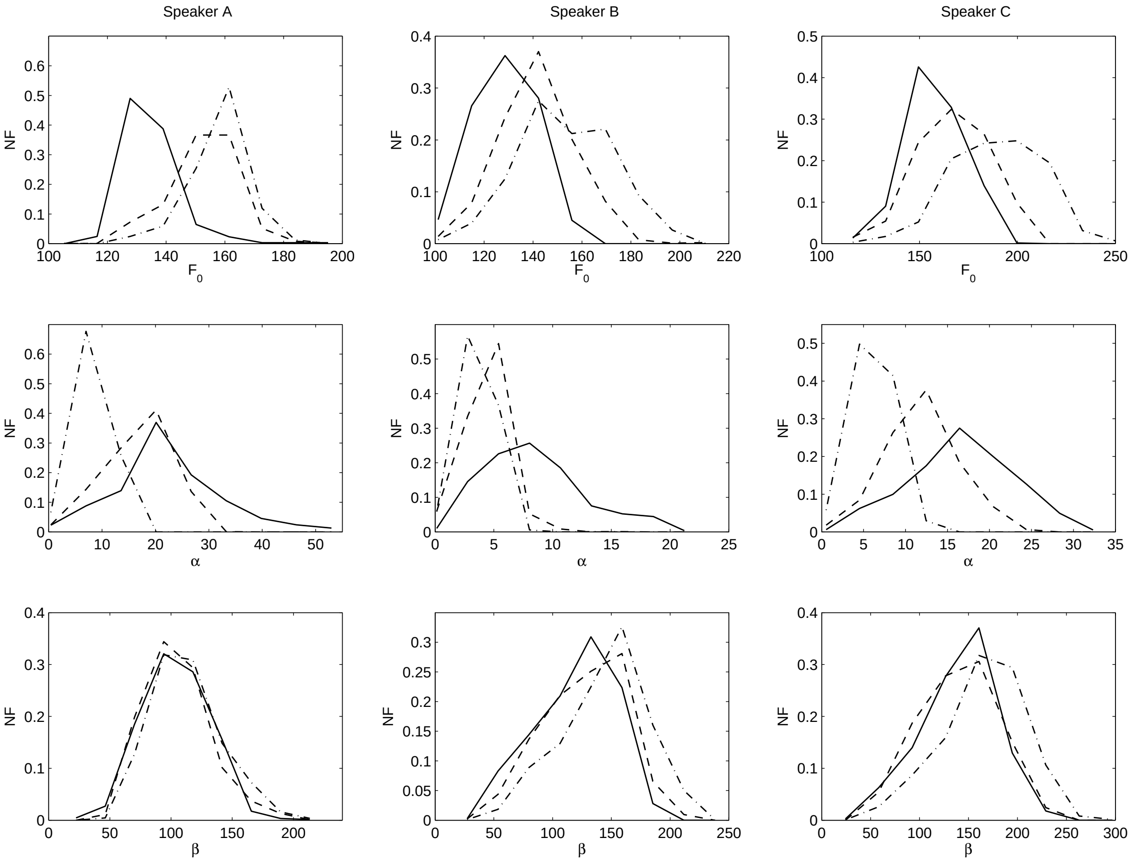

Fig. 4.5: Distribution of F_0, α and β for normal speech (solid lines), speech with single ear feedback with pink noise-70 dB (dashed lines) and Lombard effect speech with pink noise-70 dB (dash-dotted lines) for 3 speakers.

Table 4.2: Mean (μ) and standard deviation (σ) of the excitation features and standard deviation of normalized energy for various cases of Lombard effect

Feature		Normal	SEF	NF	NE
F_0	μ	143.41	149.01	151.52	148.10
	σ	17.34	18.29	20.63	20.27
α	μ	21.29	17.04	17.96	14.85
	σ	7.56	6.15	5.67	6.62
β	μ	105.23	106.36	107.97	107.99
	σ	28.69	29.05	29.37	30.09
γ	σ	8.09	8.45	8.25	8.55

4.5 No feedback case

In this case, earphones are fixed to the speaker which will mask the self feedback from reaching the speaker. No external feedback is presented to the speaker. By blocking the ears, the self feedback is not hampered as it is still perceived by the speaker through bone conduction. Thus the speech produced is not changed by a large factor. In this case the speaker doesn't have an idea of his normal speech as he can perceive his own speech only through the internal vibrations in his bones. In the process he/she adjusts his/her speech so as to perceive it as better as possible. It is found that depending on the feedback through bone conduction, few speakers even spoke softer than their normal speech. This is not a case of Lombard effect as sufficient self feedback is still present and no external feedback is presented to the speaker. Fig. 4.6 shows the distributions of the excitation features for normal speech, no-feedback speech and Lombard effect speech (under an external feedback of pink noise-70 dB) for 3 speakers. Speaker C has shown a considerable increase in loudness where as the change in features in case of speakers A and B is minimum. Speaker A has even shown a decrease in loudness. The change in the mean and standard deviation of the excitation features are not significant in this case compared to normal speech as seen in the Table 4.2.

We perceive our own speech by two conduction mechanisms which are air conduction and bone conduction. Speech through air conduction has high frequencies and that through bone conduction has low frequencies. Air conduction dominates bone conduction in the speech perception mechanism. Thus if the self feedback through air conduction is hampered our perception of speech is reduced there by forcing the speech production mechanism to modify the speech. Bone conduction has a minimum role in this case. If the air conduction is blocked by closing our ears, the perception of our self feedback is completely dependent on bone conduction. Signal-to-noise ratio (snr) value is high in case of bone conduction speech as it is not hampered by any external feedback.

Speech produced is guided by the environment we are in. By blocking our ears, the speech produced is no longer dependent on the environment. Thus, theoretically speech spoken by blocking the ears should be the actual normal speech of the person as it

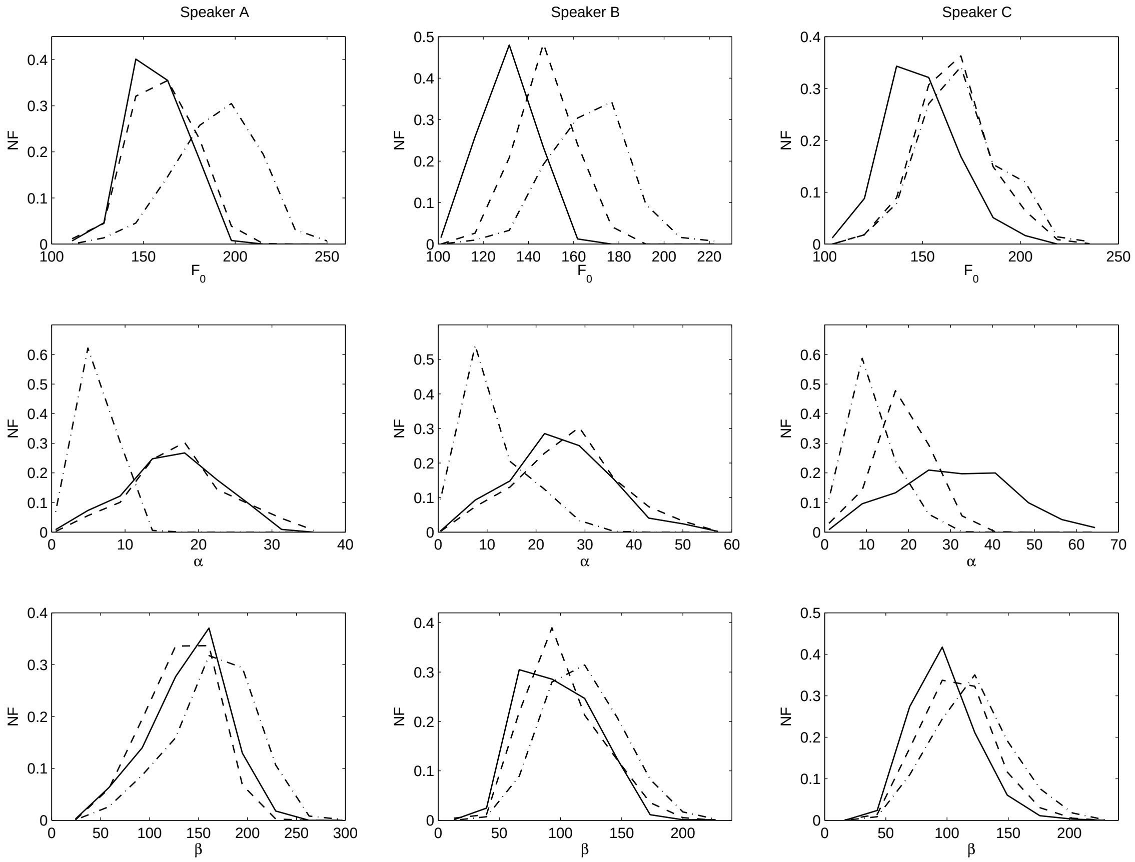

Fig. 4.6: Distribution of F_0, α and β for normal speech (solid lines), no-feedback speech (dashed lines) and Lombard effect speech with pink noise-70 dB (dash-dotted lines) for 3 speakers.

avoids any external feedback from hampering the self feedback through bone conduction. Speech produced in this case is completely speaker dependent. Experiments have shown that soft speech of a person is perceived much better by himself when he closes his ears than normal hearing. With ears closed, the perceptual quality of the self feedback is seen to decrease with increase in the intensity of his speech. The intelligibility of the perceived speech is seen to decrease with increase in vocal intensity, when the ears are closed. The reason why we close our ears in public places when we are speaking on a phone is not only to hear the other person clearly but also to protect our self feedback from getting hampered.

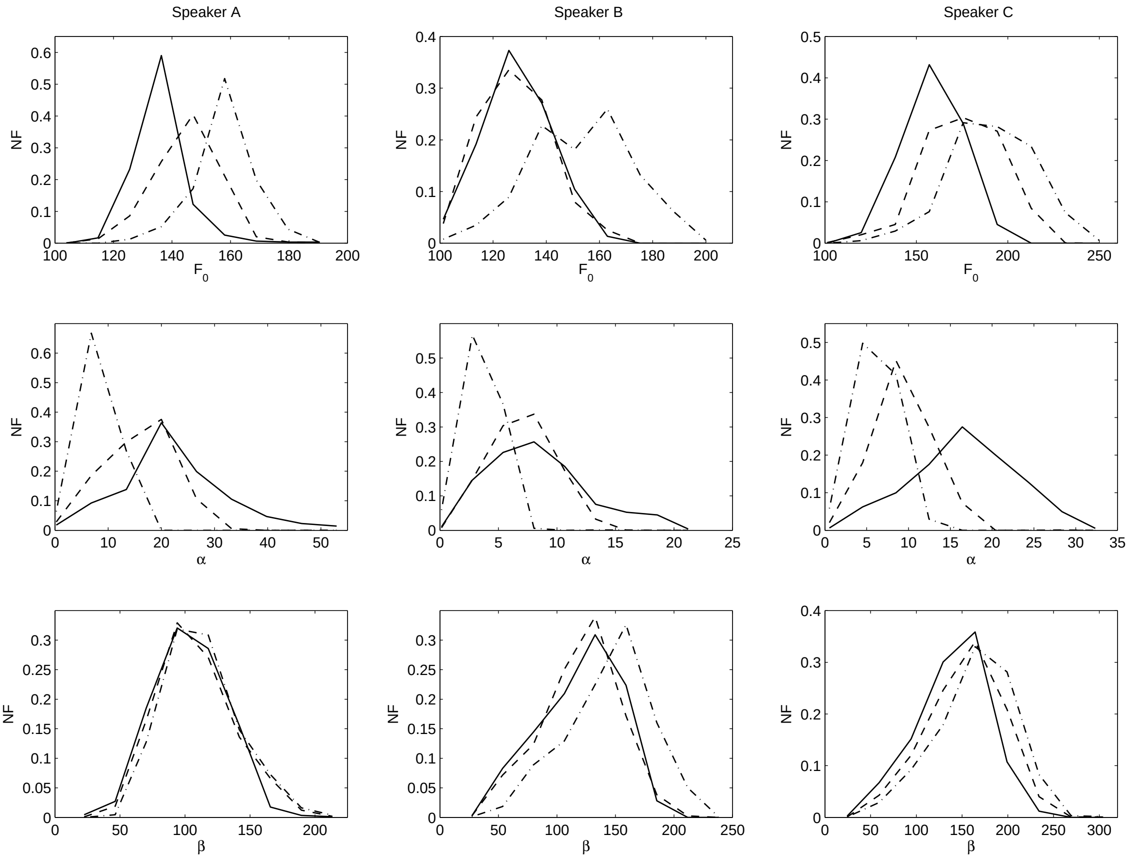

Fig. 4.7: Distribution of F_0, α and β for normal speech (solid lines), no-effect speech (dashed lines) and Lombard effect speech with pink noise-70 dB (dash-dotted lines) for 3 speakers.

4.6 Lombard effect for no effect case

In another experiment, the speaker is asked to maintain normal speech, even though he/she is under the influence of an external feedback. The speech produced in this case is voluntary. When a person tries to maintain his normal speech, when he hears his own voice after a certain delay (like an echo), an increase in duration was observed [67]. The results in Table 4.2 shows that the person cannot speak normally under the influence of external feedback. He can modify his speech and bring it closer to his speech under normal conditions, but he cannot completely attain his normal speech. The extent to which he can get close to his normal speech is seen to differ from speaker to speaker.

From Fig. 4.7, we can see that speakers A and C are able to get their speech in presence of noise close to their normal speech but in case of speaker B, he has spoken softer than his normal speech. Thus the speaker doesn't have an idea about his speech under normal

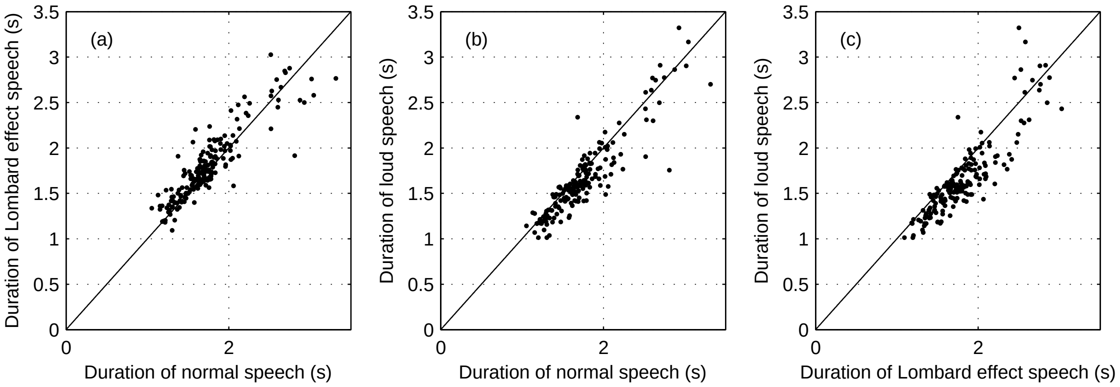

Fig. 4.8: Scatter plots of (a) durations of normal speech vs durations of Lombard effect speech, (b) durations of normal speech vs durations of loud speech, and (c) durations of Lombard effect speech vs durations of loud speech.

conditions and will only try to speak softer, which can even cause him to speak softer than his normal speech. It was reported in [68] that the changes due to Lombard effect can be controlled by instructing a person to speak as they would in silence. But this is contradicted in our study. For perfectly controlling our speech in presence of an external feedback, the speaker has to train himself extensively. With change in the environment, normal speech is affected. These characteristics might be helpful in forensic cases, or in some speech systems where a person might try to speak normally even under the influence of a feedback to cheat the system.

4.7 Lombard effect speech vs loud speech

So far we have studied the characteristics of the Lombard effect speech in comparison with normal speech. It is interesting to study how the Lombard effect speech differs from the loud speech, as the Lombard effect speech is also perceived to be loud. Loud speech is a self controlled phenomenon, whereas the Lombard effect speech depends on external feedback. There are several common features for both loud and Lombard effect speech. Both of them show an increase in F_0 and β, and decrease in α.

Loud speech is a controlled process where the person controls his vocal effort to produce speech as loud as he desires. The Lombard effect speech is also a loud speech, where the vocal effort is increased involuntarily. In the case of the Lombard effect speech a per-

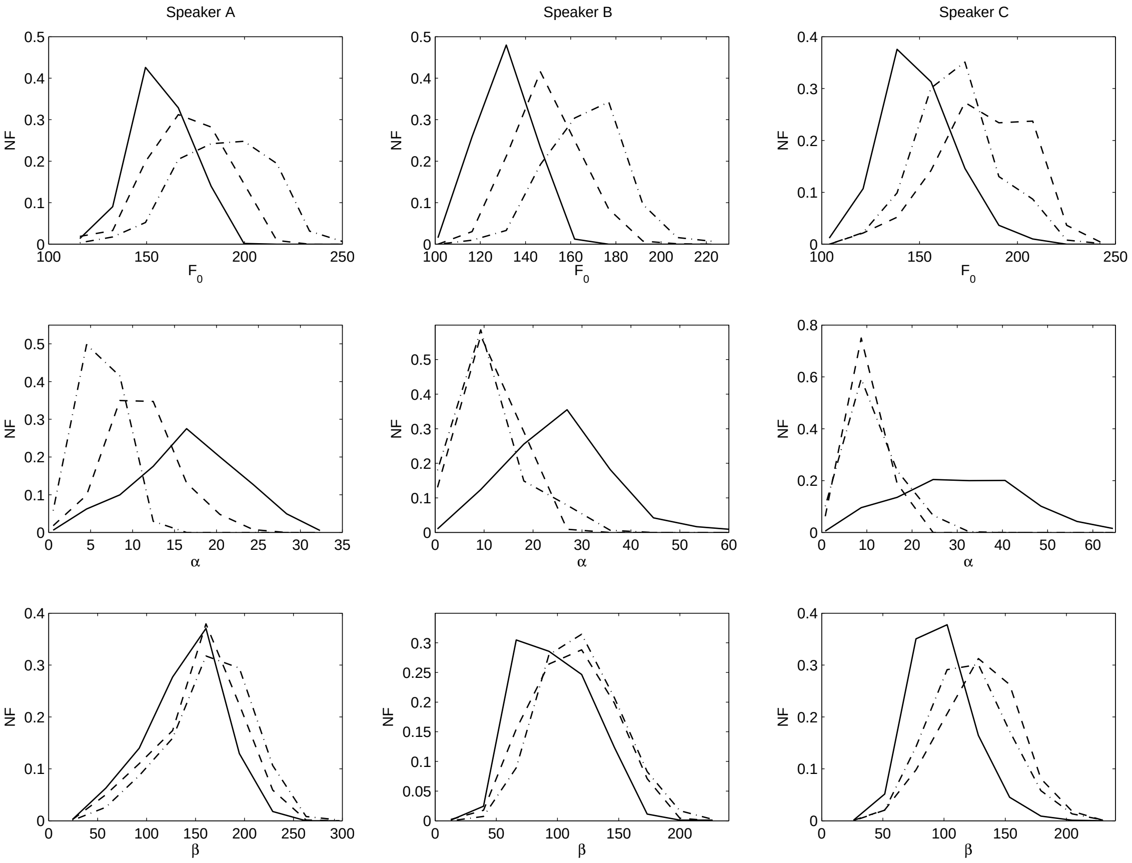

Fig. 4.9: Distribution of F_0, α and β for (a) normal speech (solid lines), (b) loud speech (dashed lines), (c) Lombard effect speech with pink noise-70 dB as feedback (dash-dotted lines) for 3 speakers.

son applies an involuntary stress on his speech. Normally the speaker is not aware of the extent of increase of their vocal effort. The increase in the vocal effort for loud speech was seen to differ from speaker to speaker. Loud speech is not as loud as the Lombard effect speech with a feedback of high intensity sound. Instead, loud speech was closer to Lombard effect speech when the feedback is of low intensity.

Loud speech was produced by two methods. Some speakers increased their speaking rate, as increasing the speaking rate seems to be a method to produce louder speech, thereby reducing the duration of utterance. Some others increased their vocal effort independent of the speaking rate. This increase in loudness is speaker dependent, which was reported in [61]. From Fig. 3.5 we have already seen that the peaks in the Hilbert envelope are sharper in case of loud speech compared to Lombard effect speech. Since loud speech is voluntary, it varies over a long range and it is difficult to distinguish loud speech from

Lombard effect speech directly. Fig. 4.8 (a), (b) and (c) show the scatter plots of durations of normal speech versus durations of Lombard effect speech, durations of normal speech versus durations of loud speech and durations of Lombard effect speech versus durations of loud speech, respectively. The external feedback for Lombard effect speech was pink noise at 70 dB. The points below the diagonal indicate that absicca is more than the ordinate, the points above indicate that ordinate is more than the absicca, and points on the diagonal indicate that the absicca and ordinate are equal. We can see that duration of loud speech is less than that of normal speech in many cases, indicating an increase in speaking rate. In other cases, the vocal effort is increased to increase the loudness. It is evident from Fig. 4.8 (c) that the duration of the Lombard effect speech is more than that of the loud speech for the same speaker. This represents a significant property that the duration of the Lombard effect speech of a speaker is greater than his loud speech. The direction of change in the excitation source features for both Lombard effect speech and loud speech with respect to normal speech is the same, whereas the direction of change in duration need not be the same.

Fig. 4.9 shows the distributions of the features for normal speech, loud speech and Lombard effect speech (under a feedback of pink noise at 70db) for 3 speakers. We can see that the increase in loudness varies from speaker to speaker in their corresponding loud speech. In case of speaker C, the change in the features for his loud speech is more than that of his Lombard effect speech, which is not the trend in case of speakers A and B. Thus a person can have his/her loudness of their loud speech more/less compared to their Lombard effect speech, but the duration of his/her loud speech is always less than the duration of his/her Lombard effect speech.

4.8 Summary

In this chapter, we have described the method of data collection and also discussed specch produced under various conditions using the proposed excitation source features. The various conditions involve speech under various types and intensities of external feedack, speech with external feedback only through a single ear, speech without feedback, speech produced in the presence of external feedback when the person pretends he is not under the influence of external feedback. Finally we have addressed the difference between Lombard efefct speech and loud speech. The main contributions in this chapter involve the study of speech produced without self-feedback, speech produced under different types of external feedback and differentiating Lombard efefct speech from loud speech.

Chapter 5

Analysis of Lombard Effect Speech

In this chapter, analysis of Lombard effect speech is performed based on the proposed features and also based on the change in duration. Analyzing Lombard effect speech enables us to study the speaker specific nature of Lombard effect and also to study the affect of feedback on the human speech production. Analysis is also performed on the perception of Lombard effect speech using the intelligibility factor. Finally the mechanism of Lombard effect is described based on the analysis. In Section 5.1, Lombard effect speech is analyzed based on the proposed features. In Section 5.2, duration analysis is performed. Section 5.3 discusses the intelligibility of Lombard effect speech. Section 5.4 describes the mechanism of Lombard effect speech. Summary of the chapter is presented in 5.5.

5.1 Feature analysis

In this section, we use the proposed features described in Chapter 3, to analyze their changes due to Lombard effect for all the speakers in the database.

5.1.1 Distribution of features in Lombard effect speech

Apart from the changes in the proposed features, the distributions of the features are also important in the characterization of Lombard effect speech relative to normal speech.

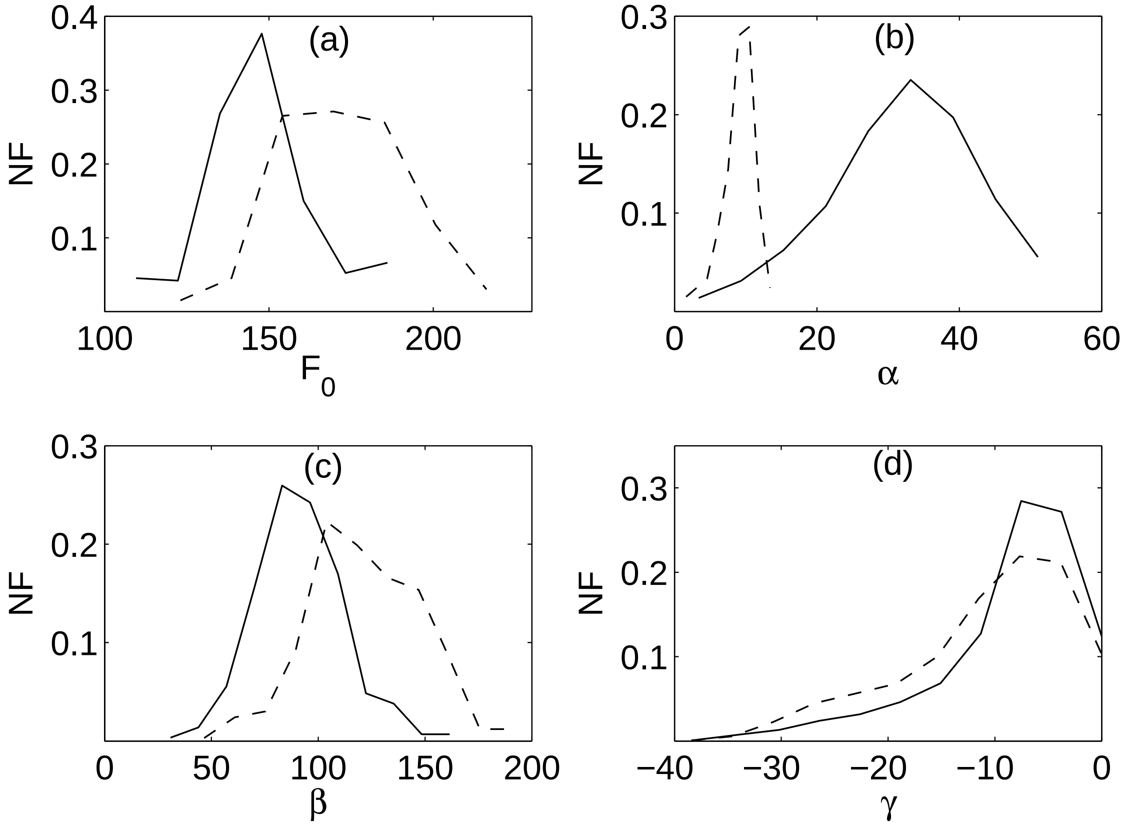

Fig. 5.1: Distributions of (a) F_0, (b) α, (c) β, and (d) γ, for normal speech (solid lines) and Lombard effect speech (dashed lines). (NF=Normalized frequency)

Fig. 5.1 shows the distributions of the features, F_0, α, β and γ, for both normal speech and Lombard effect speech. We can see that the distributions of F_0 and β have higher mean and spread, whereas α has the opposite trend for Lombard effect speech. In case of the distribution of γ, we can see that Lombard effect speech has more lower normalized energy frames and less higher normalized energy frames than for the normal speech. Thus the distribution of normalized energy is another feature which can differentiate Lombard effect speech from normal speech. This observation can be used for voice conversion of Lombard effect speech to normal speech, by adjusting the relative energies of the sound units. It may also be useful for feature compensation of Lombard effect speech for a better performance of speech systems. These distributions reflect the variations seen in the contours of the features shown in Figs. 3.2, 3.3, 3.7 and 3.8.

The variations in the features in Lombard effect speech are due to sudden changes in the features, especially in the vowel regions. Thus the variations in a given Lombard effect speech signal is dependent on the text and also on the articulation of the speaker. The deviations of the feature contours in Lombard effect speech relative to the normal speech is generally high in the consonant-vowel transition regions. The amount of deviation can be measured by the standard deviation of the feature contour for a given speech signal. This deviation is another feature which can be used to differentiate the Lombard effect speech from normal speech.

5.1.2 Variation of features across speakers

The Lombard effect was found to be different for different speakers. Fig. 5.2 shows the distribution of F_0, α, β and γ of normal speech and Lombard effect speech (under pink noise at intensity 70db) for 3 different speakers. The distributions of the features do discriminate between normal speech and Lombard effect speech. The degree of separation between the features indicates the extent of Lombard effect. The degree of discrimination between features is less in the case of Speaker B, compared to Speakers A and C, indicating that the Lombard effect on Speaker B is less compared to speakers A and C. The distribution of the excitation features also indicate the range of variations in speech that a speaker can produce. The reason for the differences among speakers is due to inter-speaker variability, since the increase of intensity from normal to Lombard effect speech varies from speaker to speaker. The inter-speaker variability may also be due to different hearing conditions of the speaker.

The extent of Lombard effect is measured as the separation between the distributions. This separation depends on the mean and standard deviation of the distributions of the features. The distributions of the features can be approximated by a Gaussian probability density function. Here we use a distance measure to approximate the separation between the distributions. The Kullback-Leibler (KL) divergence [69] is used to measure the deviation between two distributions. The KL divergence between two sets $\mathcal{X}$ and $\mathcal{Y}$ described by univariate Gaussian probability density functions is measured as [54]

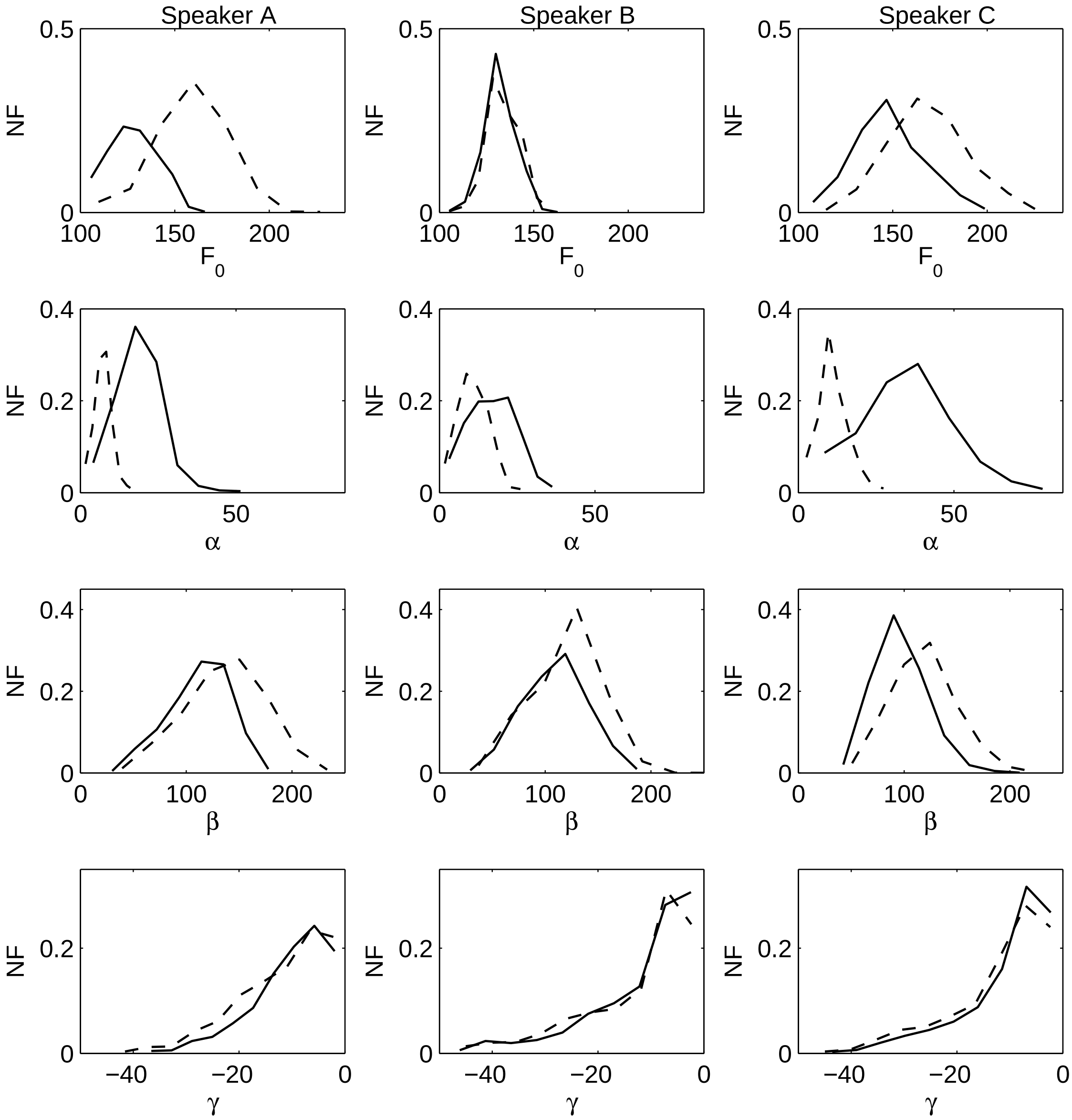

Fig. 5.2: Distribution of F_0, α, β and γ for normal speech (solid line), Lombard effect speech (dashed line) with pink noise of intensity 70dB as external feedback for 3 speakers. (NF=Normalized frequency)

$$d_{KL}(\mathcal{X},\mathcal{Y}) = \frac{1}{2}\left\{\frac{\sigma_{\mathcal{X}}{}^2}{\sigma_{\mathcal{Y}}{}^2}+\frac{\sigma_{\mathcal{Y}}{}^2}{\sigma_{\mathcal{X}}{}^2}\right\}-1 + \frac{1}{2}\left\{\mu_{\mathcal{X}}-\mu_{\mathcal{Y}}\right\}^2\left\{\frac{1}{\sigma_{\mathcal{X}}{}^2}+\frac{1}{\sigma_{\mathcal{Y}}{}^2}\right\}, \tag{5.1}$$

where $\mu_{\mathcal{X}}$, $\mu_{\mathcal{Y}}$ represent the means and $\sigma_{\mathcal{X}}$, $\sigma_{\mathcal{Y}}$ represent the standard deviations of

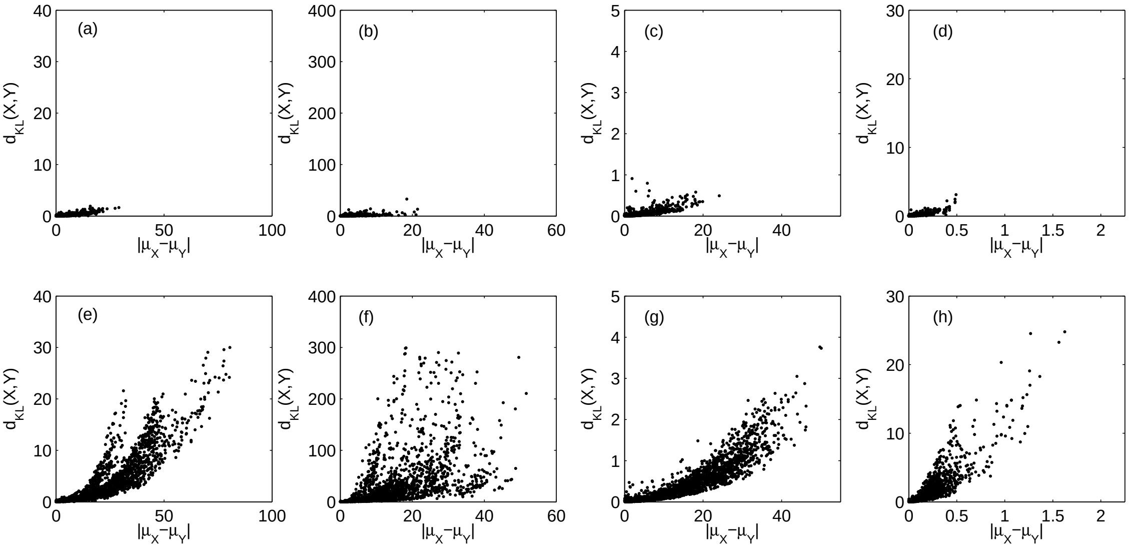

Fig. 5.3: Illustration of variation of the features F_0, α, β and γ. The plots (a), (b), (c) and (d) correspond to variations of the features F_0, α, β and γ for intra class comparisons, respectively. The plots (e), (f), (g) and (h) correspond to variations of the features F_0, α, β and γ for inter-class comparisons, respectively.

the sets $\mathcal{X}$ and $\mathcal{Y}$, respectively. The absolute value of the difference between the mean values $|\mu_{\mathcal{X}} - \mu_{\mathcal{Y}}|$ is computed. In this study we consider the samples in the sets $\mathcal{X}$ and $\mathcal{Y}$ as the values of the features F_0, α, β and γ. We consider the Lombard effect speech produced as a result of an external feedback of pink noise at 70 dB, babble noise at 65 dB and factory noise at 65 dB. We have data of five sentences from each of 18 speakers, with each sentence spoken three times under normal conditions and under the above three Lombard effect conditions. We consider the following two cases: (a) When the features in both the sets $\mathcal{X}$ and $\mathcal{Y}$ are derived from normal speech. (b) When the features in both the sets $\mathcal{X}$ and $\mathcal{Y}$ are derived from Lombard effect speech. Both the cases represent intra-class comparisons. In both the cases $d_{KL}(\mathcal{X}, \mathcal{Y})$ and $|\mu_{\mathcal{X}} - \mu_{\mathcal{Y}}|$ are small. Inter-class comparisons are those where the values of the features in $\mathcal{X}$ and $\mathcal{Y}$ are derived from the normal speech and Lombard effect speech, respectively. In this case $d_{KL}(\mathcal{X}, \mathcal{Y})$ and $|\mu_{\mathcal{X}} - \mu_{\mathcal{Y}}|$ are expected to be larger than in the case of intra-class comparisons. The ordered pair $(|\mu_{\mathcal{X}} - \mu_{\mathcal{Y}}|, d_{KL}(\mathcal{X}, \mathcal{Y}))$ is used to discriminate between normal and Lombard effect speech of a given speaker, as described below.

Let $\mathcal{N}$ denote the set of values of instantaneous F_0 for a given speaker, derived from

15 (5 sentences × 3 repetitions) utterances of normal speech. Let us consider a single sentence with 3 repetitions. Let $\mathcal{N}_{\infty}$, $\mathcal{N}_{\in}$ and $\mathcal{N}_{\ni}$ denote three distant subsets of $\mathcal{N}$, such that the values of instantaneous F_0 in each subset is derived from the 3 repetitions of the sentence. For the same speaker and sentence, let P_1 (B_1, F_1), P_2 (B_2, F_2) and P_3 (B_3, F_3) denote the set of values of instantaneous F_0 for the speech produced under the influence of Lombard effect with an external feedback of pink noise at 70dB (babble noise at 65dB, factory noise at 65dB) for the 3 repetitions of the sentence. For each speaker, sentence and an external feedback, the following ordered pairs are computed: a) $(|\mu_{\mathcal{N}_i} - \mu_{\mathcal{N}_j}|, d_{KL}(\mathcal{N}_i, \mathcal{N}_j))$ for i=1, 2, 3, j=1, 2, 3, $i < j$. (b) $(|\mu_{\mathcal{P}_i} - \mu_{\mathcal{P}_j}|, d_{KL}(\mathcal{P}_i, \mathcal{P}_j))$ for i=1, 2, 3, j=1, 2, 3, $i < k$. (c) $(|\mu_{\mathcal{N}_i} - \mu_{\mathcal{P}_j}|, d_{KL}(\mathcal{N}_i, \mathcal{P}_j))$ for i=1, 2, 3, j=1, 2, 3. The ordered pairs in (a) and (b) correspond to intra-class comparisons, while those in (c) correspond to inter-class comparisons. Thus, for a specific speaker, a specific sentence and a specific external feedback case, 6 points (3 normal + 3 Lombard) are computed due to intra-class and 9 points are computed due to inter-class comparisons. Generalizing this to all the 3 external feedbacks we get 12 points (3 normal + 3 × 3 Lombard) of intra-class comparisons and 27 points (9 × 3) of inter-class comparisons. Thus for all the 5 sentences we get 60 intra-class comparison points and 135 inter-class comparison points. Note that $d_{KL}(\mathcal{X}, \mathcal{Y}) = d_{KL}(\mathcal{Y}, \mathcal{X})$. We perform the same procedure to the features α, β and γ.

Each ordered pair can be plotted in a two-dimensional plane. Fig. 5.3 (a), (b), (c) and (d) show the intra-class points, and Figs. 5.3 (e), (f), (g) and (h) show the inter-class points of the features F_0, α, β and γ, respectively. The intra-class points are clustered closer to the origin compared to the inter-class points, which moved farther from the origin. Thus the divergence obtained by comparing normal speech and Lombard effect speech is high. This indicates that the distributions of the features does help in distinguishing normal speech and Lombard effect speech. Among all the four features, the spread in α is more due to Lombard effect. The points in intra-class comparisons which are away from the origin are due to high intra-speaker variability.

5.1.3 Perceptual Evaluation using loudness

Perceptual evaluation was carried out by conducting subjective test with 10 listeners in the age group of 21-23 years. Experiments were carried out in a laboratory environment. Two speech files, a normal speech and a Lombard effect speech of the same sentence spoken by the same person were played to the subjects through headphones. The listeners were asked to choose the louder of the two. If they perceive the pair as equally loud, they were asked to choose the option “can’t say”.

Fig. 5.4 shows the distribution of perception evaluation results for four speakers. The results for these four speakers were selected here to illustrate the variety of responses that can be obtained in perceptual studies on Lombard effect speech. Row 5 in Fig. 5.4 show the results of perceptual evaluation. We can see that the loudness of normal speech and Lombard effect speech was perceived to be similar (can’t say (CS) response) in the cases of speaker A and speaker B. For speaker C, the perception results are varied, indicating that the Lombard effect may be small. For speaker D all the subjects perceived the Lombard effect speech as loud.

These observations can also be interpreted in terms of the distributions of the parameters. The distribution of F_0 and β are distinctly different for Lombard effect speech compared to that for normal speech for Speaker D which is also reflected in the perception results. In case of speakers A, B and C, there is no significant difference in the distribution of F_0 and β. This is again reflected in the perceptual results where the listeners had failed to recognize Lombard effect speech. The strength of excitation (α) at epochs is lower for Lombard effect speech compared to normal speech for all the speakers. α can differentiate between normal speech and Lombard effect speech even though the F_0 and β are not distinctly different. This is reflected mainly in case of speakers B and C. Thus α is an important feature which can distinguish between normal speech and Lombard effect speech even though perceived loudness (β) fails to differentiate both. Note that α gives only the amplitude of the impulse-like excitation at each epoch, and it need not necessarily indicate loudness. The loudness is due to sharpness of the impulse-like behavior in excitation around the epochs, and it is better represented by the parameter β.

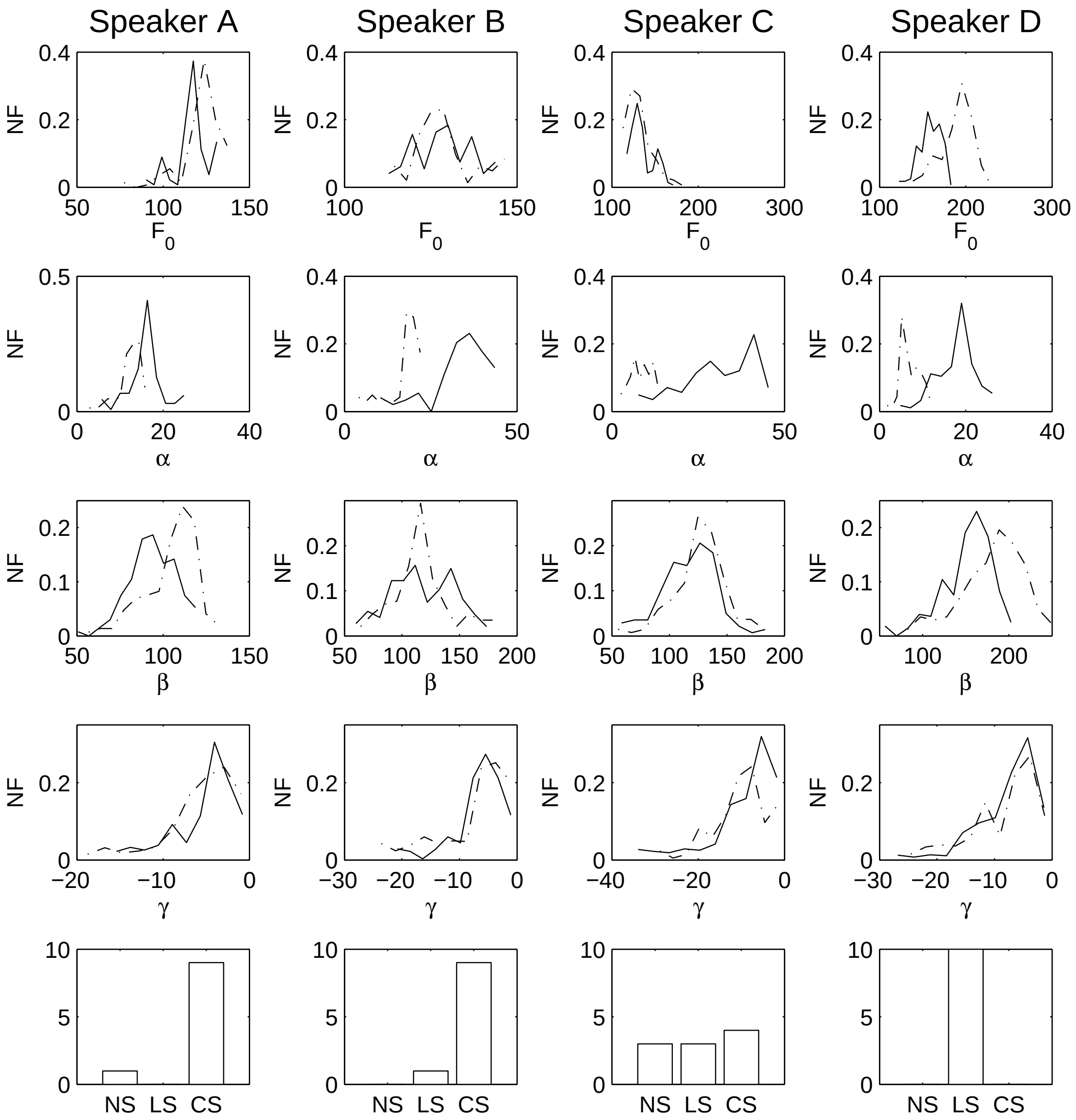

Fig. 5.4: Distribution of F_0, α, β, γ and perceptual evaluation results for normal speech (solid lines) and Lombard effect speech (dash-dotted lines) for 4 speakers. (NS=normal speech, LS=Lombard effect speech, CS=can't say, NF=Normalized frequency).

5.2 Duration analysis

Apart from the excitation features and system features, Lombard effect also causes changes in the individual sound units. The significant changes are in the duration. Some studies reported increase in the duration of vowels and decrease in the duration of stops, fricatives and silence regions [70]. Here we examine the change in the duration and the relative energy in case of nasals and fricatives. We also analyze the change in duration at a sentence level for all the cases described.

5.2.1 Duration and relative energy of nasals and fricatives

Durations of the nasals and fricatives were found to decrease for Lombard effect speech compared to normal speech. Their relative energies were also seen to decrease compared to other voiced parts for Lombard effect speech. Figs. 5.5(a) and (b) show normal speech signal and its normalized energy for the word “companions”, and Figs. 5.5(c) and (d) show Lombard effect speech and its normalized energy, respectively for the same word. We can see that the relative energy for the nasals /m/, /n/ and fricative /s/ decrease in the case of Lombard effect speech compared to normal speech. An analysis on loudness has shown that increase in loudness for nasals due to Lombard effect is less than the increase in loudness for other voiced sound units.

5.2.2 Duration of sentence

The change in durations of sentences in the case of Lombard effect speech depends on the sound units in the given text, and also on the speaker. Thus the duration change is speech and speaker dependent. The part of speech which is important in speech communication, from the talker and listener point of view, is stressed more by increasing the loudness and duration. In the part that is not so significant, the speaker may even try to rush through it. Since the durations of vowels increase and that of nasals, stops, fricatives and silence regions decrease, the overall change in duration depends on the number and nature of the different sound units in a given sentence. The change also depends on the articulation of the speaker. It was found that the duration of a sentence increases due to Lombard effect in most of the cases. This is contrary to the observations reported in [63], where the average duration of a sentence was found to decrease due to Lombard effect. It might be because there are very few silence regions in the sentences chosen in our case, whereas in [63] it was reported that the decrease in duration was mainly due to decrease in the silence regions. Thus the duration depends on the sound units and the silence regions in the sentence.

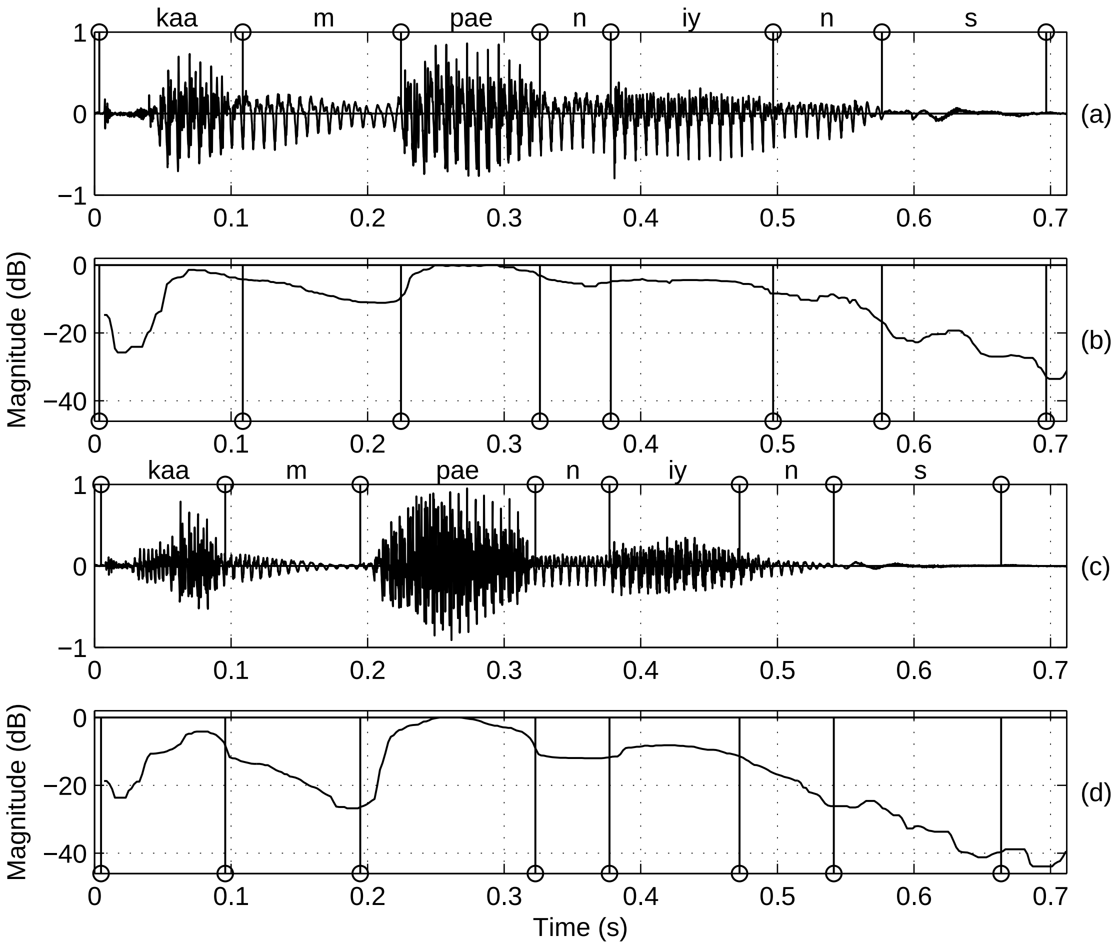

Fig. 5.5: (a) normal speech, (b) normalized energy of the normal speech, (c) Lombard effect speech, (d) normalized energy of the Lombard effect speech, for the utterance, "companions".

The regions of a given speech signal can be classified into 3 parts:

- Voiced regions
- Unvoiced regions
- Silence regions

The percentage of voiced speech is calculated by performing voiced/nonvoiced segmentation using the method proposed in [71]. Voiced speech is produced when the vocal folds are vibrated continuously due to the glottal excitation. In the absence of vocal fold vibration, the vocal tract system is considered to be excited by random noise, which forms the unvoiced speech. The energy of the random noise excitation is distributed both in time and frequency domains, where as the energy of an impulse is distributed uniformly in the frequency domain. As a result, the filtered signal exhibits significantly lower amplitude for random excitation compared to the impulse-like excitation. Fig. 5.6(b) shows the zero-frequency filtered signal of the speech signal shown in Fig. 5.6(a). Fig. 5.6(c) shows the energy of the zero-frequency filtered signal. It can be seen that the energy of the zero-frequency filtered signal is high in the voiced regions compared to non-voiced regions. The binary voiced-nonvoiced signal is computed as

$$d_{vnv}[n] = \begin{cases} 1, & \text{if } y_{zfr}[n] > 0.5 \\ 0, & \text{otherwise,} \end{cases} \tag{5.2}$$

where $y_{zfr}[n] = 1 - e^{(-10 \times v_{zfr}[n])}$. Fig. 5.6(d) shows the voiced, nonvoiced regions of the speech signal shown in Fig. 5.6(a).

The percentage of voiced speech was found to increase due to external feedback. Table 5.1 shows the percentage duration of voiced regions for 5 speakers for the sentence “I know Sir John will go, though he was sure it would rain cats and dogs” for normal speech and Lombard effect speech. We can see that there is increase in the percentage duration of the voiced region in the case of Lombard effect speech which varies from speaker to speaker. The percentage decrease in the duration of the nonvoiced region is generally more than the percentage increase in the duration of the voiced region.

Table 5.1: Percentage duration of voiced region for normal speech and Lombard effect speech for the utterance ‘I know Sir John will go, though he was sure it would rain cats and dogs’ for 5 speakers.

	Normal speech	Lombard effect speech
	Voiced	Voiced
Speaker 1	84	89.67
Speaker 2	85.35	89
Speaker 3	73	86.5
Speaker 4	77.7	86
Speaker 5	75.77	77.9

Duration and percentage of voiced speech are measured for different cases of external feedback for different utterances. Table 5.2 shows the durations and percentage voiced speech for the utterance of a sentence under different conditions. We can see that the

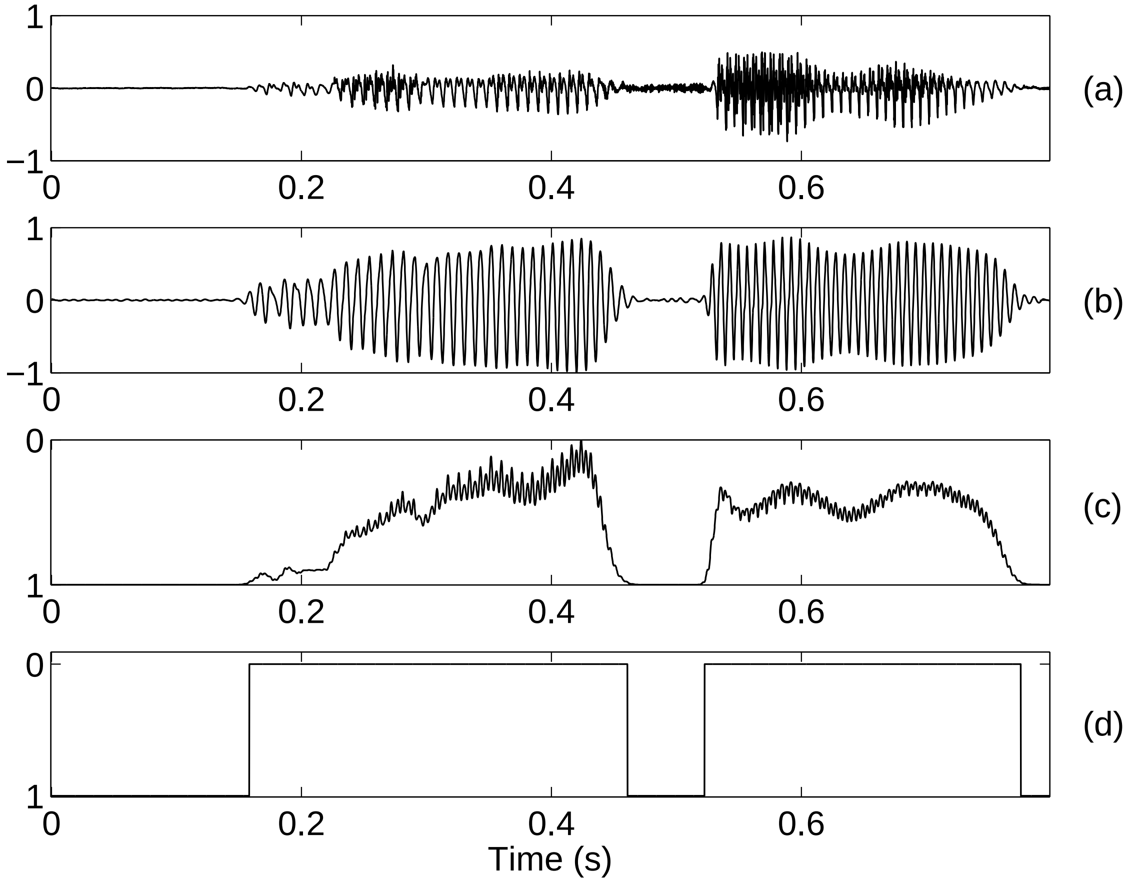

Fig. 5.6: Illustration of voiced-nonvoiced segmentation of a speech signal using zero frequency filter. (a) Segment of a speech signal, (b) zero frequency filtered signal, (c) energy of the zero frequency filtered signal, and (d) binary voiced-nonvoiced signal.

percentage of voiced speech decreases with decrease in intensity of feedback. Another interesting observation is that the percentage of voiced speech decreases even for no-feedback case. An increase in the duration for the no-effect case signifies that the speaker reduces his speaking rate while pretending that he is not under the influence of an external feedback, which might be one useful factor to detect such cases. In the case of loud speech the duration is less, and the percentage voiced region is more than that for normal speech.

It was reported in [21] that in the presence of noise the speaker tends to reduce his speaking rate and increase the duration of the utterance. But this is not necessarily true always, as the duration of an utterance depends on several other factors.

Table 5.2: Duration and percentage voiced speech for the utterance 'They remained lifelong friends and companions' in different speaking conditions for all the 18 speakers in the database

Condition	Duration	% Voiced
Normal	1.97	83.9
PN-70	2.02	88
PN-60	2.02	85.6
PN-50	2.01	84.5
BN-65	2.02	87.2
BN-55	2.00	86
BN-45	1.99	83.1
FN-65	2.04	87
FN-55	2.02	84.5
FN-45	1.95	84.4
SEF (PN-70)	1.98	83.1
NF	1.99	83.3
NE (PN-70)	2.02	83.2
Loud	1.93	85.9

5.3 Intelligibility and perception of Lombard effect speech

Intelligibility of speech refers to the accuracy with which a normal listener can understand a spoken word or phrase. Speech intelligibility is reduced due to noise (i.e., low signal to noise ratio). In noisy environment, for effective communication, the speaker tends to increase his intensity of speaking, so as to increase the signal to noise ratio, thus maintaining the intelligibility. The speaker increases his vocal effort (loudness) to maintain intelligibility of his speech under noisy conditions. It was reported in [29] that, for the same signal to noise ratio, isolated words or continuous speech produced in noise are more intelligible than speech produced in quiet. Normal connected speech can be understood even if some of the syllables are unintelligible, as the listener can still deduce the meaning from the context of the sentence. It was reported in [72] that word intelligibility is higher in a sentence context than in an individual form. The duration of content words are prolonged to a greater degree in noise than function words [73]. This ensures that intelligibility is not lost even though the utterance is a word or a short phrase. Changes in duration and energy at sentence and sound unit levels ensure that the intelligibility is not reduced due to additional noise. A person increases the signal to noise ratio, duration of certain sound units like the vowels, so that the utterance spoken is understood well. Since, Lombard effect speech is more intelligible than the speech under normal conditions, the changes in the features due to Lombard effect can be estimated as a cause for the increase in intelligibility. Note that change in features above a certain level reduces intelligibility, where the speech corresponds to shouted speech.

Thus Lombard effect speech is perceived to be louder than the normal speech so as to preserve the intelligibility even under noisy conditions. Another common perceptual observation in differentiating normal speech and Lombard effect speech is that in normal speech we observe that the person reduces his loudness at the end of the utterance in many cases, whereas under the influence of Lombard effect, the level of loudness is maintained throughout the utterance. The relative energy at the end of the sentence is seen to be low for normal speech in most of the cases. Speech produced highly depends on the environment which can be explained by the intelligibility factor. We increase the intensity of speaking in the presence of noise. In the presence of an echo, we tend to decrease the intensity of speaking and also speech rate. The reason is intelligibility. With increase in intensity of speaking in noise, we are increasing the intelligibility, with the same in reverberant environment we are decreasing the intelligibility.

Several evaluation tests were performed to show that the intelligibility of Lombard effect is more than that of normal speech in presence of noise. The vocabulary generally used were digits [8], confusable words [7]. But in a real time environment, it is essential to study the intelligibility using a spoken utterance. It is also essential to study intelligibility under different noisy conditions. Utterances from various speakers were randomly selected and a set is formed which contains two utterances each of speech added to PN-70, BN-65, FN-65, 2 utterances each of Lombard effect speech produced under the external feedback of PN-70, BN-65, FN-65 added to the respective noises. Thus a set contains 12 utterances. All the random selected utterances are taken from the database. A listener is alloted a set and asked to recognize what is spoken by the speaker. Intelligibility is then calculated as the percentage of sentence recognized correctly. The number of utterances are limited so as to avoid the listener in getting familiar with the utterances as the total unique text sentences are only five.

Table 5.3: Intelligibility (in %) of normal speech and Lombard effect speech under different types of noise

	Speaking condition			
Noise	Normal	PN-70	BN-65	FN-65
PN-70	59.72	87.61	-	-
BN-65	53.17	-	80.82	-
FN-65	42.27	-	-	75.35

The goal of the perceptual evaluation is to find the effect of noise on speech intelligibility and also to characterize the intelligibility of speech produced under different types of noise. Table 5.3 shows the intelligibility of speech under different conditions. These conditions are classified into two types:

- Normal speech added to different types of noise.
- Lombard effect speech added to noise under which it is produced.

We do not consider the other case of Lombard effect speech added to noise under which it is not produced, as the changes in intelligibility due to this is minimum and it requires a large database and a large number of listener tests to accurately study this effect. We can clearly see that intelligibility affected by factory noise is more than that of pink noise and babble noise and pink noise is shown to affect the least. Though the intelligibility of Lombard effect speech in noise is high, it is still not close to 100%. The reason for this depends on the method of data collection which will be discussed in the next section.

5.4 Mechanism of Lombard effect

In this section we describe the mechanism of Lombard effect from the experiments performed. Firstly we need to understand the recording procedure during data collection. Recording procedure is important as it will be used to describe the mechanism of Lombard effect. The speaker is given a set of sentences written on a paper and is asked to speak them by looking at the paper. Thus the speaker is communicating with himself

(self-communication) and not with others i.e., a self-regulatory feedback system. But from the analysis we have found that the intelligibility of this Lombard effect speech is more than that of normal speech in the presence of noise which was previously explained as a reason for effective communication. But in our case, the speaker has spoken in self-communication mode and not in effective communication mode. Thus hampering of the self feedback does result in Lombard effect irrespective of the mode of communication. For simplicity we refer the self-communication mode (self-regulatory feedback system) as private loop and effective communication mode as public loop.

Another inference from the analysis of Lombard effect in our case (which is the private loop), is the speaker dependence. Large range of variations between normal speech and Lombard effect speech of various speakers can be observed from Fig. 5.2. Few speakers have shown high variations and few have shown less variations. Since increase in intelligibility of the Lombard effect speech depends on the change in the features, increase in intelligibility is not the same for all speakers. In public loop, the speaker will ensure intelligible communication and in private loop, the speaker exhibits Lombard effect whose extent depends on the speaker. Since Lombard effect speech is produced in private loop in our case, high intelligibility is not ensured, due to which the intelligibility of Lombard effect speech is not close to 100%.

It is commonly seen that during description of Lombard effect speech data, the mode of recording (private loop or public loop) is not mentioned. Thus it is necessary to consider the process of Lombard effect speech data collection for drawing inferences from the analysis of the data. The following are the inferences of Lombard effect speech produced under private loop and public loop:

- Lombard effect is found both in private loop and public loop.
- Lombard effect in private loop is highly speaker dependent.
- Intelligibility is always ensured in public loop.
- Intelligibility is not always ensured in private loop.

5.5 Summary

In this chapter, we have analyzed Lombard effect speech using the proposed features and also duration. Variation of the features across speakers due to Lombard effect is studied using the KL-divergence between the distributions which shows the speaker specific nature of Lombard effect. Perceptual evaluation of loudness for normal speech and Lombard effect speech is performed. Perceived loudness of Lombard effect speech is seen to be higher than that of normal speech. Intelligibility tests were also performed to show that Lombard effect speech is more intelligible than normal speech and also to study the intelligibility under various environment conditions. Based on the analysis performed, the mechanism of Lombard effect is described.

Chapter 6

Imposter Detection in Speaker Verification System Using Lombard Effect

Speaker recognition is the task of recognizing a person by a machine from the characteristics of his/her speech. Speaker recognition can be classified into two categories: Speaker identification and speaker verification. Speaker identification is the task of recognizing a person from a given list of persons whose speech characteristics are already stored by the system. Speaker identification becomes complex if the number of persons in the list is high. Speaker verification is the process of accepting or rejecting the claim of the speaker. Thus the complexity of speaker verification does not depend on the number of persons enrolled in the system. Speaker verification is further classified into three categories:

- Text-dependent: Text used for verification is the same as the text used during enrollment of the speaker.
- Text-independent: Text used for verification is independent of the text used during enrollment.
- Text-prompting: Here the system prompts a text to be spoken by the speaker during verification.

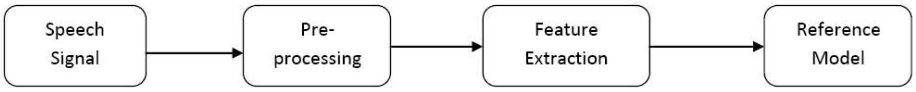

Fig. 6.1: Block diagram to show the process of enrollment phase.

In this chapter, we describe the text-dependent speaker verification system, its performance due to Lombard effect, types of imposter attacks and a method is proposed to detect imposters to improve the performance of the speaker verification system. In Section 6.1, we describe the text-dependent speaker verification system used. In Section 6.2, the performance of speaker verification using Lombard effect speech is discussed. Section 6.3 describes the types of imposter attacks. Section 6.4 shows the performance of speaker verification system due to imposter attacks. In Section 6.5, a method is proposed using Lombard effect to detect imposters using speaker verification. In Section 6.6, presents the summary of the chapter.

6.1 Text-dependent speaker verification system

A speaker recognition system consists of two modules: Enrollment and testing. In case of speaker verification system, the testing phase can also be called as verification phase. In the enrollment phase a person is registered into the system, his speech is recorded and a number of features are extracted and stored in the system to form a voice print or model or template which uniquely represents the speaker. Fig. 6.1 shows the process of enrollment. During verification phase, speech of the speaker is recorded and the features extracted are compared with the voice print of the claimed speaker who is already enrolled. Finally a binary decision (accept or reject) is made. Fig. 6.2 shows the process of verification. The following are the steps performed in the text-dependent speaker verification system:

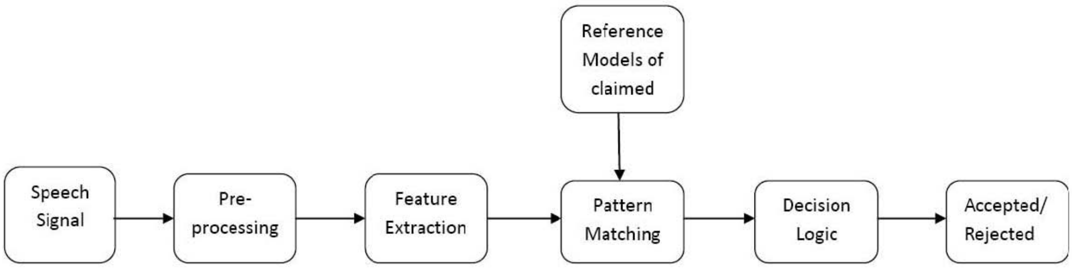

Fig. 6.2: Block diagram to show the process of verification phase.

6.1.1 Preprocessing

Preprocessing step consists of two steps: Preemphasis and segmentation. In the preemphasis step the speech signal is passed through a first-order FIR filter, to spectrally filter the signal. Preemphasis is necessary because, in the spectrum of a human speech signal, the energy in the signal decreases as the frequency increases. Preemphasis increases the energy in parts of the signal by an amount inversely proportional to its frequency. Thus, as the frequency increases, pre-emphasis raises the energy of the speech signal by an increasing amount. This process therefore serves to flatten the signal so that the resulting spectrum consists of formants of similar heights. The preemphasis system used here is the fixed first-order system:

$$H(z) = 1 - z^{-1}. \tag{6.1}$$

This is equivalent to differentiating the speech signal which is given by

$$\tilde{s}(n) = s(n) - s(n-1), \tag{6.2}$$

where $s(n)$ is the original speech signal, $\tilde{s}(n)$ is the speech signal obtained after preemphasis. The next step involves segmenting the speech signal into voiced, nonvoiced regions. The reason behind this is to consider only the voiced regions which has high speaker information and to omit silence regions which do not have any speaker information and unvoiced regions in which the speaker information is very low.

6.1.2 Feature Extraction

In this step features are extracted which form the voice print of the speaker. We use two different features.

- Mel-frequency cepstral coefficients (MFCC)
- Linear prediction cepstral coefficients (LPCC)

Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCC) is a robust feature used commonly for speaker verification [74]. It is a representation of a short-term power spectrum of a sound based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The advantage of considering mel frequency cepstrum over normal cepstrum is that in a mel frequency cepstrum, the frequency bands are equally spaced on a mel scale which approximates the human auditory system's response more closely than the linearly spaced frequency bands used in the normal cepstrum. A frame size of 20 ms and a frame shift of 10 ms is taken. 20 triangular bandpass filters were used. The MFCC are computed as follows:

$$MFCC_i = \sum_{k=1}^{20} X_k cos[i(k-1/2)\pi/20], i = 1, 2, ...M, \qquad (6.3)$$

where M is the number of cepstral coefficients and X_k, k=1,2,...,20 represents the log-energy output of the k^{th} filter. The size of the feature vector is 13. Cepstral mean subtraction is used to reduce channel variation [75].

Linear prediction cepstral coefficients

Linear prediction coefficients (LPCC) are obtained by linear prediction analysis of speech signal which predicts a given speech sample at time instant n as a linear weighted sum of the previous p samples and the predicted sample is given by [76][77]

$$\hat{s}(n) = \sum_{k=1}^{p} a_k s(n-k), \tag{6.4}$$

where *s(n)* is the speech sample at time n, and a_k, $k = 1, 2, ...p$, is the set of predictor coefficients. The prediction error *e(n)* is defined as

$$e(n) = s(n) - \hat{s}(n-1). \tag{6.5}$$

The mean square of the prediction error over an analysis frame of N samples is given by

$$E = \sum_{n=0}^{N-1} e^2(n). \tag{6.6}$$

Minimizing E with respect to the set of predictor coefficients a_k results in a set of p normal equations. The set of predictor coefficients is obtained by solving the p normal equations. p is taken as 13 in our case.

Linear prediction cepstral coefficients were obtained from the linear prediction coefficients (a_k) directly as [78]

$$LPCC_i = a_i + \sum_{k=1}^{i-1} \frac{k-i}{i} LPCC_{i-k} a_k, i = 1, 2, ..., 13. \tag{6.7}$$

6.1.3 Pattern Matching

After obtaining the sequence of features for the enrolled speech and verification speech, the task is to match the sequences. Since the time intervals of the enrolled speech and verification speech varies, the number of feature vectors varies. We need to match the similarity of two feature vector sequences which differ in time. Euclidean distance is used to estimate the difference between the features. The Euclidean distance between the reference feature vector of frame i and test feature vector of frame j is given by

$$\sum_{n=1}^{p}(r_i(n) - t_j(n))^2, \tag{6.8}$$

where p is the length of the feature vector, r_i is the reference feature vector of frame i and t_j is the test feature vector of frame j. We use the dynamic time warping (DTW) algorithm to align the samples of the two signals. A matching score is then obtained which indicates the similarity between the two speech signals. Lesser the score, more is the similarity between the signals. Matching score is zero for two identical signals.

6.1.4 Decision Logic

Based on the matching score obtained we need to make a decision which is binary, i.e., accept or reject. A threshold is used to make a binary decision. If the matching score is below a threshold, the decision is accept, else reject. Threshold is decided based on the equal error rate (EER). To explain EER we define two terms:

- Missed detection rate (MDR): It is the rate of false rejection.
- False alarm rate (FAR): It is rate of false acceptance.

EER is the rate at which both MDR and FAR are equal. The threshold at the point where MDR and FAR are equal is taken as the threshold of the system for decision making with the most optimal results.

6.1.5 Results

Data described in Chapter 4 is used for verification. A speech sample of a person is taken as reference speech and another speech sample of the same speaker and same utterance is taken as verification speech for a genuine speaker testing and a speech sample of a

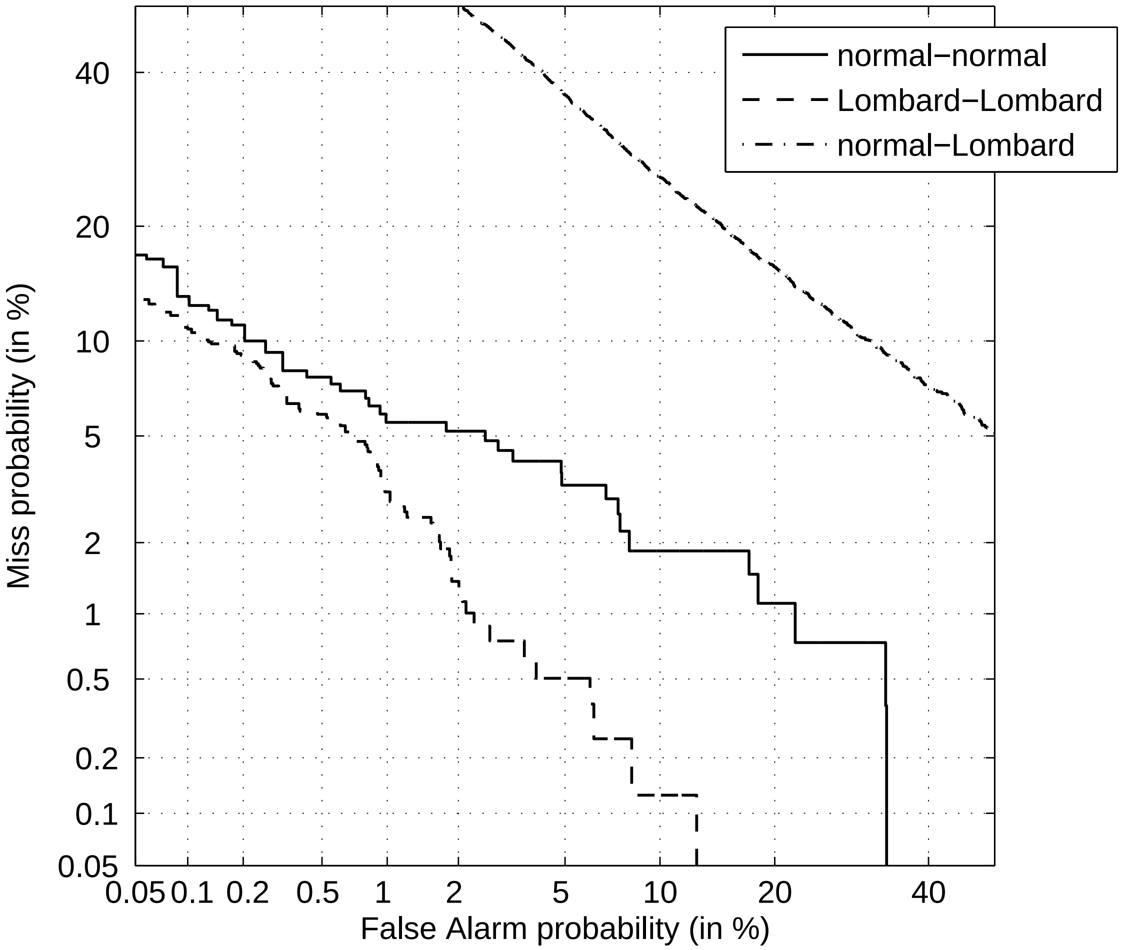

Fig. 6.3: DET curves indicating the performance of speaker verification based on MFCC for the three conditions normal-normal, Lombard-Lombard and normal-Lombard.

different speaker of the same utterance is taken as verification speech for imposter testing. Three sets of experiments were performed for speaker verification:

- Normal speech as both reference and test speech.
- Lombard effect speech as both reference and test speech.
- Normal speech as reference speech and Lombard effect speech as test speech (or vice-versa).

The performance of speaker verification for the above 3 cases is indicated in the detection error tradeoff (DET) curves in Fig. 6.3 for MFCC as the feature and Fig. 6.4 for LPCC as the feature.

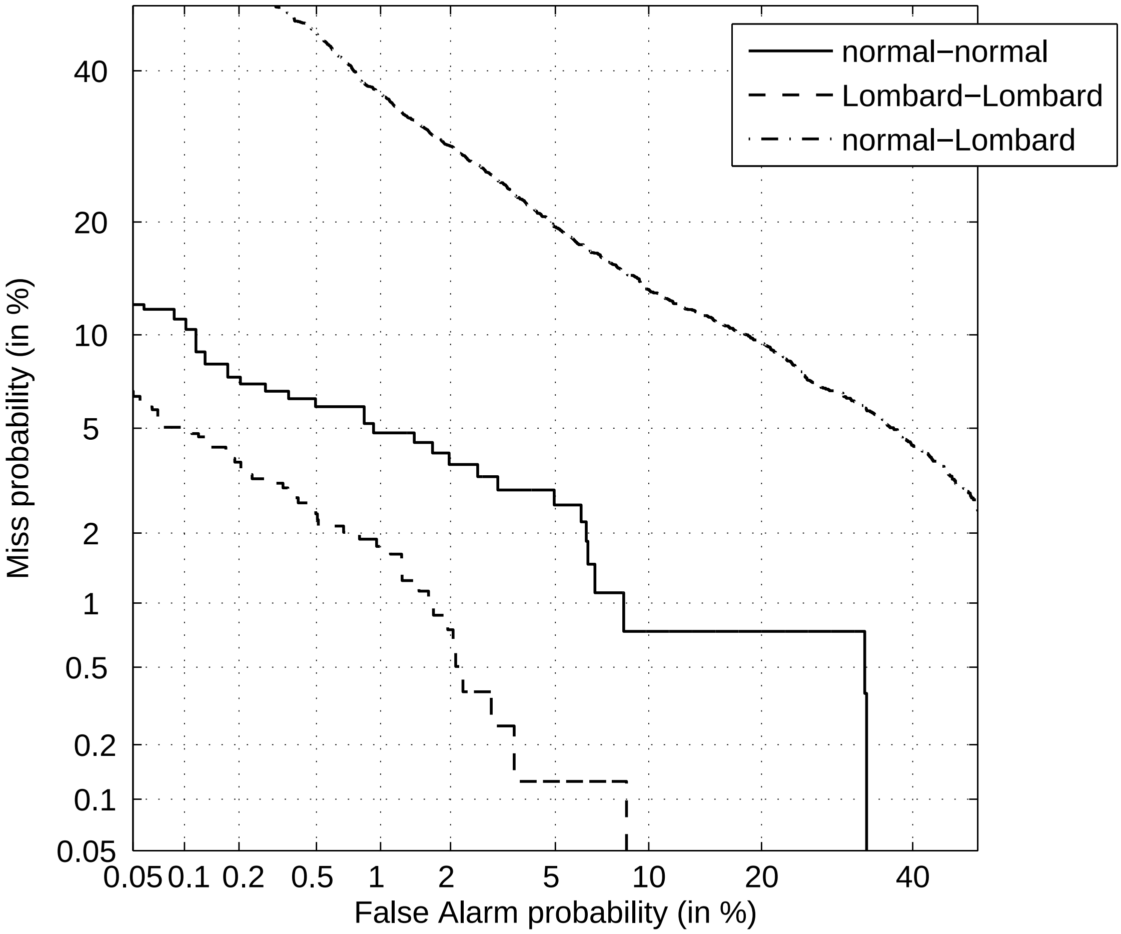

Fig. 6.4: DET curves indicating the performance of speaker verification based on LPCC for the three conditions normal-normal, Lombard-Lombard and normal-Lombard.

Table 6.1: EER (in %) for speaker verification using MFCC under different conditions.

Test	Reference	
	Normal	Lombard
Normal	4.1	17.5
Lombard	17.5	1.8

Table 6.1, 6.2 shows the equal error rate (ERR) for the speaker verification system using MFCC and LPCC as feature vectors, respectively for the matched conditions and unmatched conditions. We can draw two inferences from the results:

- LPCC perform better than MFCC in the text-dependent speaker verification system.
- In matched conditions (i.e., the condition of reference speech is the same as that of

Table 6.2: EER (in %) for speaker verification using LPCC under different conditions.

Test	Reference	
	Normal	Lombard
Normal	3.1	13
Lombard	13	1.2

test speech), Lombard effect speech as both reference and test speech has shown better performance compared to normal speech as both reference and test speech. Note that the performance is always low under mismatched conditions.

6.2 Speaker verification using Lombard effect speech

The reason for the better performance of speaker verification in case of Lombard effect speech as both reference and test speech is due to the change in the feature vector from frame to frame, which is less in case of Lombard effect speech. This shows that the intra-speaker variability is less in case of Lombard effect speech compared to normal speech. Figs. 6.5 (a), (b), (c), 6.6 (a), (b), (c), show the speech signal, spectrogram of the MFCC feature vectors and the change in the MFCC feature vectors from frame to frame for normal speech and Lombard effect speech, respectively. The change in the feature vector from frame to frame is measured as the Euclidean distance between the successive frames. This change in the feature vector determines the change in the variability of the speech of the speaker.

Fig. 6.7 shows the distribution of the change in features from frame to frame shown in Figs. 6.5 (c) and 6.6 (c). We can see that the change is less between most of the successive frames in case of Lombard effect speech compared to normal speech. Thus the intra-speaker variability can vary over a large scale for a normal speech of a person, but his/her Lombard effect speech will show less intra-speaker variability. A similar observation can be found out even by taking LPCC as the feature vectors. The idea of the less change in the feature vectors between the successive frames evolved from the spectrogram of the feature vectors which shows less variations across the speech in case of Lombard effect speech compared to normal speech.

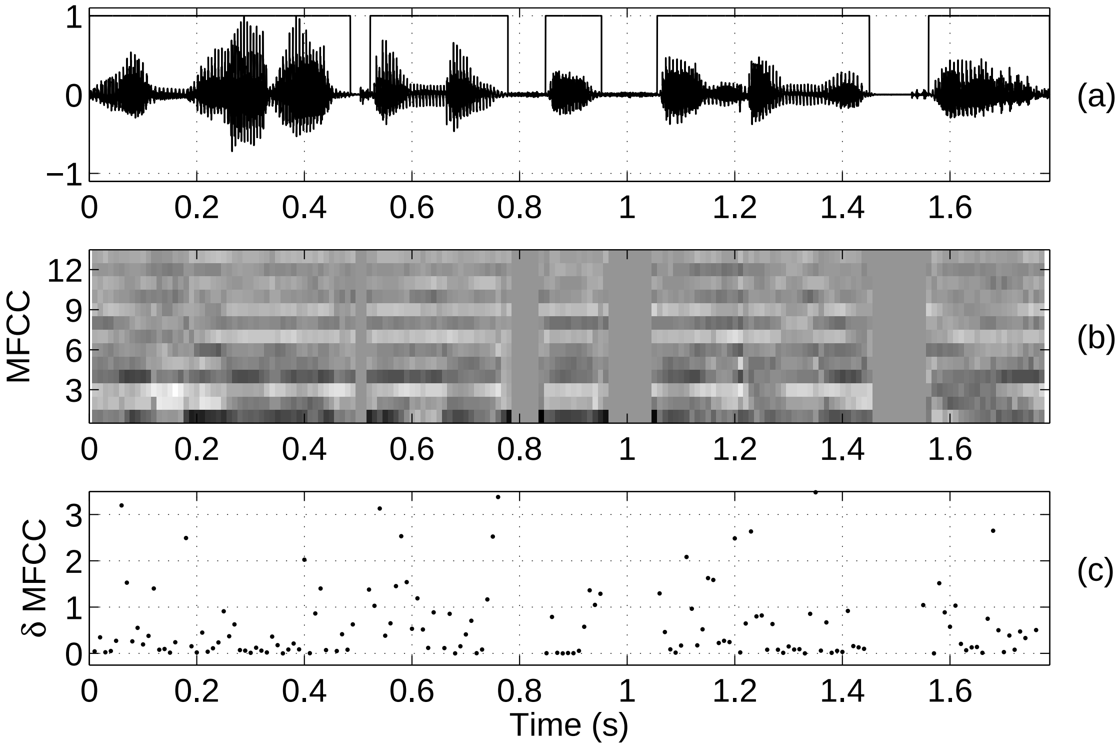

Fig. 6.5: (a) Speech signal, (b) spectrogram of the MFCC feature vectors of the speech signal, (c) change in the MFCC feature vectors from frame to frame.

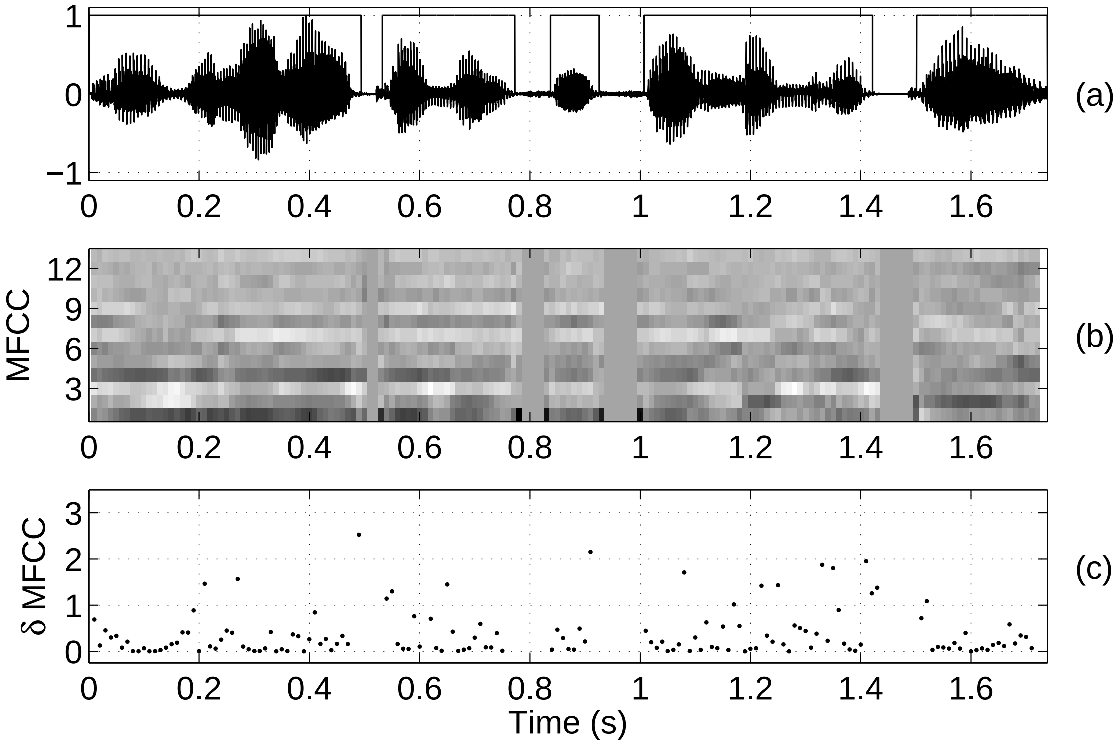

Fig. 6.6: (a) Lombard effect speech signal, (b) spectrogram of the MFCC feature vectors of the speech signal, (c) change in the MFCC feature vectors from frame to frame.

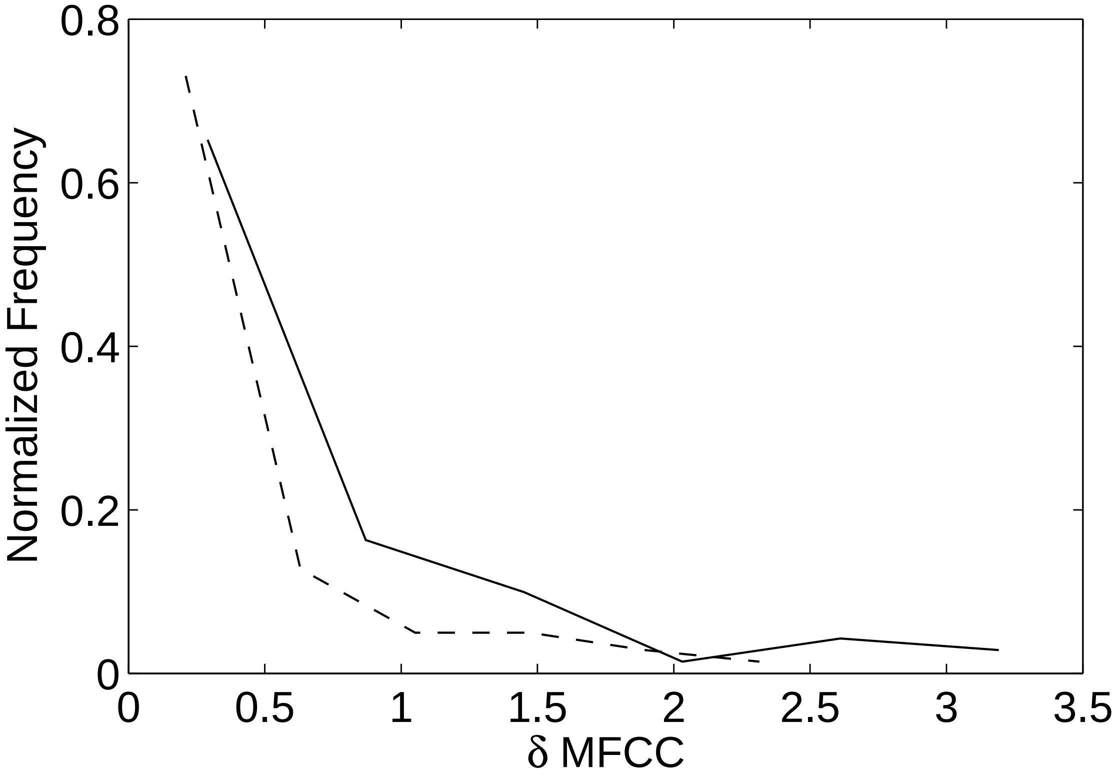

Fig. 6.7: Distribution of change in the feature vectors between successive frames for normal speech (solid lines) and Lombard effect speech (dashed lines).

This better performance of speaker verification system using Lombard effect speech as both reference and test speech also draws inroads to use of the Lombard effect in the future for robust speaker verification.

6.3 Types of imposter attacks

Imposters are the intruders who claim the identity of genuine speakers and try to gain access to the system. Such cases must be considered more than the case of rejecting a genuine speaker. We can classify the imposters into two categories:

a. Impersonation

b. Playback attack

6.3.1 Impersonation

Impersonation is the process of mimicking a person using the speaker-specific nature of his/her voice to gain access to the speaker verification system. The person who mimics other person is called an impersonator. An impersonator claims himself as a genuine speaker who is already registered in the system and mimics his voice to gain access to the system.

6.3.2 Playback attack

Playback attack is a situation where an intruder obtains the recording of a genuine speaker and plays it by claiming himself as that genuine speaker. With advancement in technology, recording and playing speech have become on easy process using the playback devices. In [79], various factors affecting the performance of playback attack detection were studied. In [80] a playback attack detector was proposed to avoid playback attacks on the speaker verification system. Here it was assumed that the data recorded while enrolling and verifying was stored in the system each time and that the intruder possesses the recording which is stored in the system. But the process becomes complex with more and more data stored in the system. Also it might be a case where the intruder might have a recording which is not stored in the system. An other case might occur where the genuine speaker might get rejected due to intra-speaker variability.

Playback speech is exposed to the environment more than once, while normal speech is exposed only once. By repeated exposure to the environment, the quality of playback speech degrades. If $x[n]$ is the actual speech, $r[n]$ is the random noise in the environment, $h[n]$ is the impulse response of the playback device,

$$y[n] = (x[n] + r[n]) * h[n], \tag{6.9}$$

where $y[n]$ is the playback speech signal. If we consider a noise-free environment, then r can be neglected. Playback speech consists of the characteristics of the playback device, as it is influenced by the impulse response of the playback device. The characteristics of the speech signal are modified due to every playback which causes degradation of the speech signal. After a number of repeated playbacks, the characteristics of speech and the speaker are completely lost and what remains is the response of the playback device. Fig. 6.8 (a),(b) shows the direct speech signal and its spectrogram, Fig. 6.8 (c),(d) shows the playback speech signal and its spectrogram obtained after playing the speech signal once and Fig. 6.8 (e),(f) shows the playback speech signal and its spectrogram obtained after repeatedly playing the speech signal 15 times. We can see the frequency band in Fig. 6.8 (f) which indicates the frequency response of the playback device which is left after repeatedly playing the speech signal 15 times.

Thus it is quite evident that the acceptance of the playback speech depends on the quality of the playback speech. This will be discussed in the next section.

6.4 Performance of the system due to imposters

We have already described the two types of imposters. Here we consider all the false acceptance cases as imposter attacks. From Tables 6.1 and 6.2, we have seen an equal error rate of 4.1% and 3.1% with normal speech as both reference and testing using the $mfcc's$ and $lpcc's$, respectively as feature vectors. This error also corresponds to the false alarm rate (FAR) where an imposter is successful in breaking into the system. In [81], it was shown that it is possible to identify and imitate another speaker's voice and speech behavior by a professional impersonator. In [82] a professional impersonator was used to imitate a person to gain access to the system. Due to the difficulty of getting a speaker who can mimic another person, we do not consider the case of trained impersonation.

Playback data was collected by playing all the samples of the normal speech which was discussed in Chapter 4 by a playback device. This playback data is tested on the speaker verification system. The acceptance rate was seen to be 15%. Thus the playback attacks do result in the degradation of the performance of speaker verification system though it is not to a large extent. The acceptance of the playback speech depends on the quality of playback speech which in turn depends on the quality of the playback device and the nature of the environment. With a high quality playback device, imposter attacks

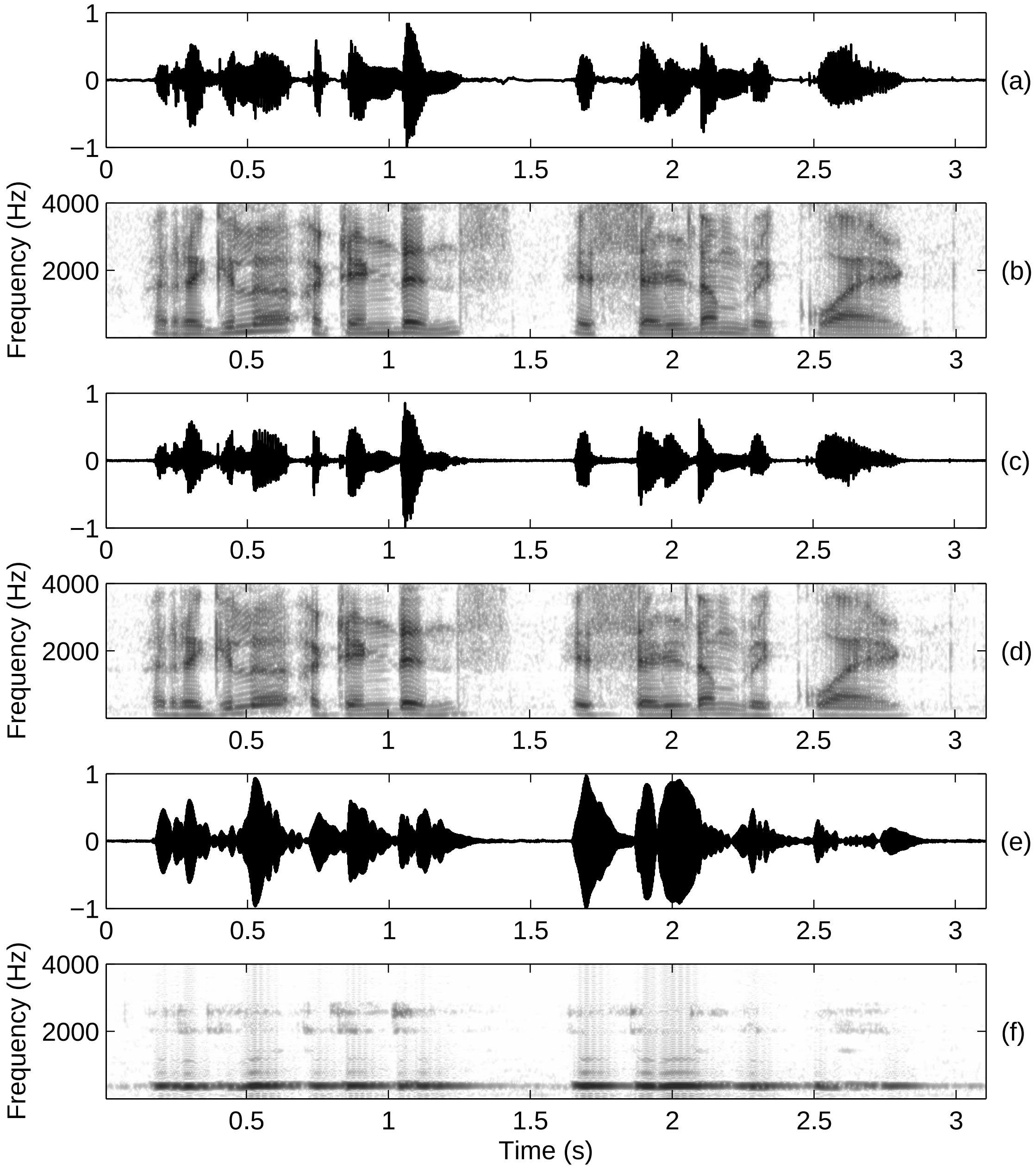

Fig. 6.8: (a) Speech signal, (b) spectrogram of the speech signal, (c) speech signal played back once, (d) spectrogram of the speech signal played back once, (e) speech signal repeatedly played back 15 times, and (f) spectrogram of the speech signal repeatedly played back 15 times.

can be more successful. Thus we require a method to handle imposters to increase the robustness of the speaker verification system.

6.5 Imposter detection using Lombard effect

Since Lombard effect changes the characteristics of the speech of a person, we can use this theory to detect imposters in speaker verification system. Firstly we assume that Lombard effect speech of the speaker is stored in the system during enrollment. Since this is a speaker verification scenario, the speaker is assumed to speak in private loop (self communication). We also make an assumption that the impersonator cannot mimic the Lombard effect speech of the person as he needs to further inherit the characteristics of Lombard effect speech of that real person as Lombard effect is speaker dependent in private loop. The speaker is asked to speak the text during verification. If the speaker is accepted by the system, the system then proceeds to imposter checking step. Here an artificial environment is created with noise and the speaker is asked to speak the same text. The system then checks for the change in features due to Lombard effect and accepts the speaker if the change in the features correspond to the speaker.

In case of playback attack we assure a better performance of the system due to two reasons:

- The recorded speech of the Lombard effect speech of the genuine speaker may not be available with the imposter.
- Even though the impostor has the recorded speech of the Lombard effect speech of the genuine speaker, since the performance of the system by verifying with playback speech is only 15%, the performance further degrades due to the second step.

After the imposter checking step, the MDR is bound to increase and that of FAR is bound to decrease. Also false acceptance cases are to be considered more threatful than false rejection cases in real time. Table 6.3 shows the MDR and FAR of the system after passing it through imposter checking step. We can see that the decrease in FAR is more

Table 6.3: MDR and FAR (in %) for speaker verification after imposter checking in case of MFCC and LPCC.

	MFCC	*LPCC*
MDR	5.1	4.6
FAR	0.3	0.2

than the increase in MDR. Thus the performance of the speaker verification has improved after the imposter detection step. Now for obtaining the performance of playback speech, Lombard effect speech is played back using the playback device and the performance is viewed. The final acceptance rate of the playback speech was found to be 3% after the testing step which was 15% initially. Thus the performance of the system is better after imposter testing.

6.6 Summary

In this chapter, we have presented the performance of a text-dependent speaker verification system and have proposed a method to detect imposters using Lombard effect. The feature vectors used in the system are linear prediction cepstral coefficients (LPCC) and mel-frequency cepstral coefficients (MFCC). LPCC were seen to perform better than MFCC. Performance of speaker verification is studied under both matched conditions and mismatched conditions of reference speech and test speech. The two conditions of speech considered are normal speech and Lombard effect speech. Lombard effect speech is seen to perform better compared to noemal speech under matched conditions which is due to less speaker variability of Lombard effect speech in an utterance. Finally a method is proposed to detect imposters during speaker verification using Lombard effect. Imposters are divided into two groups (a) Impersonation, (b) Playback attacks. The performance of he system due to imposter attacks is discussed. It is difficult to mimic the Lombard effect speech of a person and we assume that the imposter do not have the Lombard effect speech of the person. Thus by creating an artificial environment which causes Lombard effect, we can check for the characteristics of Lombard effect which is speaker-dependent and thus detect imposters.

Chapter 7

Summary and Conclusions

7.1 Summary of the work

The main contributions in this thesis centered around analyzing the Lombard effect speech to study the characteristics of Lombard effect mainly in terms of the excitation source features. We have described the excitation source features, (a) fundamental frequency (F_0), (b) strength of excitation (α), (c) perceived loudness (β), to study the characteristics of Lombard effect. Fundamental frequency and strength of excitation are obtained from the zero-frequency filter and loudness is obtained from the Hilbert envelope of the linear prediction (LP) residual. Studies have clearly shown that the impulse-like excitation is affected due to Lombard effect. Fundamental frequency (F_0) and perceived loudness (β) tend to increase and the strength of excitation (γ) tends to decrease due to Lombard effect. A significant increase in energy is found in a few regions like the vowel regions compared to other regions due to which the number of frames with low normalized energy are more and high normalized energy are less in the case of Lombard effect speech compared to normal speech.

The speaker-specific nature of the Lombard effect is illustrated using the distributions of the features. Dependence of Lombard effect on the type and intensity of feedback are studied. Relative energy in the case of nasals and fricatives was found to decrease. The change in duration of a sentence due to Lombard effect depends on the sound units in the sentence. A decrease in the rate of speech is observed when a person tries to maintain his normal speech in the presence of noise. The changes in the duration of an utterance were analyzed to differentiate Lombard effect speech and loud speech. The direction of change in the features for Lombard effect speech and loud speech are same. But this is not true in the case of duration. A given Lombard effect speech is defined with respect to the type and intensity of the external feedback, whereas loud speech is a controlled phenomenon.

Intelligibility of Lombard effect speech in noise is seen to be more than that of normal speech in noise. Intelligibility tests were performed at a sentence level. Lombard effect speech in factory noise is seen to be less intelligible compared to pink noise and babble noise. The mechanism of Lombard effect is studied using the analysis performed. Two cases should be considered to describe the mechanism of Lombard effect: private loop (self-communication), public loop (effective communication). In public loop, it is quite evident that the speaker increases his vocal effort based on the external feedback, whereas in private loop, the extent of Lombard effect depends on the speaker. A person cannot escape from the Lombard effect when his self feedback is hampered. He can only change the extent of Lombard effect. Lombard effect speech produced due to public loop is always intelligible while Lombard effect speech produced due to private loop need not be intelligible compared to normal speech.

The performance of text-dependent speaker verification in both matched conditions and mismatched conditions are noted. Mel-frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC) are taken as feature vectors. LPCC's were found to perform better than MFCC's. Performance was the system was better when both reference and test speech correspond to Lombard effect speech. Imposters are classified into two types: Impersonators and playback attacks. The degradation in the performance of the speaker verification system due to imposters in seen and a method to detect imposters is proposed using Lombard effect.

7.2 Major contributions of the work

The important contribution of the work reported in this thesis is the analysis of Lombard effect speech using the excitation source information. This is analyzed based on the change in the excitation source features due to Lombard effect. The major contributions of this thesis are:

- Analysis of Lombard effect speech using the excitation source features.
- Study on Lombard effect due to different types and intensities of external feedback.
- Studying the characteristics of speech produced without feedback.
- Differentiating Lombard effect speech and loud speech.
- Change in duration due to Lombard effect.
- Intelligibility of Lombard effect speech and normal speech in noisy conditions.
- Describing the mechanism of Lombard effect based on the analysis performed.
- Performance of speaker verification using Lombard effect speech as both reference and test speech.
- Method to detect imposters during speaker verification using Lombard effect.

7.3 Directions for future work

- Since the intelligibility of Lombard effect speech is seen to be high, synthesis of Lombard effect speech has significance in applications like spoken dialog systems where the listener expects to hear an intelligible speech which also depends on the environment he is in.
- Since it is not possible to maintain a silent environment always, the performance of speech and speaker recognition systems will be effected. The analysis done in this

work may help us to model the Lombard effect speech and build some compensation to transform it to normal speech.

- Lombard effect speech as both reference and test speech can be used in speaker verification system for a better performance.
- The method proposed to detect imposters is not perfect and needs to be generalized for a better performance to build a robust speaker verification system. Since speech produced varies with the changes in the type and level of external feedback, we need to model the change in speech due to change in the type and intensity of noise. This can be used further to improve the method.

References

[1] E .Lombard, “Le signe de l’elevation de la voix, annals maladiers oreille,” *Larynx, Nez, Pharynx*, vol. 37, pp. 101–119, 1911.

[2] C. Rivers and M. Rastatter, “The effects of multitalker and masker noise on fundamental frequency variability during spontaneous speech for children and adults,” *Journal of Auditory Research*, vol. 25, pp. 37–45, 1985.

[3] J. Fricke, “Syllabic duration and the Lombard effect,” *International Audiology*, vol. 19, pp. 53–57, 1970.

[4] G. Fairbanks, “Systematic research in experimental phonetics: 1. A theory of the speech mechanism as a servosystem,” *Journal of Speech and Hearing Disorders*, vol. 19, no. 2, pp. 133–139, 1954.

[5] I. Fonagy and J. Fonagy, “Sound pressure level and duration,” *Phonetica*, vol. 15, pp. 14–21, 1966.

[6] H. L. Lane and B. Tranel, “The lombard sign and the role of hearing in speech,” *Journal of Speech and Hearing Research*, vol. 14, pp. 677–709, 1971.

[7] J. C. Junqua, “The Lombard reflex and its role on human listeners and automatic speech recognizers,” *J. Acoust. Soc. Amer.*, vol. 93, no. 1, pp. 510–524, Jan. 1993.

[8] WV. Summers, DB. Pisoni, RH. Bernacki, RI. Pedlow and MA. Stokes, “Effects of noise on speech production: acoustic and perceptual analyses,” *J. Acoust. Soc. Amer.*, vol. 84, no. 3, pp. 917–929, 1988.

[9] J. Black, “Systematic research in experimental phonetics:2. Signal reception: Intelligibility and sidetone,” *Journal of Speech and Hearing Disorders*, vol. 19, no. 2, pp. 140–146, 1954.

[10] J. C. Junqua and Y. Anglade, “Acoustic and perceptual studies of Lombard speech: Application to isolated-words automatic speech recognition,” in *ICASSP*, vol. 2, Apr. 1990, pp. 841–844.

[11] T. Applebaum, B. Hanson and P. Morin, “Recognition strategies for Lombard speech,” *STL Research Reports*, vol. 5, pp. 69–75, 1996.

[12] AL. Winkworth and PJ. Davis, “Speech breathing and the lombard effect.” *Journal of Speech and Hearing Research*, vol. 40, pp. 159–169, Feb 1997.

[13] B. J. Stanton, L. H. Jamieson and G. D. Allen, "Acoustic-phonetic analysis of loud and Lombard speech in simulated cockpit conditions," in *ICASSP*, vol. 1, New York, Apr. 1988, pp. 331–334.

[14] J. C. Junqua, "The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex," *Speech Communication*, vol. 20, pp. 13–22, Nov. 1996.

[15] E. Vatikiotis-Bateson, V. Chung, K. Lutz, N. Mirante, J. Otten and J. Tan, "Auditory, but perhaps not visual, processing of Lombard speech," *J. Acoust. Soc. Amer.*, vol. 119, no. 5, p. 3444, May 2006.

[16] A. Castellanos, J. M. Benedi and F. Casacuberta, "An analysis of general acoustic-phonetic features for spanish speech produced with the lombard effect," *Speech Communication*, vol. 20, pp. 23–35, 1996.

[17] Sunhee Kim, "Durational characteristics of Korean Lombard speech," in *INTERSPEECH*, Lisbon, Portugal, Sept. 2005, pp. 2901–2904.

[18] Z. Bond, T. Moore and B. Gable, "Acoustic-phonetic characteristics of speech produced in noise and while wearing an oxygen mask," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 907–912, 1989.

[19] D. Pisoni, "Word identification in noise," *Language and Cognitive Processes*, vol. 11, no. 6, pp. 681–687, 1996.

[20] I. Pollack and J. Pickett, "Masking of speech by noise at high sound levels," *J. Acoust. Soc. Amer.*, vol. 30, no. 2, pp. 127–130, 1958.

[21] J. J. Dreher and J. J. O'Neill, "Effects of ambient noise on speaker intelligibility for words and phrases," *J. Acoust. Soc. Amer.*, vol. 29, pp. 1320–1323, 1957.

[22] H. Lane, "Foreign accent and speech distortion," *J. Acoust. Soc. Amer.*, vol. 35, no. 4, pp. 451–453, 1963.

[23] V. Hazan and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Communication*, vol. 24, pp. 211–226, 1998.

[24] J. M. Picket, "Effects of vocal force on the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 28, no. 5, pp. 902–905.

[25] D. Rostolland and C. Parant, "Distortion and intelligibility of shouted voice," in *Symposium: Speech Intelligibility*, Linoitalic, 1973, pp. 293–304.

[26] C. Haagen, "Intelligibility measurements," *Speech monographs*, vol. 12, no. 2, pp. 4–7, 1946.

[27] T. Hanley and M. Steer, "Effect of level of distracting noise upon speaking rate, duration and intensity," *Journal of Speech and Hearing Disorders*, vol. 14, no. 4, pp. 363–368, 1949.

[28] M. Munro, “The effects of noise on the intelligibility of foreign-accented speech ,” *Studies on Second Language Acquisition*, vol. 20, pp. 139–154, 1998.

[29] John. W. Black, “The effect of room characteristics upon vocal intensity and rate,” *J. Acoust. Soc. Amer.*, vol. 22, p. 174, 1950.

[30] S. Tonkinson, “The lombard effect in choral singing,” *Journal of Voice*, vol. 8, no. 1, pp. 24–29, Mar. 1994.

[31] CI. Johnson, HL. Pick, SR. Garber and GM. Siegel, “Intensity of guitar playing as a function of auditory feedback,” *J. Acoust. Soc. Amer.*, vol. 63, no. 6, p. 1930, June 1978.

[32] PM. Scheifele, S. Andrew, RA. Cooper, M. Darre, FE. Musiek and L. Max, “Indication of a Lombard vocal response in the St. Lawrence river Beluga,” *J. Acoust. Soc. Amer.*, vol. 117, no. 3, pp. 1486–1492, Mar. 2005.

[33] H. S. and M. Peet, “Ecology: Birds sing at a higher pitch in urban noise,” *Nature*, vol. 424, no. 6946, Jul. 2003.

[34] K. Manabe, EI. Sadr and RJ. Dooling, “Control of vocal intensity in budgerigars (Melopsittacus undulatus): differential reinforcement of vocal intensity and the Lombard effect,” *J. Acoust. Soc. Amer.*, vol. 103, no. 2, pp. 1190–1198, Feb. 1998.

[35] S. Nonaka, R. Takahashi, K. Enomoto, A. Katada and T. Unno, “Lombard reflex during PAG-induced vocalization in decerebrate cats,” *Neurosci.*, vol. 29, no. 4, pp. 283–289, Dec. 1997.

[36] B. Brumm, R. Schmidt and L. Schrader, “Noise-dependent vocal plasticity in domestic fowl,” *Animal Behaviour*, vol. 78, pp. 741–746, 2009.

[37] H. Brumm, K. Voss, I. Kllmer and D. Todt, “Acoustic communication in noise: regulation of call characteristics in a New World monkey,” *J. Exp. Biol.*, vol. 207, pp. 443–448, 2004.

[38] SE. Egnor, MD. Hauser, “Noise-induced vocal modulation in cotton-top tamarins (Saguinus oedipus),” *Am. J. Primatol.*, vol. 68, no. 12, pp. 1183–1190, Dec. 2006.

[39] LM. Potash, “Noise-induced changes in calls of the Japanese quail,” *Psychonomic Science*, vol. 26, pp. 252–254, 1972.

[40] H. Brumm, “Causes and consequences of song amplitude adjustment in a territorial bird: a case study in nightingales,” *An. Acad. Bras. Cienc.*, vol. 76, no. 2, pp. 1190–1198, June 2004.

[41] JM. Sinnott, WC. Stebbins and DB. Moody, “Regulation of voice amplitude by the monkey,” *J. Acoust. Soc. Amer.*, vol. 58, no. 2, pp. 412–414, Aug. 1975.

[42] J. Cynx, R. Lewis, B. Tavel B and H. Tse, “Amplitude regulation of vocalizations in noise by a songbird, Taeniopygia guttata,” *Anim Behav*, vol. 56, no. 1, pp. 107–113, July 1998.

[43] SR. Hage, U. Jurgens U and G. Ehret, "Audio-vocal interaction in the pontine brainstem during self-initiated vocalization in the squirrel monkey," *Eur. J. Neurosci.*, vol. 23, no. 12, p. 32973308, June 2006.

[44] B. J. Stanton, L. H. Jamieson and G.D. Allen, "Robust recognition of loud and Lombard speech in the fighter cockpit environment," in *ICASSP*, Glasgow, U.K., May 1989, pp. 675–678.

[45] H. Steeneken and J. H. L. Hansen, "Speech under stress conditions:overview of the effect on speech production and on system performance," in *ICASSP*, vol. 4, Mar. 1999, pp. 2079–2082.

[46] P. Rajasekaran, G Doddington and J. Picone, "Recognition of speech under stress and in noise," in *ICASSP*, 1986, pp. 733–736.

[47] Roman Goldenberg, Arnon Cohen and Ilan Shallom, "The Lombard effect's influence on automatic speaker verification systems and methods for its compensation," in *International conference on Information Technology, Research and Education*, 2006, pp. 233–237.

[48] A. Ikeno and J. H. L. Hansen, "Lombard Speech Impact on Perceptual Speaker Recognition," in *INTERSPEECH*, 2007, pp. 414–417.

[49] C. E. Mokbel and G. F. A. Chollet, "Automatic word recognition in cars," *IEEE Trans. Speech, Audio Process.*, vol. 3, no. 5, pp. 346–356, Sep. 1995.

[50] H. Boril, P. Fousek and P. Pollak, "Data-driven design of front-end filter bank for Lombard speech recognition," in *INTERSPEECH*, 2007, pp. 414–417.

[51] A. Wakao, K. Takeda and F. Itakura, "Variability of Lombard effects under different noise conditions," in *Proceedings of the International Conference on Spoken Language Processing*, vol. 4, Philadelphia, Pa, USA, Oct. 1996, pp. 2009–2012.

[52] Sang-Mun Chi and Yung-Hwan Oh, "Lombard effect compensation and noise suppression for noisy Lombard speech recognition," in *ICASSP*, vol. 4, Oct. 1996, pp. 2013–2016.

[53] J. H. L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation(MCE-ACC) for speech recognition in noise and Lombard effect," *IEEE Trans. Speech, Audio Process.*, vol. 2, no. 4, pp. 598–614, Oct. 1994.

[54] B. Hanson and T. Applebaum, "Robust speaker-independent word recognition using static, dynamic and acceleration fetaures: Experiments with lombard and noisy speech," in *ICASSP*, 1990, pp. 857 – 860.

[55] H. Boril, P. Fousek and H. Hoge, "Two-stage system for robust neutral/Lombard speech recognition," in *INTERSPEECH*, 2007, pp. 1074–1077.

[56] Fu Jie Huang and Tsuhan Chen, "Consideration of Lombard effect for speechreading," in *IEEE Fourth Workshop on Multimedia Signal Processing*, 1991, pp. 613 – 618.

[57] B. Yegnanarayana and K. Sri Rama Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 614–624, May 2009.

[58] K. Sri Rama Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[59] B. Yegnanarayana, K. Sri Rama Murty and S. Rajendran, "Analysis of stop consonants in indian languages using excitation source information in speech signal," in *ISCA-ITRW Workshop on Speech Analysis and Processing for Knowledge Discovery*, Aalborg university, Denmark, June 4-6 2008.

[60] K. Sri Rama Murty, B. Yegnanarayana and M. Anand Joseph, "Characterization of glottal activity from speech signals," *IEEE Signal Processing Letters*, vol. 16, no. 6, June 2009.

[61] Guruprasad Seshadri and B. Yegnanarayana, "Perceived loudness of speech based on the characteristics of excitation source," *J. Acoust. Soc. Amer.*, vol. 126, no. 4, pp. 2061–2071, Oct. 2009.

[62] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice Hall, 1975.

[63] J. H. L. Hansen and Vaishnevi Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, pp. 366–378, Feb. 2009.

[64] J. S. Garofalo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett and N. L. Dahlgrens, "The darpa timit acoustic-phonetic continuous speech corpus cdrom," in *Linguistic Data Consortium*, Philadelphia, PA, USA, 1993.

[65] *Noisex-92*, http://speech.cs.cmu.edu//comp.speech/Section1/Data/noisex.html.

[66] G. Bapineedu, B. Avinash, Suryakanth V. Gangashetty and B. Yegnanarayana, "Analysis of Lombard speech using excitation source information," in *INTERSPEECH*, Brighton, UK, Sept. 2009, pp. 1091–1094.

[67] B. S. Lee, "Effects of delayed speech feedback," *J. Acoust. Soc. Amer.*, vol. 22, no. 6, pp. 824–826, 1950.

[68] H. Pick Siegel, Jr., P. Fox, S. Garber and J. Kearney, "Inhibiting the lombard effect," *J. Acoust. Soc. Amer.*, vol. 85, no. 2, pp. 894–900, 1989.

[69] S. Kullback, *Information Theory and Statistics*. Mineola, New York: Dover Publications Inc., 1968.

[70] V. Varadarajan and J. H. L. Hansen, "Analysis of the Lombard effect under different types and levels of noise with application to in-set speaker ID systems," in *INTERSPEECH*, Sept. 2006, pp. 937–940.

[71] N. Dhananjaya, S. Rajendran and B. Yegnanarayana, “Features for automatic detection of voice bars in continuous speech,” in *INTERSPEECH*, Brisbane, Australia, Sept. 2008, pp. 22–26.

[72] G. Miller, G. Heise and W. Lichten, “The intelligibility of speech as a function of the context of the test materials,” *Journal of Experimrental Psychology*, vol. 41, no. 5, pp. 329–335, 1951.

[73] R. Patel and K. Schell, “The influence of linguistic content on the lombard effect,” *Journal of Speech and Hearing Research*, vol. 51, no. 1, pp. 209–220, Feb 2008.

[74] Steven B. Davis and Paul Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. on Acous., Speech, Signal Process.*

[75] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.

[76] J. Makhoul, “Linear prediction: A tutorial review,” *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr. 1975.

[77] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Prentice-Hall, Englewood Cliffs and N.J., 1993.

[78] J. D. Markel and A. H. Gray, *Linear Prediction of Speech.* New York, USA: Springer-Verlag, 1976.

[79] W. Shang and M. Stevenson, “A preliminary study of factors affecting the performance of a playback attack detector,” in *CCECE*, Niagara Falls, Canada, May 2008.

[80] ——, “A playback atack detector for speaker verification systems,” in *ISCCSP*, Malta, March 2008.

[81] E. Zetterholm, “Same speaker - different voices. A study of one impersonator and some of his different imitations,” in *Int. Conf. Speech Sci. & Tech.*, Auckland, New Zealand, Dec. 2006, pp. 70–75.

[82] E. Zetterholm, M. Blomberg and D. Elenius, “A comparison between human perception and a speaker verification system score of a voice imitation,” in *Tenth Australian Int. Conf. Speech Sci. & Tech.*, Sydney, Australia, Dec. 2004, pp. 393–397.

List of Publications

Refereed Journals

1. G. Bapineedu and B. Yegnanarayana, ‘Analysis of Lombard effect speech’, to be submitted to Speech Communication journal.

Conferences

1. G. Bapineedu, B. Avinash, S. V. Gangashetty, B. Yegnanarayana, ‘Analysis of Lombard speech using excitation source information’, in Proc. INTERSPEECH 2009, Brighton, UK, pp. 1091-1094.