ANALYSIS OF LAUGH SIGNALS FOR AUTOMATIC DETECTION AND SYNTHESIS

by

SUDHEER KUMAR K

200402023

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 Center

International Institute of Information Technology

Hyderabad, India

To My Parents

Poleswara Rao *and* **Lakshmi**

and My Guide

Prof. B. Yegnanarayana

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY

Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "**Analysis of Laugh Signals for Automatic Detection and Synthesis**" by **Sudheer Kumar K(200402023)**, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. B. Yegnanarayana

Acknowledgements

I would like to express my deepest respect and sincere gratitude to my guide Prof. B. Yegnanarayana, for his constant guidance and encouragement at all stages of my work. I am fortunate to have had technical discussions with him, from which I have benefited enormously. His motivational lectures in the monday meetings were undoubtedly my best source of inspiration. I thank him for the excellent research environment he has created for all of us to learn.

I am grateful to Dr. Rajendran, Mr. S.P.Kishore and Dr. Suryakanth for their encouragement and support all through my research work. I am extremely grateful to Dr. Rajendran for his technical discussions on phonetics, which gave me some insight as well as interest into the topic. I am thankful to Kishore for his attitude towards students. I also would like to convey my thanks to Dr. Suryakanth for providing excellent infrastructure in the labs.

I am very thankful to the senior lab members Anand, Anil bhai, Dhanu, Guru bhai and Murty for their valuable advices and suggestions in various aspects of my research work. I will forever remember the wonderful time I have had with my dearest friends and lab mates Amby (Harish), Bappi (Bapineedu), Boppay (Avinash) and Yella (Harsha). I thank my batchmates for all the wonderful moments we shared together. I thank all the past and present lab-mates for the friendly and conductive atmosphere in SVL.

Needless to mention the love and moral support of my parents, sister and brother-in-law. This work would not have been possible but for their support.

Sudheer Kumar Kovela

Abstract

Laughter is a nonverbal vocalization that occur often in continuous speech. It is produced by the speech production mechanism using a highly variable physiological process. The vocalized expression of laughter varies across gender, individuals and context. Despite its variability, laughter is perceived naturally by humans.

Since laughter is produced by the human speech production mechanism, spectral features were generally used for the study of laughter acoustics. This work mainly aims at showing the significance of excitation information for analysis of laugh signals. We proposed acoustic features which are motivated from the production characteristics of laughter. The features are based on pitch period (T_0), the strength of excitation derived from zero-frequency filtered signal, amount of breathiness and loudness measure.

It is observed that there will be sudden bursts of air flow through the vocal tract in the case of laughter. This will result in faster vibration of the vocal folds, and hence reduction in the pitch period. Apart from decrease in the pitch period, there is also a raising pattern in the pitch period contour. Similar observations are also made in case of strength of excitation. The strength of excitation rises sharply and then falls almost at the same rate. Since there will be more air flow, laughter is typically accompanied by some amount of breathiness. It will be reflected perceptually as less louder and more noisy. These loudness and non-determenistic component (noise) in the signal are estimated using measures based on Hilbert envelope. A method is proposed using these features for detecting laughter in continuous speech. The method is tested on a noisy data (TV show data) and a clean data (AMI Corpus) and the results are reported.

A method is also proposed for synthesizing isolated laughter by modifying the above

features. The perceptual significance of each of the features is studied using an analysis by synthesis based approach. The affect of each of the features are studied by deemphasizing them from some original laugh signals, and performing perceptual evaluation on the obtained signals. The study indicated that pitch period pattern is the most significant factor, followed by call-intercall duration, breathiness and strength of excitation.

Keywords: *Laughter, epoch, pitch period, strength of excitation, breathiness, loudness, zero-frequency resonator, synthesis.*

Contents

Abstract **iv**

List of Tables **ix**

1 Introduction **2**

1.1 Laughter terminology 2

1.2 Acoustics and Phonetics of Laughter 3

1.3 Types of Laughter 4

1.4 Thesis organization 4

2 Previous Studies on Laughter - A Review **6**

2.1 Studies on laughter analysis 6

2.2 Studies on automatic laughter detection 9

2.3 Studies on laughter synthesis 10

2.4 Issues with existing methods 11

2.5 Issues analyzed in the thesis 11

3 Source analysis of laughter **14**

3.1 Zero-frequency analysis 14

3.2 Acoustic features for analysis of laugh signals 19

3.2.1 Call level analysis 20

3.2.2 Bout level analysis 30

3.3 Distributions of the features for laughter and speech 32

3.4 Summary 34

4 Proposed method for detection of laughter in continuous speech 36

4.1 Block diagram and algorithm of proposed method 36

4.1.1 Pre-processing 37

4.1.2 Features 39

4.1.3 Decision Logic 40

4.1.4 Post-processing 43

4.2 Data 45

4.3 Results 45

4.4 Summary 46

5 Synthesis of laughter 48

5.1 Modeling the features 48

5.1.1 Pitch Period 49

5.1.2 Excitation Strength 50

5.1.3 Call and Inter-call Duration 52

5.1.4 Frication 52

5.2 Incorporating feature variation 53

5.2.1 Pitch period and Duration modification 53

5.2.2 Strength of Excitation modification 54

5.2.3 Frication Incorporation 54

5.3 Laughter Synthesis Procedure 55

5.4 Perceptual significance of features 57

5.5 Experiment for subjective evaluation 59

5.6 Evaluation Results 62

5.7 Summary 62

6 Discussions and Conclusion **64**

6.1 Contributions of the Work 65

6.2 Scope for future Work 65

References **67**

List of Publications **72**

List of Tables

3.1 Table showing the condition on slope values for grouping. 24

3.2 Frequency of occurrences of calls in each configuration. 24

4.1 Value and fraction thresholds for each feature. 43

4.2 Results of laughter spotting on AMI corpus and TV Show data. 46

5.1 Perceptual evaluation scores obtained for the modified versions of an original laugh signal. 58

5.2 Results of the experiment on perceptual significance of features. 59

5.3 Results of the laughter synthesis system. 61

List of Figures

1.1 A bout of a typical voiced laughter showing the bout, call and inter-calls. 3

3.1 Illustration of epoch extraction and their strengths from the zero-frequency filtered signal. (a) A segment of speech signal. (b) Filtered signal (c) DEGG signal. (d) Epoch locations. (e) Strength of excitation (α). (f) Fundamental frequency (F_0). 15

3.2 (a) A segment of speech signal. (b) zero-frequency filtered signal with window length of 3 ms for trend removal. (c) Voiced/non-voiced decision based on strength of excitation (α) values at the epoch locations. (d) Filtered signal obtained with adaptive window length for trend removal. (e) Strength of excitation (α). (f) Pitch period obtained from epoch locations. 17

3.3 (a) A segment of speech signal. (b) Zero-frequency filtered signal obtained with adaptive window length for trend removal. (c) Filtered signal with window length for trend removal obtained from the complete signal. (d) Pitch period obtained from modified epoch extraction. (e) Pitch period obtained from original epoch extraction. 19

3.4 A bout of a typical voiced laughter is showed in (b) and its spectrogram in (a). 20

3.5 Laughter calls of different laughs with the corresponding pitch period contours below. 22

3.6 Laughter calls of different laughs with the corresponding pitch period (T_0) contours and pitch slope (p') contours. 25

3.7 (a) A voiced laughter with epoch locations. (b) Pitch period contour derived from epoch locations. (c) Strength of excitation derived from the zero-frequency filtered signal. (d) Ratio of strength of excitation to pitch period. (e) Normalized slope of pitch period. (f) Normalized slope of strength of excitation. 26

3.8 Illustration of effect of breathiness on LP residual and hilbert envelope. Speech signal, LP residual and Hilbert envelope for (a) modal vowel /a/ and (b) breathy vowel /a/. 28

3.9 (a) Speech signal containing both laughter and speech, (b) Hilbert envelope, (c) η, (d) γ, (e) η/γ. 29

3.10 Pattern of call and inter-call durations for different call numbers. 30

3.11 Distribution of duration of calls of laughter. 31

3.12 Illustration of differences in the durations of voiced regions (calls in case of laughter) and unvoiced regions (intercall interval in case of laughter) between laughter and normal speech using the distributions of the differenced call and intercall durations. (a) differenced call durations (b) differenced inter-call durations. 32

3.13 Illustration of differences in the excitation source features of Laughter and Normal speech using frequency distributions (first column). It also shows the variations of features across the calls within a laughter bout using the distributions of first and last calls (second and third columns). In each plot the distributions of corresponding features of laughter and normal speech are represented with solid and dashed lines respectively. 33

3.14 Illustration of differences in the excitation source features of Laughter and Normal speech using frequency distributions (first column). It also shows the variations of features across the calls within a laughter bout using the distributions of first and last calls (second and third columns). In each plot the distributions of corresponding features of laughter and normal speech are represented with solid and dashed lines respectively. 34

4.1 (a) Voiced laugh signal containing glottal frication with voiced non-voiced decision (b) zero-frequency filtered (zff) signal with 4ms trend removal window, (c) energy of the strength of excitation contour derived from the slope of the zff signal. 38

4.2 (a) Voiced laugh signal containing glottal frication with corrected voiced non-voiced decision, (b) correlation values (c), (c) lag values (l), (d) $c - \hat{l}$. 39

4.3 Illustration of estimation of the value and fraction thresholds for all the features. (a) T_0, (b) α, (c) β, (d) δT_0, (e) $\delta\alpha$, (f) η and (g) γ. First column, second column and third column shows the values of dr, far and ($dr - far$) respectively. 42

4.4 Block diagram of proposed laughter spotting algorithm. 44

5.1 (a) A segment of laughter signal. (b) Pitch period derived from the epoch locations. (c) Strength of excitation at the epochs. 49

5.2 Illustration on comparison of original and modeled pitch period contours. Two laugh calls are shown in (a) and (b) with their corresponding pitch period contours in (c) and (d), respectively. In (c) and (d), the actual pitch period contour is showed in thin dots (.) and modeled one is shown in thick dots (*) 51

5.3 Scatter plot demonstrating the correlation between impulse strength and strength of excitation obtained from zero-frequency filtered signal. . . . 55

5.4 Block diagram of the synthesis system. 57

Chapter 1

Introduction

The phenomenon of laughter is common in human communication as a way of expressing the emotion of happiness. It is produced by the speech production mechanism using a highly variable physiological process. The vocalized expression of laughter varies across gender, individuals and context. Despite its variability, laughter is perceived naturally by human listeners. In recent years much of the research done in the area of speech recognition has been mainly concentrated on natural data. This requires data collected in natural environment which contain many non-speech elements like laughter and other non-linguistic sounds. Automatic detection of such elements helps in increasing the accuracy of recognition. There are also theories on laughter which say that we not only laugh at humor but also at surprise, when embarrassed etc. So, spotting laughter also helps us to know the possible emotional states of the speaker which makes us easy to converse with them.

1.1 Laughter terminology

Laughs were analyzed at three levels: bout, call and segment levels [1]. The entire laugh is referred to as an episode which consists of laughter bouts that are produced during one exhalation. Calls are the discrete acoustic events that constitute a bout, and each call of a voiced laughter consists of a voiced part followed by an unvoiced glottal fricative/silence

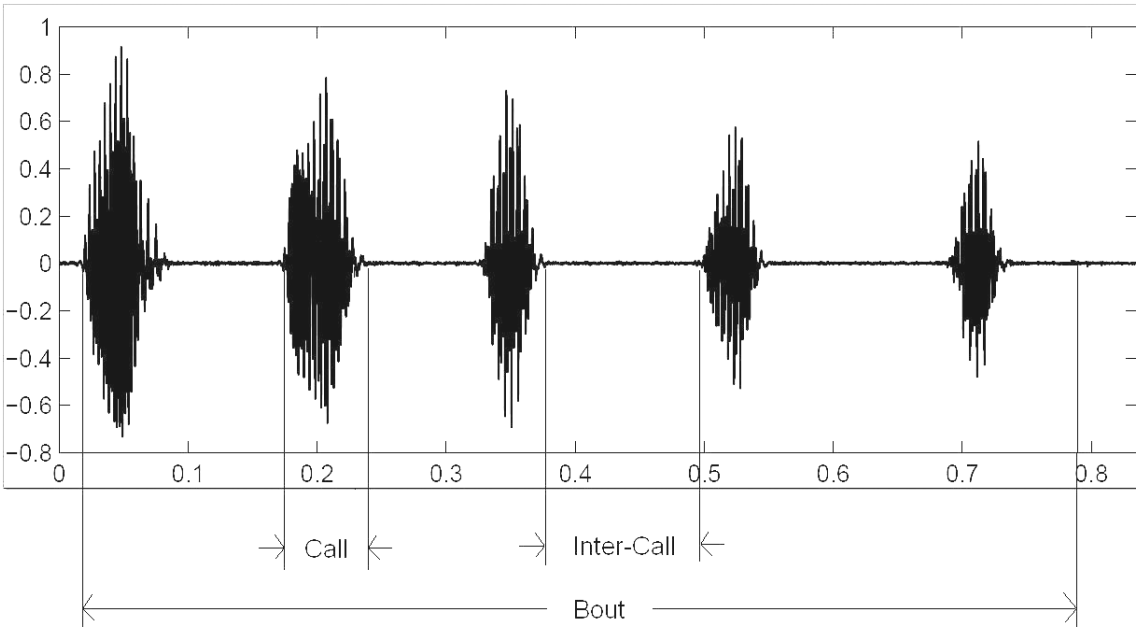

Fig. 1.1: A bout of a typical voiced laughter showing the bout, call and inter-calls.

part. Segments are the audibly reflected changes in the production within a call. It is assumed that each laughter bout contains several calls, so that isolated calls are not considered as laughter.

Fig. 1.1 shows the bout, call and inter-call of a voiced laughter.

1.2 Acoustics and Phonetics of Laughter

Laughter is not a very controlled natural non-verbal vocalization. While laughter has very distinct perceivable pattern, its production is not guided by any rules of grammar as in the case of speech. It is typically produced by series of sudden bursts (outflows) of air, keeping the vocal tract in a neutral position. The main difference between speech and laughter is that normal speech does not disrupt breath whereas laughter may [2]. In normal speech most of the role is played by the articulators of the vocal tract system, whereas in case of laughter major role is played by the lungs and the vocal folds (source). Laughter can be broadly divided into two types based on its glottal activity: a) voiced laughter and

b) unvoiced laughter. In a voiced laughter the air bursts flows mostly through the mouth, and in some cases it may even pass through the nose. There will be more air pressure build up in lungs, as a result of which the vocal folds may vibrate in a different way. Also, since there is more air flow, there will be turbulence generated within the vocal folds, and hence the signal may be breathy when compared to normal speech. Unvoiced laughter may also have the same overall structure as that of a voiced laughter, but the major difference is that there is no voicing. The durations of the calls will be less, and there will be more damping when compared to voiced laughter.

1.3 Types of Laughter

There were many attempts done on categorizing laughter from perspective of production mechanism. The primary division made by Grammer and Eibl-Eibesfeldt [3] based on the glottal activity as voiced and non-voiced laughter. Bachorowski and his group [4] gave a more clearer division based on the broad characteristics of bout. Bouts comprising of voiced calls were called as 'song-like', which are also commonly called as giggle or chuckles. They further divided unvoiced laughter into two types, 'snort-like' having perceptually salient nasal-cavity turbulence and 'unvoiced grunt-like' which are acoustically noisy produced with turbulence arising in either the laryngeal or oral cavity.

1.4 Thesis organization

The contents of the thesis are organized as follows:

Chapter 2 gives an overview of the existing approaches for laughter detection and laughter synthesis. The drawbacks of the existing approaches and hence the need for alternate approaches are discussed in the chapter.

In chapter 3, we highlight the significance of excitation characteristics for the analysis of laugh signals. We proposed modified version of an already existing epoch extraction method [5] for capturing the rapid variations of F_0 in laughter. We also proposed features

based on excitation information for analyzing laugh signals. Using the frequency distributions we showed that these features can indeed be used to discriminate laughter from speech.

In chapter 4, we proposed a method for spotting laughter in spontaneous speech using the features described in chapter 3. The performance of the proposed method is evaluated and compared with the previous works.

In chapter 5, we showed the significance of the proposed features by conducting perceptual evaluation tests on the synthesized laughter. Features were modeled and their modification was done using the methods explained. Isolated laughter was synthesized by incorporating the feature variations.

In chapter 6, we summarize the contributions of the present work, and discuss some issues that are still to be addressed.

Chapter 2

Previous Studies on Laughter - A Review

Although, it has been more than 100 years since attempts were being made to understand the psychological theory behind laughter, the studies on analysis of laugh signals were started no sooner than 10 years from now, due to the unavailability of processing tools and techniques. Previous studies on laughter can be categorized into three a) on laughter analysis b) automatic laughter spotting and c) laughter synthesis.

2.1 Studies on laughter analysis

Laughter has been studied in different perspectives like studies in philosophy of laughter, acoustics and phonetics of laughter, speech pathology, time and spectral domain analysis etc.

Analysis of laughter has started with understanding the philosophy behind it. The studies on philosophy of laughter try to explain why and when humans laugh. If we want to synthesize laughter in spontaneous speech, we have to first determine where it should come and hence there is a need for understanding the philosophy behind it. There were many theories which tried to explain laughter completely but none of them were successful in doing so. The oldest, and probably still the most widespread theory of laughter

is ‘superiority theory’ or ‘Hobbesian theory’, which says that laughter is an expression of a person’s feelings of superiority over others. This theory goes back at least as far as Plato and Aristotle [6] [7]. Hobbes gave a revised version of it laughter expresses ”a sudden glory arising from some conception of some eminency in ourselves, by comparison with the infirmity of others, or with our own formerly” [8]. In our century many have adopted versions of the superiority theory. Albert Rapp, for example, claims that all laughter developed from one primitive behavior in early man, ”the roar of triumph in an ancient jungle duel” [9]. Konrad Lorenz and others treat laughter as a controlled form of aggression []. Later some [10] have responded to the ‘superiority theory’ by denying the reality of hostile and derisive laughter. The second popular theory ‘incongruity theory’, given by Kant and Schopenhauer. The theory as defined by Kant was ”Laughter is an affection arising from the sudden transformation of a strained expectation into nothing” [11] and as defined by Schopenhauer as ”laughter is a mismatch between our concepts and the real things that are supposed to be instantiations of these concepts” [12]. Like the ‘superiority theory’, the ‘incongruity theory’ is also not comprehensive enough to cover all the non-humorous cases. The laugh of the 5-month-old baby at being tickled involves no incongruity as some examples. The next most popular theory was the ‘relief theory’ proposed by Herber Spencer. It says that ”laughter occurs when some emotion has built up but then is suddenly seen to be inappropriate” [13]. Spencer’s theory influenced many subsequent theorists of laughter, including Dewey and Freud [14] [15]. Freud’s theory says that ”in all laughter situations we save a certain quantity of psychic energy, energy that is usually employed for some psychic purpose but which turns out not to be needed and hence discharges it in the form of laughter” [15]. Apart from some extra cases which can be explained by ‘relief theory’, it is also proved to be not very general. There is a one more theory given by John Morreall which says that ”laughter is rather the physical activity which is caused by, which expresses, the feeling produced by a pleasant psychological shift” [16]. The author also says that the theory is the most generalized theory given till now.

There are also studies done in understanding the dynamics of the lungs, vocal tract, vocal folds, larynx, tongue etc. while laughing [17] [18] [19] [20] [21]. Filippelli, in his work [17] has analyzed the volume and pressure drops in lungs between every call

and between bout. He concluded that there is sudden and substantial decrease in the lung volume after every call. John Esling studied the possible states of larynx in laughter. He also described the states of larynx for different types of laughter.

Laughter was studied for its practical importance in speech pathology [18] [19] [20]. Citardi et. al analyzed laughter of people suffering from hyperadductive dysphonias (the patient has not much control on the vocal folds). They found that this is not the case in laughter and hence concluded that laughter can be used to correct these kinds of problems.

There are many studies on understanding the characteristics of laughter [22], [23], [4], [24], [1]. The authors differed in the extent to which they characterized laughter, with some claiming laughter has very specific attributes while others emphasized that laughter is a variable signal [4]. The differences in how specific the characterization of laughter was could be due to the vastly different number of subjects and laugh bouts studied in each experiment. For example, there were 2 subjects and 15 laugh bouts analyzed in [23], 51 laugh bouts investigated in [24], and 97 subjects and 1024 laugh bouts analyzed in [4]. Not surprisingly, due to the small sample size of some of the studies, the conclusions of these works varied and sometimes contradicted one another. Many agreed that laughter has a repetitive breathy consonant-vowel structure (i.e. ha-ha-ha or ho-ho-ho) [23], [24], [1]]. One work went further and concluded that laughter is usually a series of short syllables repeated approximately every 210 ms [24]. Yet, others found laughter to be highly variable [4] [1] particularly due to the numerous bout types (i.e. voiced song-like, unvoiced snort-like, unvoiced grunt-like, etc.) [4], and thus difficult to stereotype. These conclusions led us to believe that automatic laughter detection is not a simple task.

Since laughter is produced by the human speech production mechanism, the laughter signal is also analyzed like a speech signal in terms of the acoustic features of the speech production. Analysis of laughter could be done for synthesis, where perceptually important characteristics need to be preserved, or for studying the acoustic features during its production. Based on analysis of large database of laughter sounds, Bachorowski and colleagues have differentiated three broad categories, namely, song-like (consisting primarily of voiced sounds), snort-like (consisting largely unvoiced calls with perceptually salient nasal-cavity turbulence) and grunt-like (with turbulence from laryngeal or

oral cavities) [4]. Typically, the acoustic analysis of laughter is carried out using duration (between onset and offset of acoustic events), F_0 (fundamental frequency of voiced excitation) and spectral features. All of these are used to describe the temporal variability, source variability and variability in production modes [4]. Formants, pitch and voice quality analysis are used to discriminate speech, speech-laughs and laugh [25]. Some of these features were studied using conventional methods of analysis for F_0 and voiced quality, but using mostly spectrum-based features, like harmonics, spectral tilt and formants [25]. The difficulty in deriving the features of excitation source using short-time spectral analysis limits the analysis significantly, especially due to the choice of the size, shape and position of the segment in relation to the acoustic events in speech production. The main problem is to extract the rapidly varying instantaneous F_0. Moreover, traditional short-time spectrum analysis masks several important subsegmental (less than pitch period) features of the glottal source that are unique in the production of laughter.

2.2 Studies on automatic laughter detection

Earlier work pertaining to automatic laughter detection focused on identifying whether a predetermined segment contained laughter using various machine learning methods including Hidden Markov Models (HMMs) [26], Gaussian Mixture Models (GMMs) [27], and Support Vector Machines (SVMs) [28][29]. Note that the objectives of these studies differed as described below.

Cai et al. used HMMs trained with Mel Frequency Cepstral Coefficients (MFCCs) and perceptual features to model three sound effects: laughter, applause, and cheer. They used data from TV shows and sports broadcasts to classify 1 second windows overlapped by 0.5 seconds. They utilized the log-likelihoods to determine which classes the segments belonged to and achieved a 92.7% recall rate and an 87.9% precision rate for laughter [26].

Truong and van Leeuwen classified presegmented ICSI Meeting Recorder data as laughter or speech. The segments were determined prior to training and testing and had variable time durations. The average duration of laughter and speech segments were 2.21 and 2.02 seconds, respectively. They used GMMs trained with perceptual linear predic-

tion (PLP), pitch and energy, pitch and voicing, and modulation spectrum features. They built models for each type of feature. The model trained with PLP features performed the best at 7.1% EER for an equal-prior test set [27]. The EER in Truong and van Leeuwen.s presegmented work was computed on the segment level, where each segment (which had variable duration) was weighted equally. Note that this is the only system that scored EER on the segment level. All other systems reported EER on the frame level, where each frame was weighted equally.

Kennedy and Ellis studied the detection of overlapped (multiple speaker) laughter in the ICSI Meetings domain. They split the data into non-overlapping one second segments, which were then classified based on whether or not multiple speakers laughed. They used SVMs trained on statistical features (usually mean and variance) of the following: MFCCs, delta MFCCs, modulation spectrum, and spatial cues. They achieved a true positive rate of 87%[28].

More recently, automatic laughter recognition systems improved upon the previous systems by detecting laughter with higher precision as well as identifying the start and end times of the segments. In particular, Truong and van Leeuwen utilized GMMs trained on PLP features with a Viterbi decoder to segment laughter. They achieved an 8% EER, where each frame was weighted equally, on an equal-prior test set [30].

2.3 Studies on laughter synthesis

Recently, there were also attempts done to synthesize laughter [31], [32], [33], [34], [35]. Though the number is very small, they are all completely different. Laughter synthesis was handled at the bout level in [31]. Laughter is modeled at two levels, the overall episode level and at the local call level. At episode level, the authors attempts to capture the overall temporal behavior of laughter with a parametric model based on the simple harmonic motion of a mass-spring system. At the call level, they relied on a standard linear prediction based analysis-synthesis model. [33] concentrates on synthesizing speech-laughs i.e laughter occurring with spontaneous speech. Different types of synthesized laughter were combined with speech in a dialogical situation and then perceptually evalu-

ated. They reported that intensity of the laughter also plays a crucial role in the perception. Lasarcyk et. al [34] tried to synthesize isolated laughter using articulatory synthesis and diphone synthesis. They modeled voiced/non-voiced durations, vowel quality, pitch etc. for synthesizing laughter. Trouvain et. al in one of their later studies [34] modeled laughter by the parameters spread of the lips, raise in the larynx and fundamental frequency. They explored the relative contributions of each of the features by conducting a perceptual evaluation test.

2.4 Issues with existing methods

While most of the previous studies on analysis was done on understanding prosody (bout and episode level) of laughter, very less number concentrated on intra-call level variations. For the same reason, the studies resulted in automatic spotting of laughter also relied mostly on the structure (correlation between successive calls etc.) of the bout in time and spectral domains. So, the less the number of calls, the more difficult the spotting would be for such methods.

2.5 Issues analyzed in the thesis

The variability in laughter production is complex in the sense that it is not guided by the production rules of speech. Hence it is difficult to describe the phenomenon of laughter precisely, although it can be perceived by the listeners. The analysis and description is also limited by the available tools for analysis of laughter signals. The objective of this study is to show that some important features of laughter acoustics can be highlighted using some new tools for analysis proposed in this paper. It is likely that these new features may help to spot the laughter regions in continuous speech communication. Some of these features are:

1. Rapid changes in the instantaneous fundamental frequency (F_0) within calls of a laughter bout.

2. Strength of excitation within each glottal cycle and its relation to F_0.

3. Loudness of speech derived from the excitation information.

4. Breathiness in signal extracted from the hilbert envelope.

5. Temporal variability of F_0, strength and formants across calls within a bout.

Chapter 3

Source analysis of laughter

In this chapter, we present the study on analysis of laugh signals. The analysis is mostly focused on the variations of acoustic features related to excitation source, extracted in a pitch synchronous manner. The features for the analysis are motivated from the production mechanism of laughter. They are based on pitch period (T_0), strength of excitation and amount of breathiness.

In Section 3.1, we discuss about the zero-frequency based epoch extraction algorithm which is a basis for the further analysis. In Section 3.2, we analyze laugh signals at call and bout level using the proposed features. Section 3.3 discusses the variations of these features across calls within a bout, and between laughter and speech using distributions.

3.1 Zero-frequency analysis

Most of the analysis on laugh signals in this work is done in a pitch synchronous manner. The discontinuities associated with windowing can be reduced if the analysis is carried out pitch synchronously. But finding the exact location of the epochs is a very difficult task, especially in the case of laughter where there are large variations in the vocal fold vibration, while at the same time we need to have a epoch detection algorithm with a good accuracy for doing such an analysis. Recently a new method is proposed for extraction of the instantaneous F_0 [36], for epoch extraction [5] and for strength of excitation at

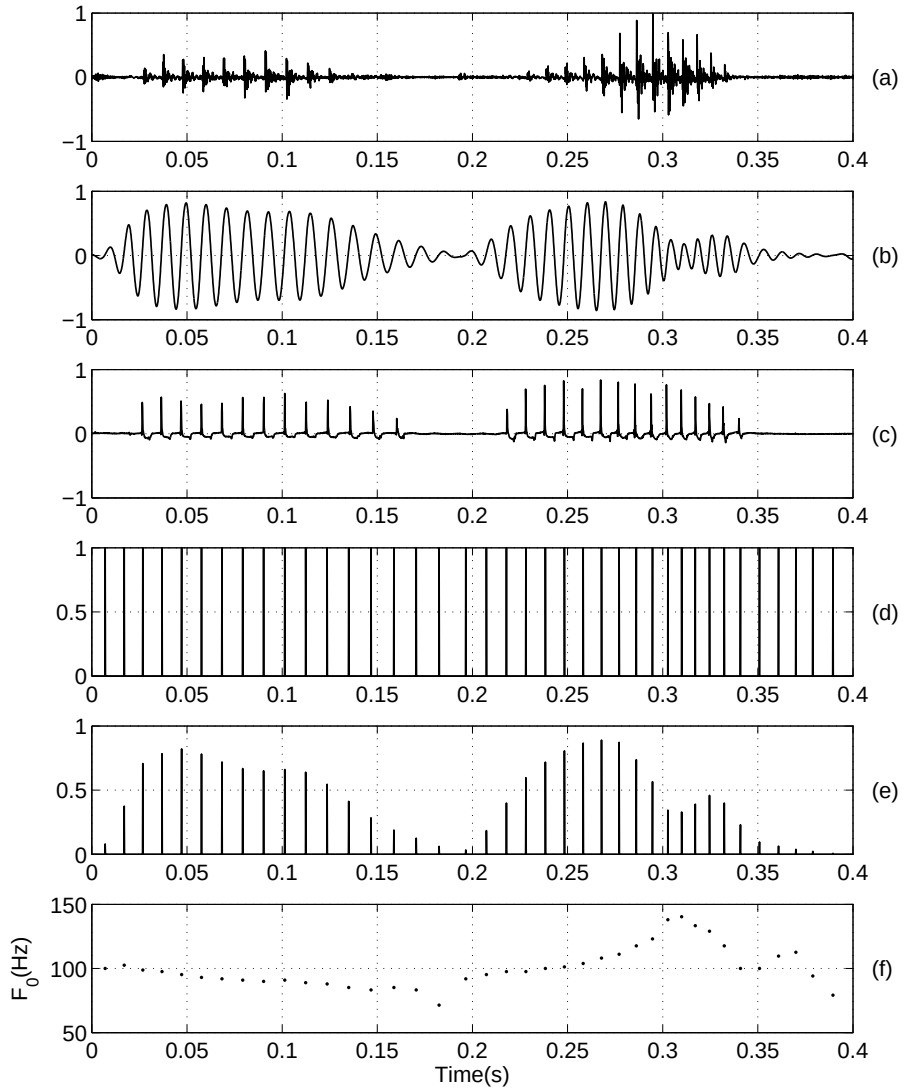

Fig. 3.1: Illustration of epoch extraction and their strengths from the zero-frequency filtered signal. (a) A segment of speech signal. (b) Filtered signal (c) DEGG signal. (d) Epoch locations. (e) Strength of excitation (α). (f) Fundamental frequency (F_0).

epochs [37]. The method uses the zero-frequency filtered signal derived from speech to obtain the epochs (instants of significant excitation of the vocal tract system) and the strength at the epochs. The following steps are involved in processing the speech signal to derive the epochs and their strengths from the filtered signals [37].

1. Difference the speech signal $s[n]$ to remove any very low frequency 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 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 $a_4 = 1$.

3. Compute the average pitch period using the autocorrelation function for every 30 ms speech segments.

4. Remove the trend in $y_o[n]$ by subtracting the local mean computed over a window obtained from (c) at each sample. The resulting signal $y[n]$ is the zero-frequency filtered signal, given by

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

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.

5. The instants of positive zero crossings of the filtered signal give the locations of the epochs.

6. The strength of the epoch (denoted as α) is obtained by taking the slope of the filtered signal around the epoch. The slope is measured by taking the difference between the positive and negative sample values around each epoch.

Fig. 3.1 illustrates extraction of epochs and their strengths from the filtered signal derived from the speech signal. The strength values are compared with the amplitudes of the peaks around the epochs in the differenced electro GlottoGraph (DEGG). While this method works well for the variations of F_0 in normal speech signals, it cannot capture the rapid changes of F_0 that occur in the calls of a laughter episode or cycle.

The critical factor in the above method is the choice of the window for trend removal from the output of the zero-frequency resonator. If the window size is too small compared to the average pitch period, then too many zero crossings occur in the filtered signal. If it

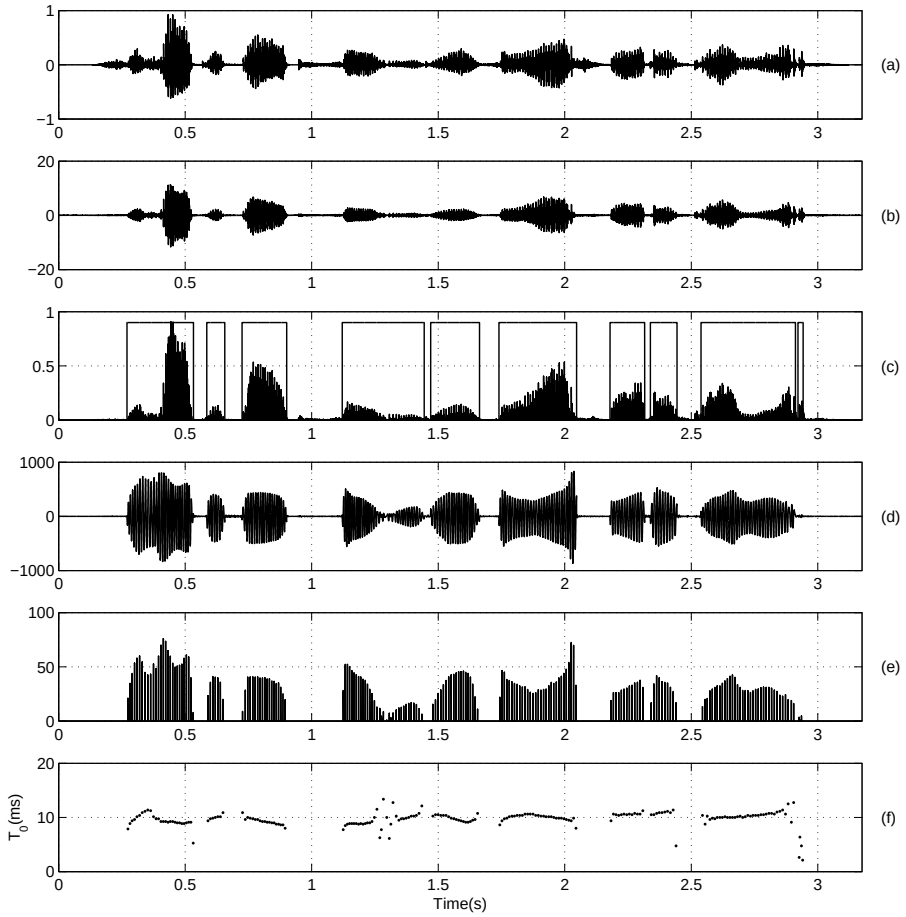

Fig. 3.2: (a) A segment of speech signal. (b) zero-frequency filtered signal with window length of 3 ms for trend removal. (c) Voiced/non-voiced decision based on strength of excitation (α) values at the epoch locations. (d) Filtered signal obtained with adaptive window length for trend removal. (e) Strength of excitation (α). (f) Pitch period obtained from epoch locations.

is too large, then the short pitch periods corresponding to high F_0 may be missed. In order to capture the rapid variations in F_0 between speech and laughter the following procedure is adopted:

1. Pass the signal through the zero-frequency resonator with window length of 3 ms for trend removal. This window length has been chosen in such a way that it gives high energy in the filtered signal in case of speech and laughter and low energy in the non-voiced and silence regions.

2. Positive zero crossings of the filtered signal gives the epoch locations, and the slope calculated as the difference of values of the samples after and before the epochs gives the strength of excitation. Mean of the strength of excitation over a window

of 10 ms is calculated, and if this value is more that 30 percent of the maximum strength value of the complete signal then that segment is considered as a voiced segment, otherwise it is a non-voiced segment.

3. After finding the voiced segments, each voiced region is separately passed through a zero-frequency resonator with window length for trend removal derived from that segment. This is done by first computing the autocorrelation of the segment with a frame size of 20 ms and a frame shift of 10 ms. Then the maximum occurring peak in the autocorrelated signals is chosen as the window length for that region.

4. The positive zero crossings of the final filtered signal give the epoch locations, and the difference in the values of the samples after and before each epoch gives the strength of excitation.

Fig. 3.2 illustrates the epochs and strength of excitation for a segment of speech signal using the modified epoch extraction method.

Fig. 3.3 illustrates the improvement of epoch extraction using the modified epoch extraction method. Because of the wide range of pitch periods in the signal, the same window length cannot be used for all the segments which can be clearly reflected in the extracted pitch periods.

Some cases where there may be improvement due to the modified epoch extraction method are

1. when there are multiple speakers speaking in turns.

2. when there is laughter in the signal.

3. when there are large pitch period variations in the signal caused by the text or emotional state of speaker.

There are also some drawbacks in the above method. The estimation of window size for trend removal may go wrong especially when the segments are very small (less than 5 pitch cycles). In such a case, we may use the information of the adjacent voiced segments.

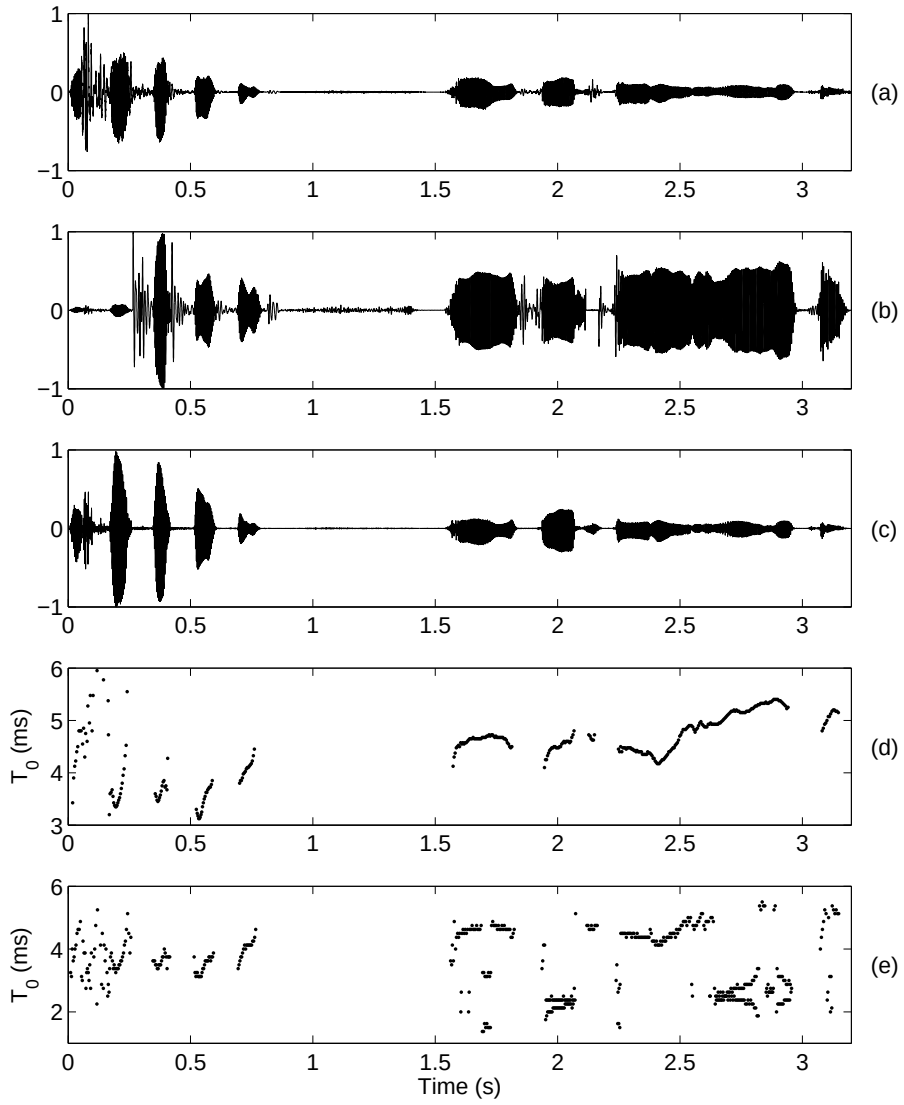

Fig. 3.3: (a) A segment of speech signal. (b) Zero-frequency filtered signal obtained with adaptive window length for trend removal. (c) Filtered signal with window length for trend removal obtained from the complete signal. (d) Pitch period obtained from modified epoch extraction. (e) Pitch period obtained from original epoch extraction.

3.2 Acoustic features for analysis of laugh signals

We keep the vocal tract in the neutral position while laughing. Also during this process we don't vary our articulators much. Though the above two characteristics are common to speech vowels, we perceive laughter very different from normal speech. So, we hypothesize that most of the laughter characteristics occur due to variations in the excitation.

A typical voiced laughter is shown in the Fig. 3.4. A voiced laughter bout typically

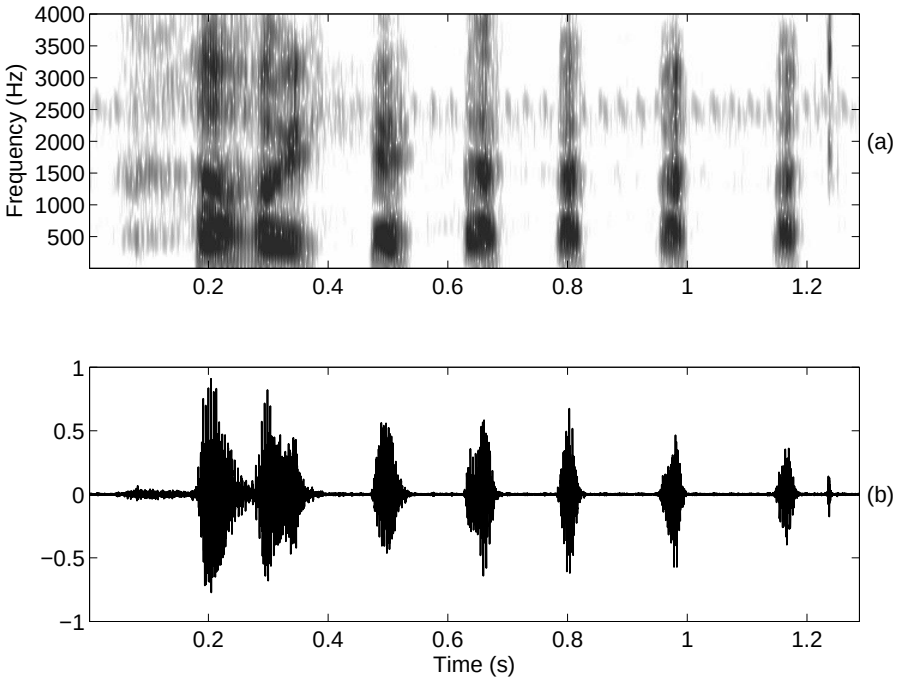

Fig. 3.4: A bout of a typical voiced laughter is showed in (b) and its spectrogram in (a).

contains about 4 to 6 calls, with each call duration ranging from 10 ms to 200 ms, and some amount of silence/frication between two consecutive calls. The energy of the calls may fall as the laughter progresses. The production mechanism of laughter is different from speech in many ways starting from the amount of air flow through the vocal tract to the shape of the vocal tract. This fact increased our interest to analyze the source features of laughter signals more closely.

The laughter signals are analyzed at two levels (a) call level and (b) bout level. Based on this kind of analysis, the features can be categorized into two groups call level features and bout level features. The call level features are used to capture the call level patterns (variations within a call) of laughter where as bout level features for capturing the high level repetitive structure (patterns between calls) of laughter.

3.2.1 Call level analysis

The source and system characteristics of laugh signals at call level are analyzed using features like pitch period (T_0), strength of excitation (α), amount of breathiness, call du-

rations and some parameters derived from them which are explained below in detail.

Pitch period (T_0)

It is observed that pitch frequency for laughter is more than that for normal speech [4]. For normal speech the pitch frequency typically ranges between 80 Hz and 200 Hz for male speakers and 200 Hz to 400 Hz for female speakers, whereas for laughter the mean pitch frequency for males is above 250Hz, and for females it is above 400 Hz [4]. As mentioned earlier, there will be more air flow through the vocal tract in the case of laughter. This will result in faster vibration of the vocal folds, and hence reduction in the pitch period. Apart from lower pitch period, there is also a raising pattern in the pitch period contour within a call. In some cases the pitch period may even start with some large value, decreases to some minimum and then increases again. This may be because this high pitch frequency (F_0) is not normal for the vibration of the vocal folds to maintain that frequency, and hence it tends to decrease. The general pattern that is observed in the pitch period within a call is that it starts with some value, decreases slightly to some minimum, and then increases rapidly to a high value. Fig. 3.5 shows this general trend of T_0 within a call for 6 different calls. We can clearly observe the pattern described above. As a result of this faster vibration of vocal folds, there will be more coupling between the source and system about which very less is known.

The main issue here is extracting the pitch period accurately. It is not easy as in case of normal speech, since the pitch variation is large in laughter and also there is a large difference between pitch values of speech and laughter. The epoch locations and the pitch period are extracted as explained in Section 3.1.

Fig. 3.5 shows the pitch period contours of different laughter calls. We can see that the above described pitch period pattern is being followed by all of them.

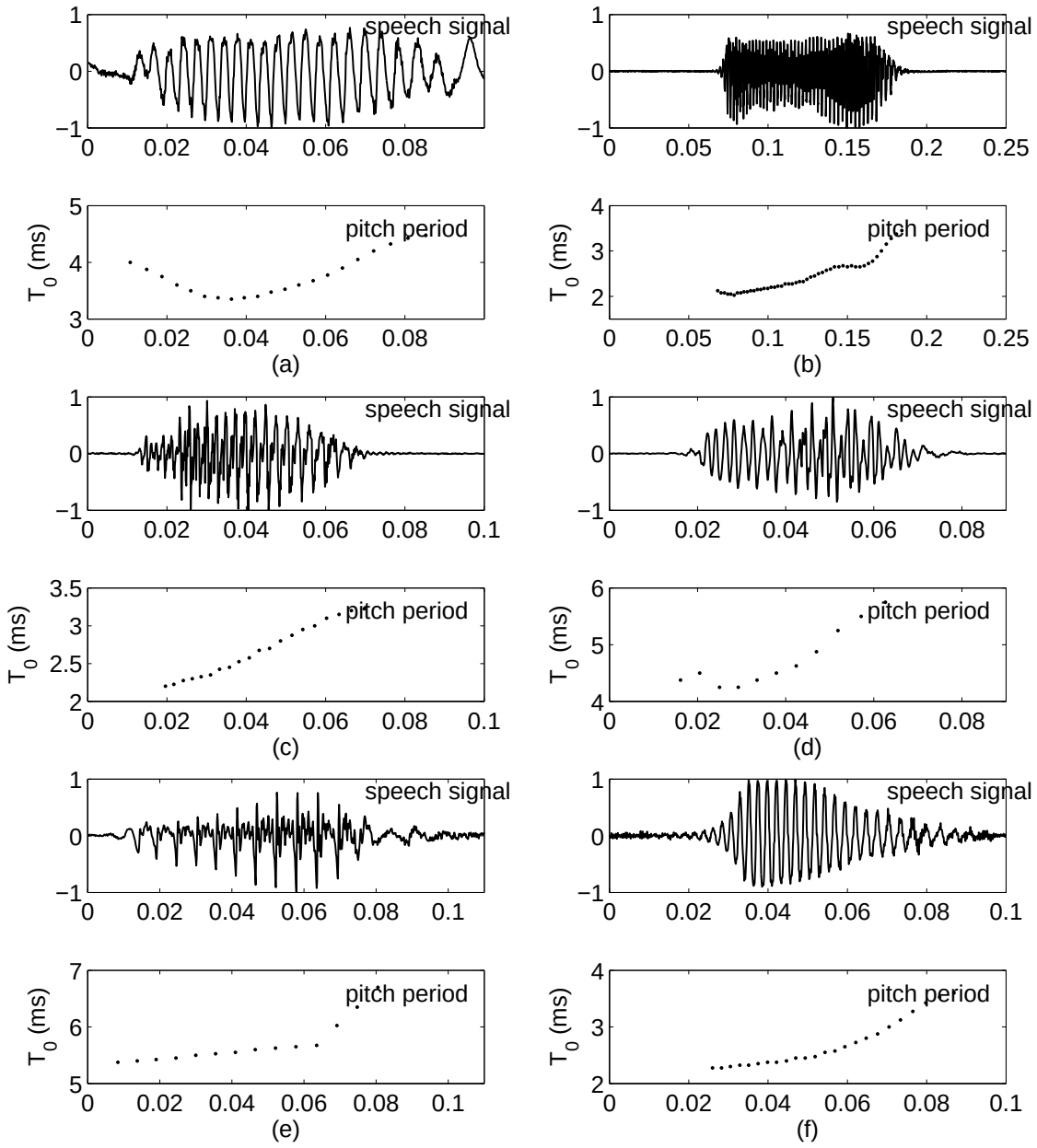

Fig. 3.5: Laughter calls of different laughs with the corresponding pitch period contours below.

Strength of excitation (α)

Since there is large amount of air pressure build up in the case of laughter, (as large amounts of air is exhaled), the closing phase of the vocal folds is very fast. This will result in an increase in the strength of excitation. Strength of excitation (α) at every epoch is computed as the difference between two successive samples of the filtered signal in the vicinity of the epoch. Fig. 5.3(c) shows this general trend of α in the calls within a bout.

Duration of the opening phase (β)

Since the closing phase of the vocal folds is fast for laughter, the corresponding opening phase will be larger in duration. So we have used the ratio (β) of the strength to excitation (α) at the epoch location and the pitch period (T_0) as an approximate measure of the relative duration of opening phase.

$$\beta = \alpha/T_0 \tag{3.4}$$

Slope of T_0

Two different measures are used for capturing the intra-call variations of pitch period contour. First one for capturing the rate of variations in pitch period contour and the second one for capturing the patterns in pitch period contour.

The pitch period contour of laughter has a unique pattern of rising rapidly at the end of a call. So, we use the slope of the pitch period contour to capture this pattern. First the pitch period contour is normalized between 0 and 1. At every epoch location the slope of the pitch period contour is obtained using a window width of 5 successive epochs. The slope is calculated by dividing the difference between the maximum and minimum of the 5 pitch period values within each window by the duration of the window. We denote this slope by δT_0. This value will almost be zero at the first half of the laughter call and a value approximately corresponding the slope of the contour during the rising phase.

For measuring the possible patterns in the pitch period contour, an experiment has been conducted. We observed that unlike normal speech, laughter calls can have only specific configurations allowed in the pitch period contour. The pitch period contour of the laughter call is divided into three equal segments (p_1, p_2 and p_3). The slope of the contour in each segment is determined by fitting them with linear polynomials. The segments are then categorized into three groups (a) rising, (b) flat or (c) falling based on the obtained slope values. The decision is done as shown in Table.3.1

The slope of contour after decision is denoted by p'_1 for 1st segment, p'_2 for 2nd segment and p'_3 for 3rd segment. So, the pitch period can have 27 (3^3) different configurations

Table 3.1: Table showing the condition on slope values for grouping.

condition	p'
slope > 0.15	1 (rising)
-0.15 < slope < 0.15	0 (flat)
slope < -0.15	-1 (falling)

as per the above grouping. An analysis is done on the number of laughter calls falling in each configuration. It is observed that only 5 of the 27 are predominant. Table 3.2 gives the frequency of occurrences in each configuration. We can see from the table that 93.6% of the laughter calls have one of the 5 configurations while the rest of the 6.4% calls are occupied by 22 configurations.

Table 3.2: Frequency of occurrences of calls in each configuration.

p'_1	p'_2	p'_3	call frequency
-1 (falling)	1 (rising)	1 (rising)	26.4%
1 (rising)	1 (rising)	1 (rising)	21.6%
0 (flat)	0 (flat)	1 (rising)	17.9%
0 (flat)	1 (rising)	1 (rising)	16.4%
-1 (falling)	0 (flat)	1 (rising)	11.3%
	rest		6.4%

From the above analysis, we can say with a high confidence that a segment cannot be a laughter call if it has a pitch period contour which does not belong to one of these five configurations. Note that speech segments can also have a pitch period contour belonging to one of the five configurations and so the values of p'_1, p'_2 and p'_3 together are used for eliminating false alarms for laughter spotting. Also note that the slope values are greatly affected by the wrong voiced non-voiced decision and spurious pitch period values. Fig. 3.6 demonstrates the pitch period contours and the line fitted pitch period contours (p'_1,p'_2,p'_3) for some laughter calls.

Slope of α ($\delta\alpha$)

As in the case of the pitch period, the strength of excitation at epochs also changes rapidly. It rises rapidly to some maximum value and again falls at the same rate. Hence the slope of the normalized strengths is calculated by dividing the difference between maximum

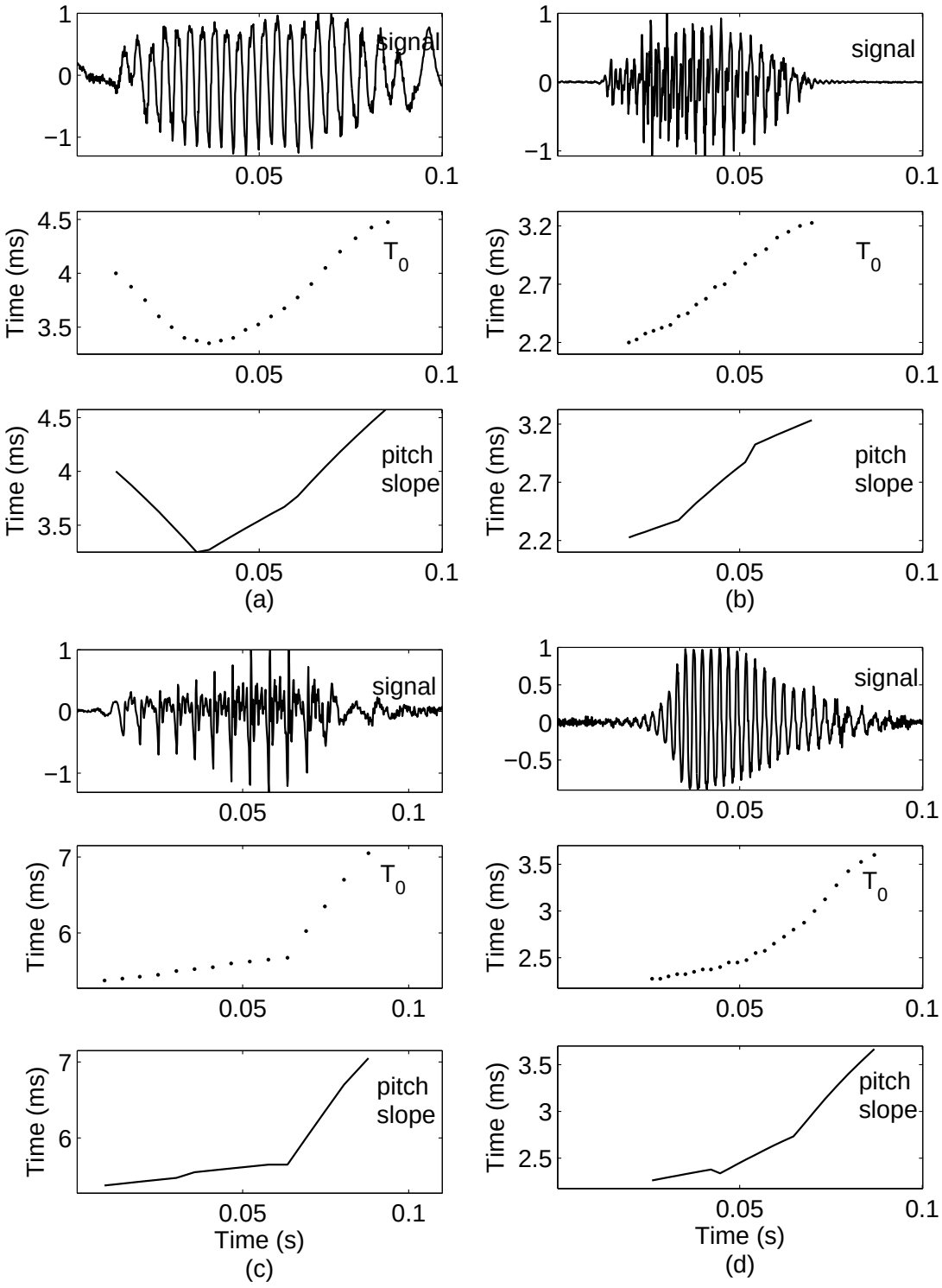

Fig. 3.6: Laughter calls of different laughs with the corresponding pitch period (T_0) contours and pitch slope (p') contours.

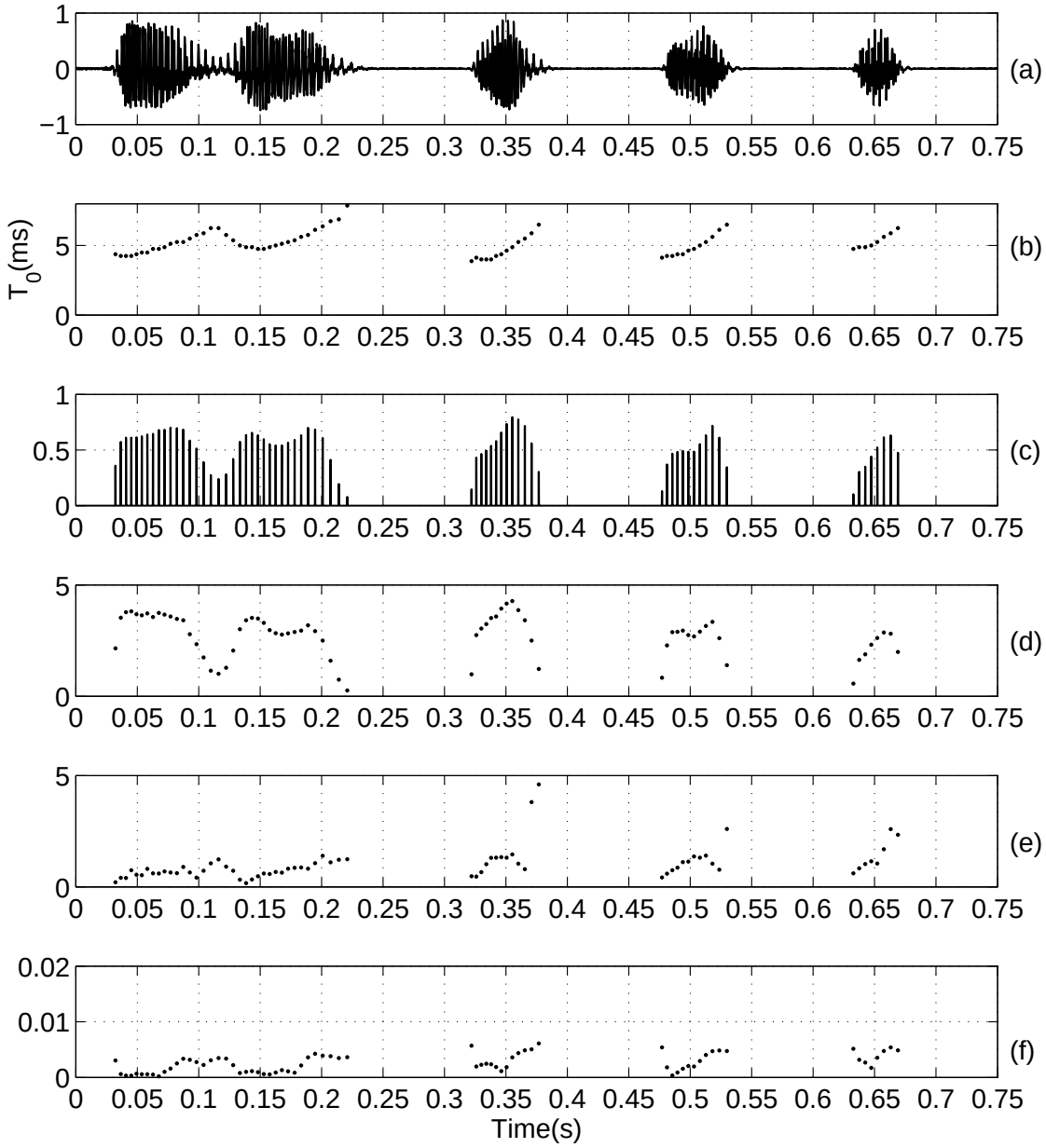

Fig. 3.7: (a) A voiced laughter with epoch locations. (b) Pitch period contour derived from epoch locations. (c) Strength of excitation derived from the zero-frequency filtered signal. (d) Ratio of strength of excitation to pitch period. (e) Normalized slope of pitch period. (f) Normalized slope of strength of excitation.

and minimum of the normalized strength values within 5 epochs window by the duration of the window. We denote this slope by $\delta\alpha$.

Breathiness

Because of high amount of airflow, laughter is typically accompanied by some amount of breathiness. Breathiness is produced with the vocal folds loosely vibrating and as a result more air escaping through the vocal tract than modally voiced sound [38]. This type of phonation is also called glottal frication and is reflected as high frequency noise (non-deterministic component) in the signal.

A breathy signal will typically have less loudness and more non-deterministic (noise) component. Measures based on hilbert envelope (HE) are used for calculating loudness and proportion of non-deterministic component in the signal. Loudness is defined as the rate of closure of vocal folds at the glottal closure instant (gci). This can be computed from excitation signal (residual) obtained from inverse filtering the signal (LP analysis [39]). But, the detection from the LP residual is difficult because of the amplitude values with either polarity occurring around the instants of gci. This difficulty can be overcome by using the Hilbert envelope of the LP residual [40]. The Hilbert envelope $r[n]$ of the LP residual $e[n]$ is given by

$$r[n] = \sqrt{e^2[n] + e_H^2[n]}, \tag{3.5}$$

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

$$e_H[n] = \text{IFT}(E_H(\omega)), \tag{3.6}$$

where IFT denotes the inverse Fourier transform, and $E_H(\omega)$ is given by (Oppenheim and Schafer, 1975[[41]])

$$E_H(\omega) = \begin{cases} +jE(\omega), & \omega \leq 0 \\ -jE(\omega), & \omega > 0. \end{cases} \tag{3.7}$$

Here $E(\omega)$ denotes the Fourier transform of the signal $e[n]$.

The idea is that the non-deterministic component will remain as noise in the residual obtained after doing an LP (Linear Prediction) analysis on the signal. Fig. 3.8 demonstrates the effect of breathiness on LP residual and hilbert envelope. It can be clearly observed from the figure that the non-deterministic part in the excitation is reflected as

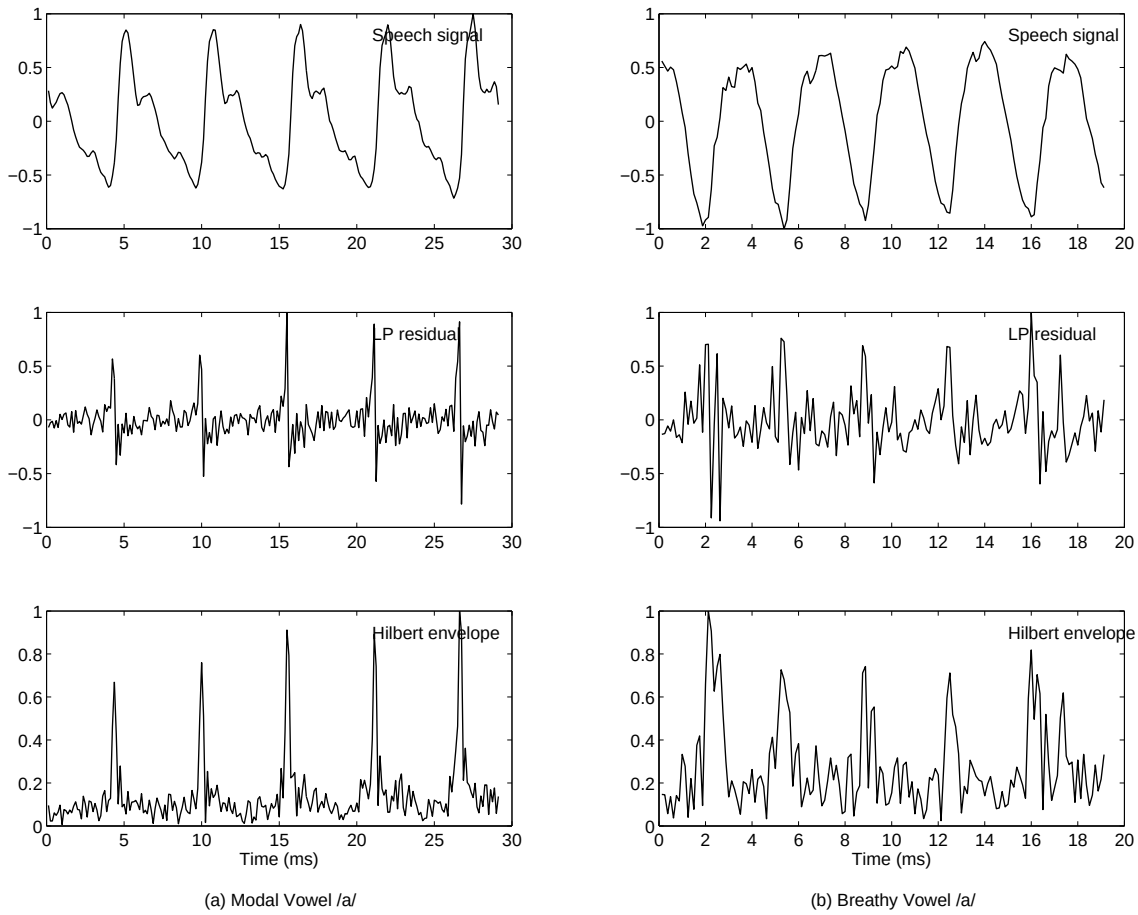

Fig. 3.8: Illustration of effect of breathiness on LP residual and hilbert envelope. Speech signal, LP residual and Hilbert envelope for (a) modal vowel /a/ and (b) breathy vowel /a/.

noise, and broader peaks in both the LP residual and hilbert envelope of Fig. 3.8(b).

We propose a quantitative measure for capturing these variations. A measure of the non-deterministic component (γ) has been computed as the ratio of the energy of the hilbert envelope in the open phase to the amplitude of the peak. The sharpness of the Hilbert envelope of the LP residual is captured by the parameter η, which is computed using a 2 ms segment of the Hilbert envelope around each epoch. The strength of these impulse-like excitation (also called the strength of excitation in [42]) is expressed by $\eta = \frac{\sigma}{\mu}$. Here μ denotes the mean of the samples of the Hilbert envelope (HE) of the LP residual in a short interval around the instants of significant excitation, and σ denotes the standard deviation of the samples of the HE [42]. The difference can be observed better in the ratio η/γ. Fig. 3.9 shows the difference in the values of η, γ and *eta*/*gamma* between laughter and speech. It can be observed that value of η (measure of loudness) is less and

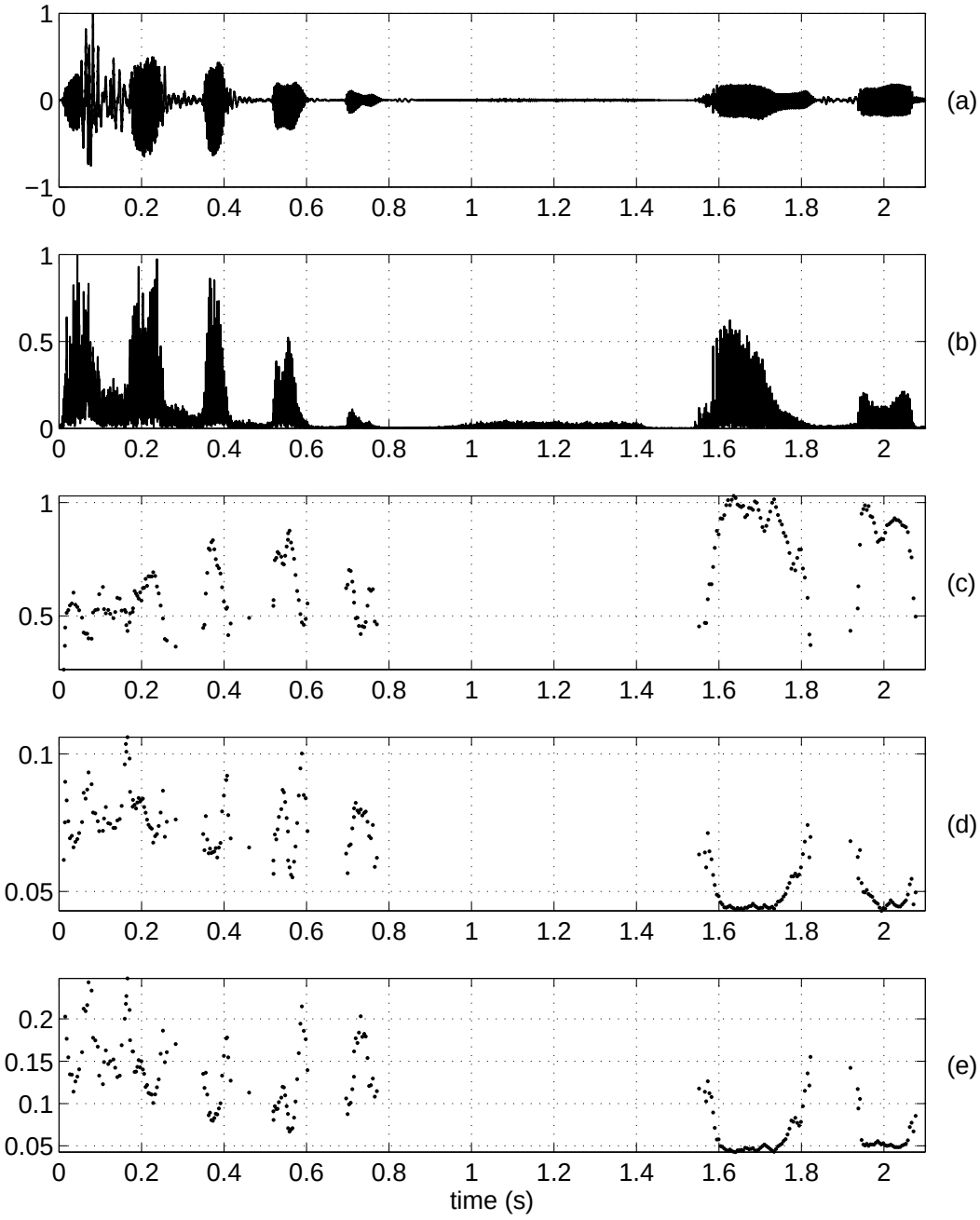

Fig. 3.9: (a) Speech signal containing both laughter and speech, (b) Hilbert envelope, (c) η, (d) γ, (e) η/γ.

value of γ (measure of breathiness) is more for laughter than that of speech.

3.2.2 Bout level analysis

The repetitive pattern in laughter bout is exploited using bout level features. Durations of calls and intercall intervals, correlation between calls are used as bout level features.

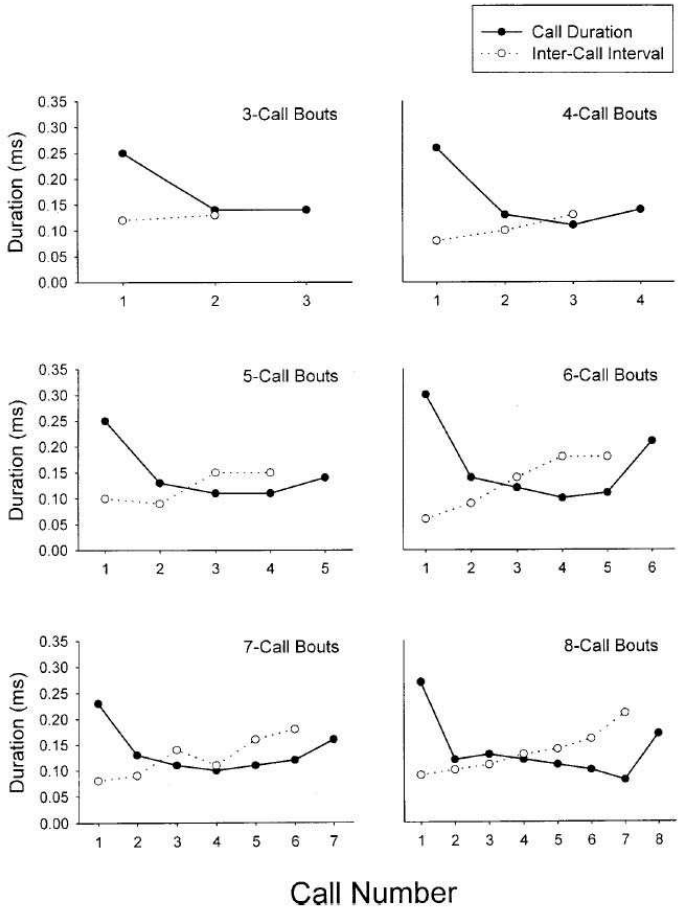

Fig. 3.10: Pattern of call and inter-call durations for different call numbers.

Call and Inter-call durations

As we have already seen, the laughter bout consists of voiced calls in between silence or unvoiced regions. An analysis is performed on the duration of laughter calls. The mean duration of the calls is observed to be 132 ms with a variance of 40 ms. Fig. 3.12 shows the distribution of calls of laughter. Bachorowski et. al [4] reported that there is also

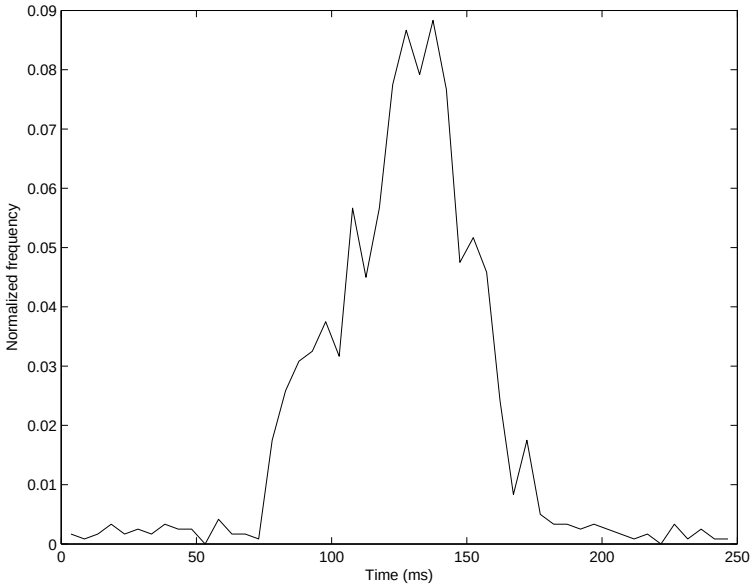

Fig. 3.11: Distribution of duration of calls of laughter.

a specific pattern followed by the call durations and intercall durations for different call numbers based on their observations. Fig. 3.10 shows the pattern followed by the call and intercall durations for different call numbers. The durations of these voiced regions (calls) and the non-voiced regions almost remain same through out the laughter bout. This pattern which is unique to laughter cannot be seen in case of normal speech. Fig. 3.12 shows the distribution of the differenced call durations and intercall durations for laughter and normal speech. The distribution shows that difference in the call durations is almost zero for most of the calls. The same trend can also be observed in the differenced intercall durations too.

Call level correlation

Because of the repetitive pattern of laughter, the laugh syllables will repeat itself for some number of times. So the signal (laughter syllables or calls) also will almost repeat itself for some number of times which is not the case in normal speech. The features like pitch period, strength of excitation etc between two laugh syllables will match closely. Correlation of these feature contours between two consecutive voiced regions (will be consecutive calls in case of laughter) is used as a feature. Since the laughter calls are almost similar, the correlation will be more in case of laughter than normal speech.

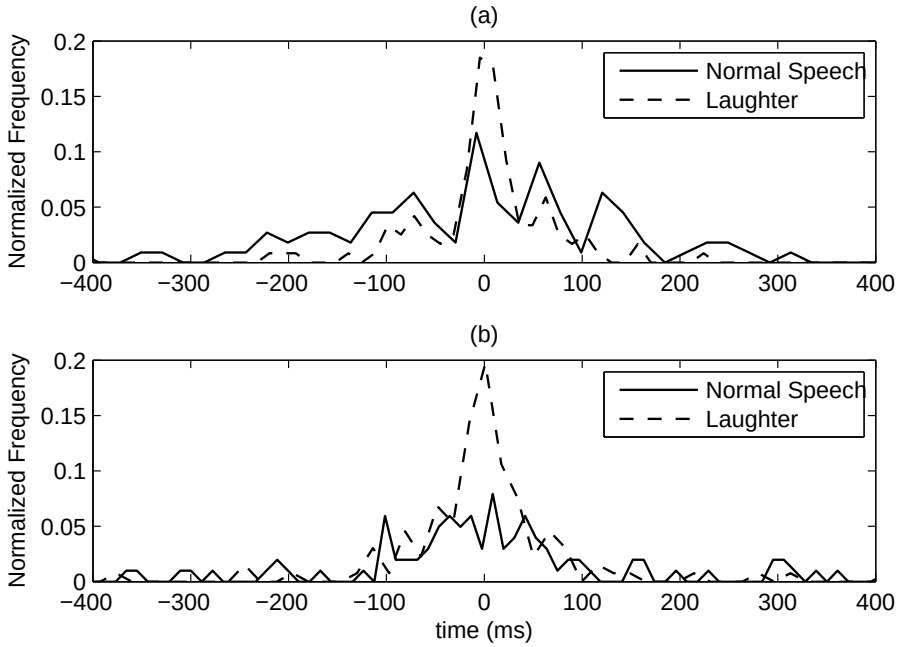

Fig. 3.12: Illustration of differences in the durations of voiced regions (calls in case of laughter) and unvoiced regions (intercall interval in case of laughter) between laughter and normal speech using the distributions of the differenced call and intercall durations. (a) differenced call durations (b) differenced inter-call durations.

3.3 Distributions of the features for laughter and speech

As mentioned earlier, the production of laughter and speech are different in many aspects. As a result, source features like pitch period (T_0), strength of excitation (α), breathiness differ. Distributions of the features T_0, α, β, δT_0, $\delta\alpha$, η and γ for laughter and speech samples of 5 male and 5 female speakers are shown in the first columns of Fig. 3.13 and Fig. 3.14. We can see from the distributions that there are certain regions where the laughter feature values are more concentrated, and there are regions where the speech feature values are more concentrated. This difference in distribution of features show that they could be used to discriminate between speech and laughter.

There is also a pattern observed in the variations of feature values across calls within a bout. The intensity of the feature values decreases as the calls progress. This is showed using the distributions of first calls and last calls of laugh bouts. The features extracted from the last call will be more closer to speech when compared to those extracted from the first call. This hypothesis can be clearly observed from the distributions showed in first

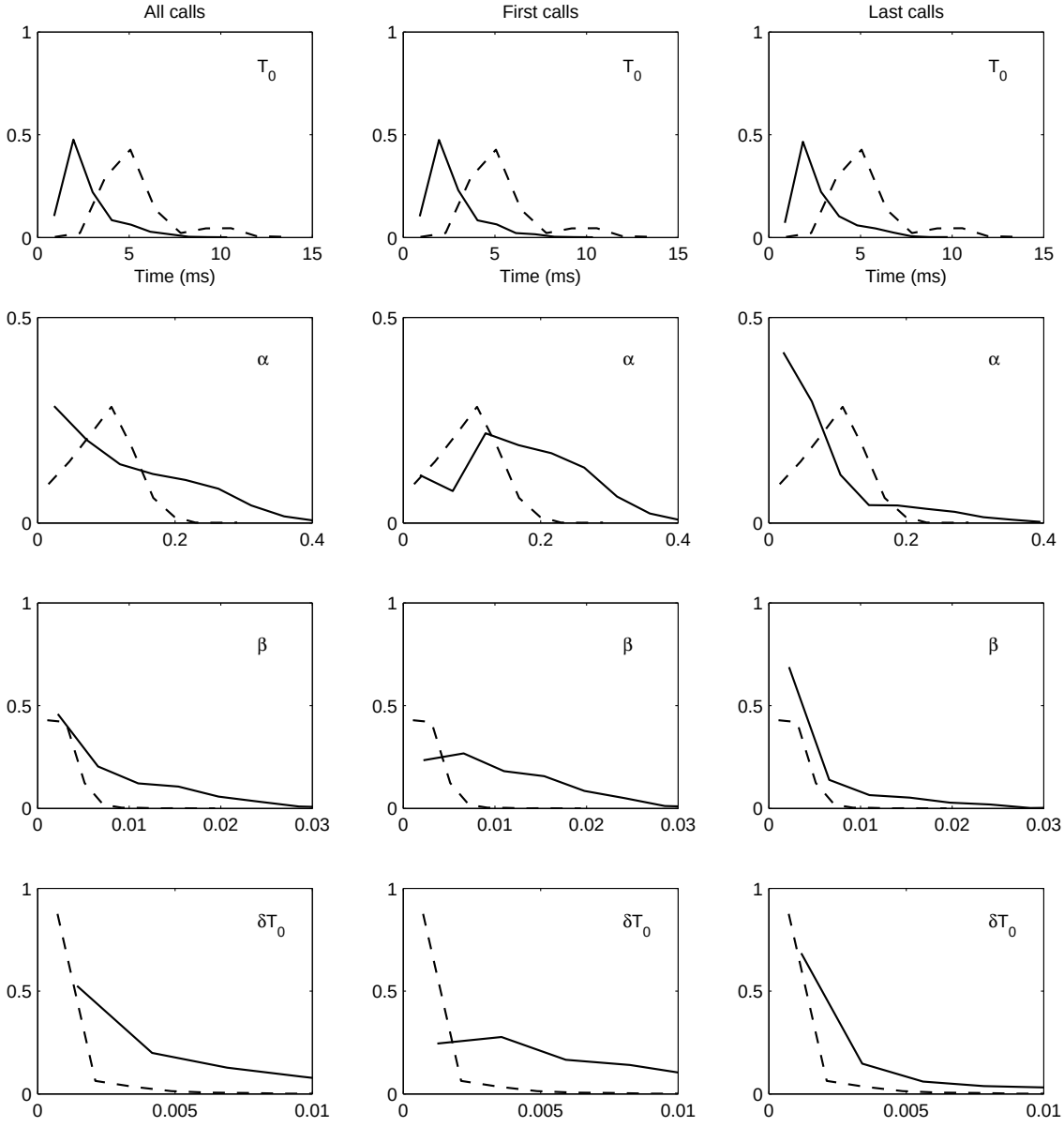

Fig. 3.13: Illustration of differences in the excitation source features of Laughter and Normal speech using frequency distributions (first column). It also shows the variations of features across the calls within a laughter bout using the distributions of first and last calls (second and third columns). In each plot the distributions of corresponding features of laughter and normal speech are represented with solid and dashed lines respectively.

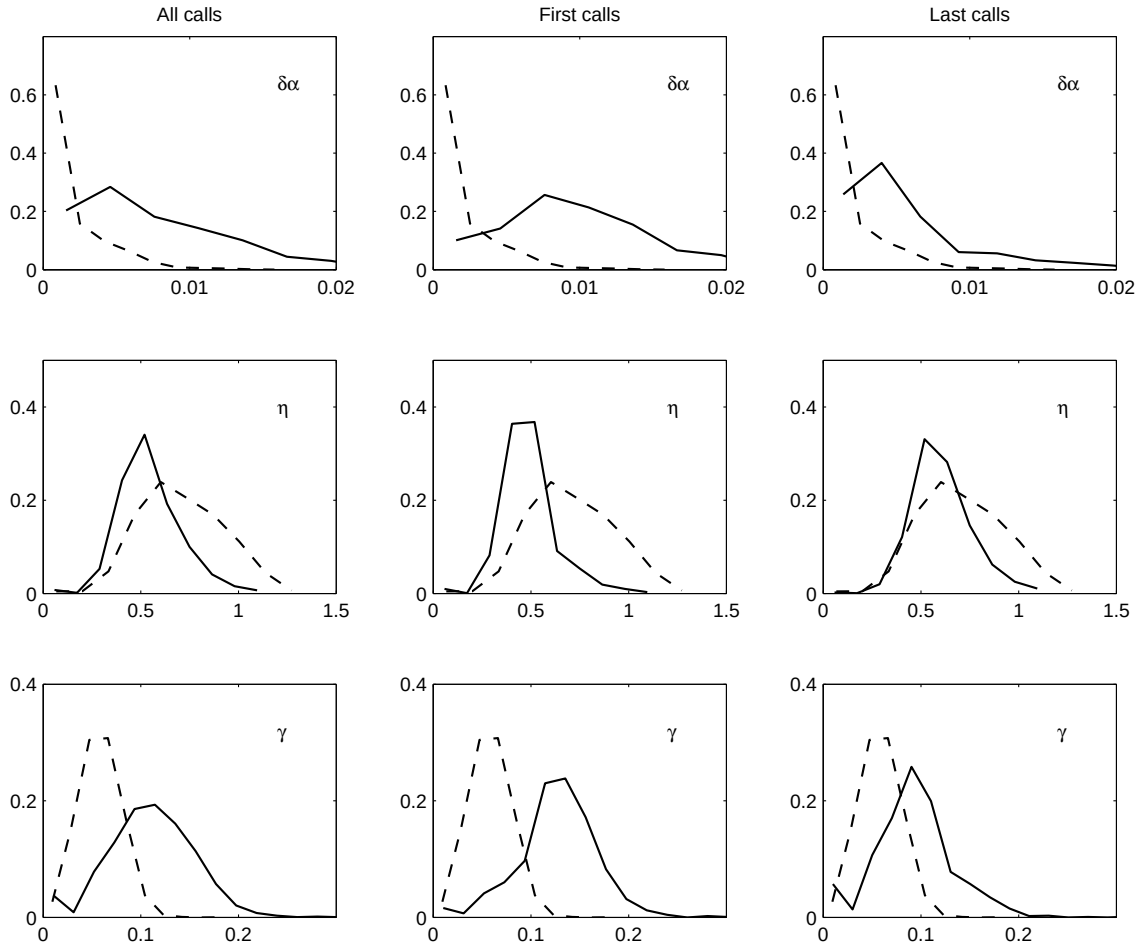

Fig. 3.14: Illustration of differences in the excitation source features of Laughter and Normal speech using frequency distributions (first column). It also shows the variations of features across the calls within a laughter bout using the distributions of first and last calls (second and third columns). In each plot the distributions of corresponding features of laughter and normal speech are represented with solid and dashed lines respectively.

call (2nd column) and last call (3rd column) of Fig. 3.13 and Fig. 3.14. We can observe from the figure that there is a clear discrimination between the distributions of features of first calls of laughter and speech while the distributions of last calls of laughter and speech are very close.

3.4 Summary

In this chapter, we presented the study on the analysis of laugh signals using features which are motivated from its production mechanism. The features are based on pitch

period, strength of excitation, amount of breathiness and loudness. We showed using the distributions that the features can discriminate between speech and laughter. We also showed that this discrimination is more predominant in the initial calls and falls as the calls progress.

Chapter 4

Proposed method for detection of laughter in continuous speech

This chapter describes the method for automatic spotting of laughter in spontaneous speech using the features proposed in the previous chapter. Unlike conventional methods where the decision is made on frames of fixed frame size, here the decision is made on each voiced segment. Call level features are used for making decision on voiced segment and bout level features for detecting the final laughter bouts.

Section 4.1 describes the components of the proposed method for laughter detection. In Section 4.2, the two data sets used for evaluating the performance of the algorithm are described. Section 4.3 explains the results and Section 4.4 deals about the error analysis and further refinements.

4.1 Block diagram and algorithm of proposed method

The main blocks involved in the system are a) pre-processing, b) feature extraction, c) decision logic and d) post-processing. Each of them are explained in detail below.

4.1.1 Pre-processing

The preprocessing step mainly involves the voiced non-voiced segmentation. The voiced non-voiced segmentation is done based on the energy of the zero-frequency resonator. The signal is first passed through zero-frequency resonator with a 3ms window for trend removal. This window length has been chosen in such a way that it gives high energy in the filtered signal in case of speech and laughter and low energy in the unvoiced and silence regions. The slope of the filtered signal at positive zero-crossings, which is an estimate of strength of excitation, is computed. This slope contour is smoothed using a mean filter with a window size of 4 values. A threshold of 30 percent of the maximum value is put on the smoothed contour. The regions crossing this threshold are considered as voiced segments and the regions falling below this threshold are considered as non-voiced segments.

The above method for voiced non-voiced segmentation has a problem of segmenting glottal fricatives (/h/) (which is very common in laughter), as voiced segment. This is because of the reason that glottal fricative also has a dominant low frequency component as in case of voiced signal. We can see in Fig. 4.1 that the glottal fricative getting segmented as voiced region after 1st level of decision.

To overcome this problem a second level of segmentation is performed on the voiced segments obtained from above the step. It exploits the similarity between successive pitch cycles in voiced region. A correlation value (c) and lag value (l) for each epoch are obtained from correlation between the short segments of the signal in the vicinity of the consecutive epoch locations. The lag values (l) are then divided by 40 to derive normalized lag values ($\hat{l}$). The values of c will be close to 1 for epochs in voiced regions and less (than 0.5) for epochs which belong to non-voiced regions. Similarly, the values of $\hat{l}$ will be close to 0 for epochs in voiced regions and more for epochs belonging to non-voiced regions. The final decision is taken based on the difference between $\hat{l}$ and c. The voiced non-voiced algorithm can be summarized as follows:

1. The input signal is first passed through the zero-frequency resonator with a 3ms window for trend removal.

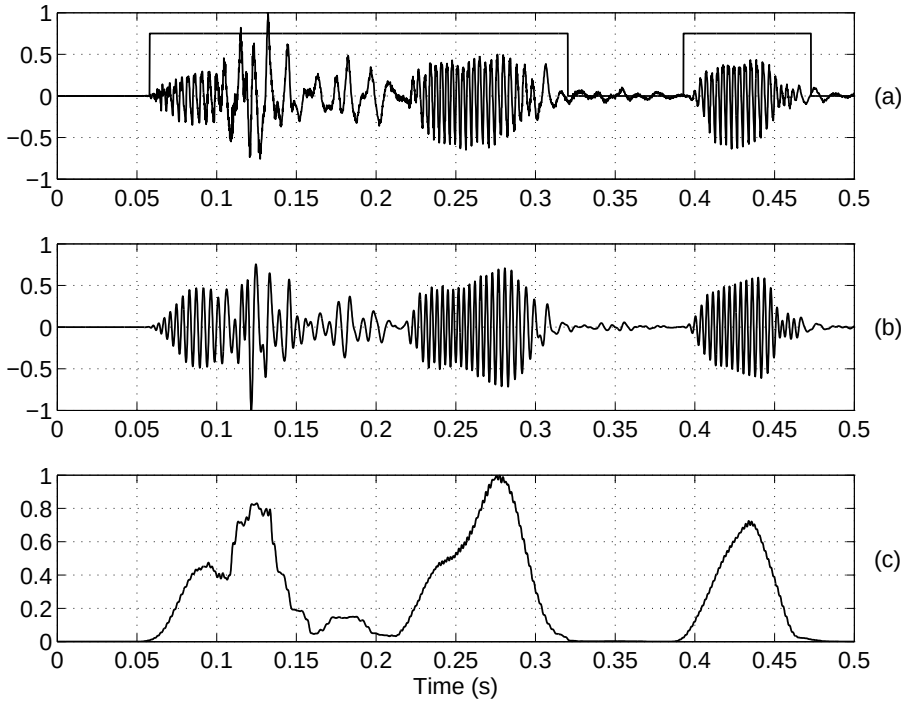

Fig. 4.1: (a) Voiced laugh signal containing glottal frication with voiced non-voiced decision (b) zero-frequency filtered (zff) signal with 4ms trend removal window, (c) energy of the strength of excitation contour derived from the slope of the zff signal.

2. The slope of the filtered signal at positive zero-crossings, which is an estimate of strength of excitation, is computed and the obtained contour is smoothed using a mean filter with a window size of 4 values. A first level of voiced non-voiced segmentation is done by putting a threshold (30% of the maximum value) on the smoothed contour.

3. For every voiced region obtained in the above step, correlation values (c) and normalized lag values ($\hat{l}$) are extracted for all the epochs in that region. The difference between c and $\hat{l}$, which is denoted by vc, is then computed for all the epochs.

4. A threshold of 0.5 is put on the vc contour. The regions in the voiced segment falling below this threshold are considered as voiced regions and the regions crossing the threshold are considered as non-voiced.

FIG: showing the pitch period diff contour, st diff contour and correlation with new v-uv decision.

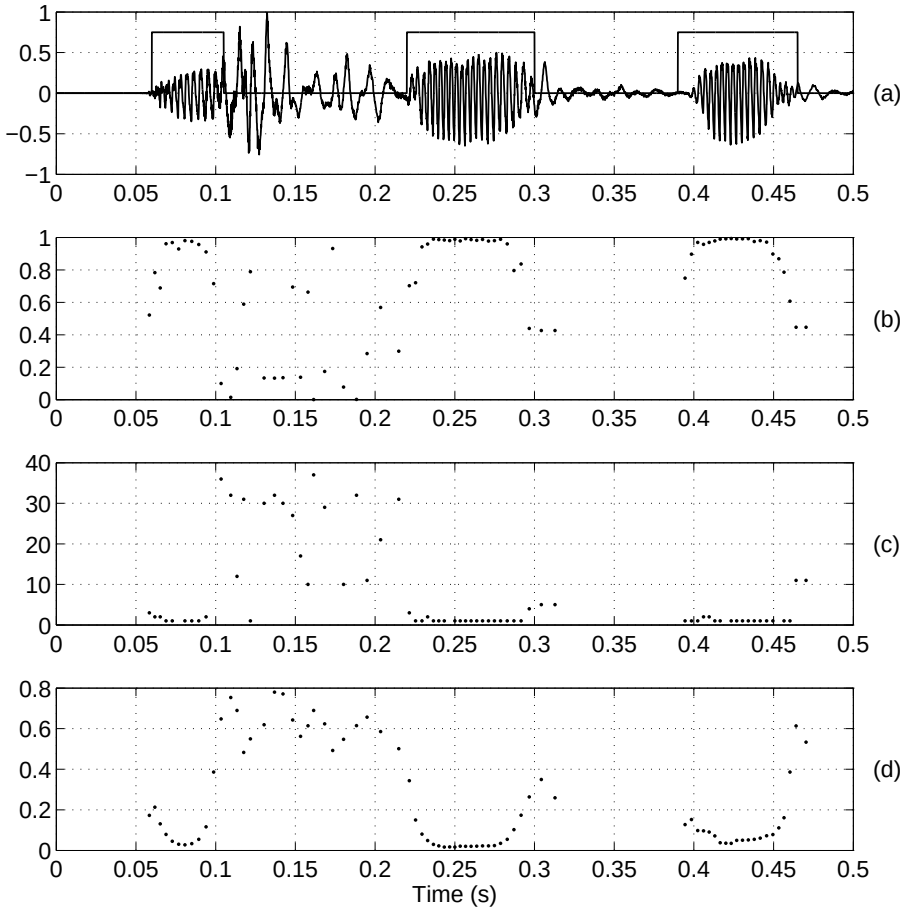

Fig. 4.2: (a) Voiced laugh signal containing glottal frication with corrected voiced non-voiced decision, (b) correlation values (c), (c) lag values (l), (d) $c - \hat{l}$.

4.1.2 Features

All the previous works in laughter spotting use conventional features like MFCC's, spectral tilt etc. [26][27][28]. The features proposed in this work are motivated from the production mechanism of laughter.

As we have already seen, the features used for laughter spotting are divided into two types a) call-level features and b) bout-level features. Call-level features are used for discriminating calls from other voiced segments of speech and bout-level features for capturing the intercall similarity within a bout. The call level features used are T_0, α, β, $\delta\ T_0$, $\delta\ \alpha$, η and γ. The bout level features used are δ duration, non-voiced duration. The extraction of each of these features are explained in detail in Chapter 3.

4.1.3 Decision Logic

After extracting the above features for every epoch location, a decision has to be finally made on the voiced segment based on these values. Note that this work concentrates only on voiced laughter and unvoiced laughter are not considered.

For every feature, a decision is first made on each epoch in the segment. This is performed by putting a threshold on the feature value (different for each feature), which is called as 'value threshold' (vt) for that feature. If the feature value of an epoch satisfies this 'value threshold', it means that the epoch belongs to laughter according to that feature. A decision is then made on the segment by putting a threshold called 'fraction threshold' (ft), which determines the percentage of epochs that should satisfy the 'value threshold' for the segment to be a laughter segment. After applying the two thresholds, separate binary decisions on the segment are obtained for all the features. Finally, the segment is considered as laughter if atleast 4 out of the 7 features gave a positive decision. The estimation of 'value threshold' and 'fraction threshold' is explained in the following subsection.

As we have already seen, a typical laughter bout will have strong laughter characteristics initially but falls as the calls progress. So, it is highly probable that the last calls may not get discriminated as laughter. To overcome this problem, a slightly modified algorithm which takes the context information of the previous call into consideration is proposed. For this purpose, call duration and intercall duration are used as the bout level features. The final decision on the segment depends on a) feature decision of the segment b) confidence of the feature values obtained for the previous segment (c_{-1}) c) percent difference in the durations between current and previous segments (δd) and d) time interval between the two segments. It is performed by relaxing the 'value threshold' of the features of a segment if the previous segment is identified as laughter. This reduction in the 'value threshold' denoted by δvt is different for all the features. The value of δvt of a

feature is computed as follows:

$$\begin{aligned} \delta vt &= -1/2 \times c_{-1} \times vt \times exp(-\delta d), \delta d < 0.1 \\ &= 0, otherwise \end{aligned}$$

where c_{-1} is a measure of feature confidence of previous segment, obtained as the difference between mean of the maximum n values and the value threshold of the segment. n is determined from the 'fraction threshold' of the feature. δd is the percentage difference between the duration of the current segment and the previous segment.
Note that $1/2 \times c_{-1}$ is the maximum attainable change in the feature threshold for a feature.

After obtaining the decisions on individual voiced segments, pitch slope values (p1,p2,p3) are used for eliminating some false alarms.

Threshold estimation

The 'value threshold' and 'fraction threshold' introduced in the previous subsection are estimated from the distributions of the feature values on the training data and the process is explained below. Feature values are normalized at some level before deriving the thresholds. For all possible combinations of feature values (minimum to maximum) and fraction values (0 to 1), the percentage of laughter segments identified as laughter (dr) and non-laughter segments identified as laughter (far), are computed. For ideal thresholds, the value of dr should be equal to 1 and value of far should be equal to 0. Selecting the best combination of thresholds is an optimization problem which try to maximize the difference ($dr - far$) in the most generic case. There may also be some additional constraints on the value of dr (should not be lesser than a particular value, which basically means that detection rate should not be too less) etc. The thresholds should be able to satisfy all these constraints.

Fig. 4.3 shows the values of dr, far and ($dr - far$) for all the features. Red color in the figure indicates the maximum value and black color indicates the minimum value and the in between colors correspondingly mapping between maximum and minimum. The

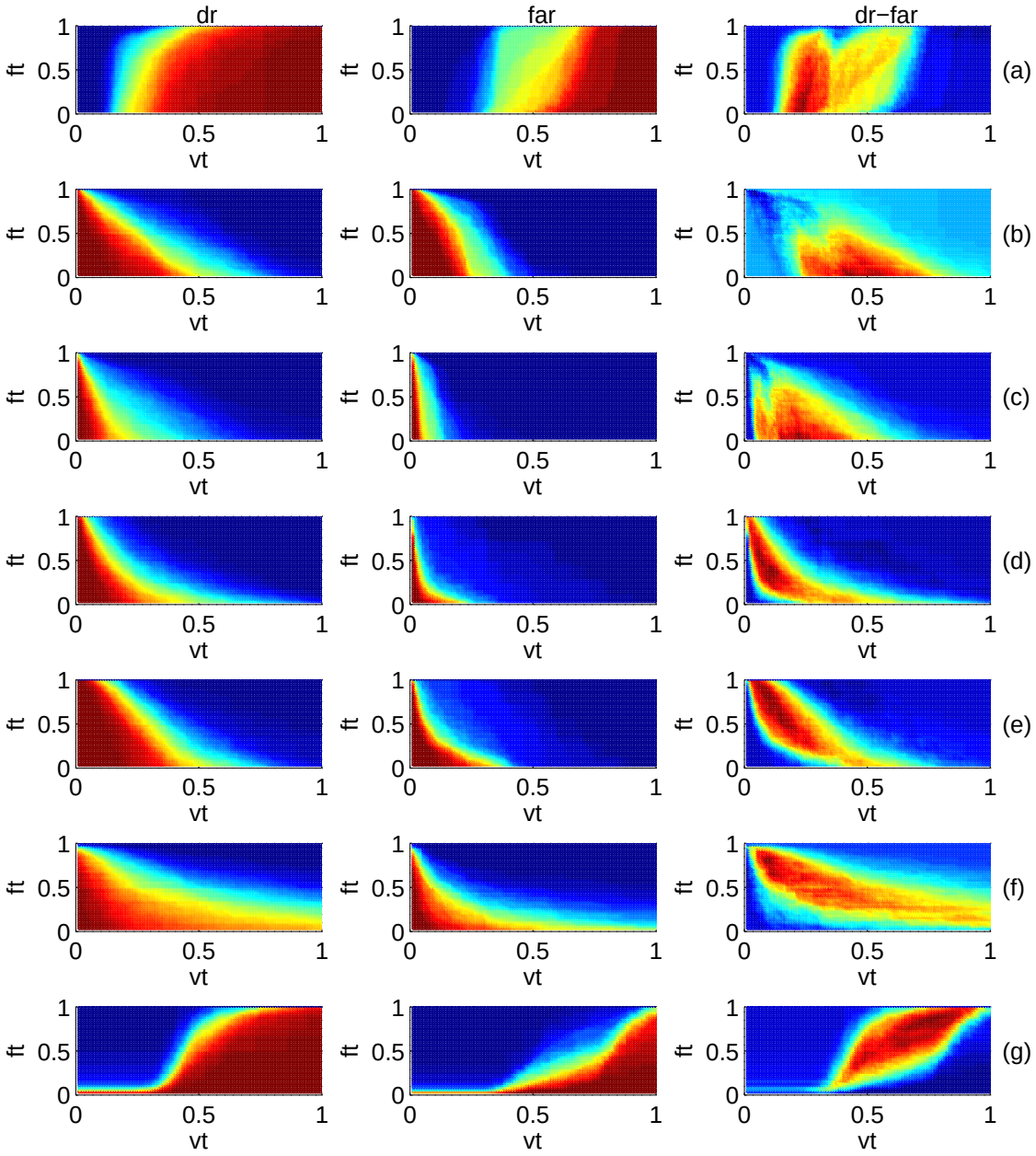

Fig. 4.3: Illustration of estimation of the value and fraction thresholds for all the features. (a) T_0, (b) α, (c) β, (d) δT_0, (e) $\delta\alpha$, (f) η and (g) γ. First column, second column and third column shows the values of dr, far and $(dr - far)$ respectively.

thresholds are selected by picking the maximum point in the $(dr - far)$ figures. Table 4.1 gives the value thresholds, fraction thresholds and value of dr, far and $(dr - far)$ at the points where the optimal thresholds are obtained, for each feature.

Table 4.1: Value and fraction thresholds for each feature.

S.No	Feature	Value Threshold	Fraction Threshold	dr	far	$(dr - far)$
1	T_0	3.6	0.3	0.6340	0.0988	0.5353
2	α	0.1548	0.05	0.7010	0.1852	0.5158
3	β	0.0059	0.05	0.5825	0.0988	0.4837
4	$\delta\, T_0$	0.0005	0.18	0.9433	0.2079	0.7156
5	$\delta\, \alpha$	0.0014	0.33	0.9072	0.1728	0.7344
6	η	0.0500	0.32	0.8170	0.0864	0.7306
7	γ	0.9750	0.33	0.8660	0.0988	0.7672

4.1.4 Post-processing

While the above decision logic gives a decision on the individual voiced segments, the final goal is to detect the boundaries of the laughter bouts. So, a post-processing block is included to attain such requirements.

Post-processing involves including some missed laugh segments, eliminating wrongly identified segments (false alarms) and finally joining the calls which belong to single laughter to form a complete bout. Non-laugh segments with duration between 50 ms and 150 ms, which has laugh segments on either side with a time gap of no more than 100 ms are included as laugh segments. Laugh segments with duration lesser than 50 ms, which occur in isolation (a time gap of 3 sec with laugh segments on either side) are eliminated. These two steps are performed in two different iterations. In the first iteration false alarms are eliminated and in the second iteration missed laugh segments are included. Boundaries of final laughter bouts are obtained by joining the adjacent laughter segments along with the intercalls.

The proposed method for laughter spotting can be summarized as follows:

1. The signal is first segmented into voiced and non-voiced regions using a two level segmentation. The first level of segmentation is performed by passing the signal through the zero-frequency resonator using a window length of 3 ms for trend removal. For every voiced segment obtained in the first level, a second level of segmentation is performed by exploiting the similarity between successive pitch cycles in voiced region.

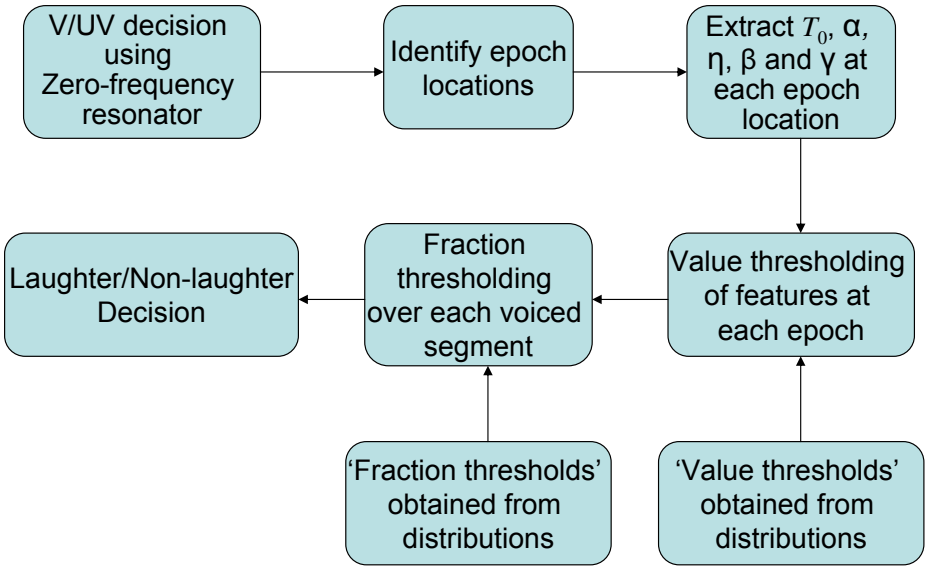

Fig. 4.4: Block diagram of proposed laughter spotting algorithm.

2. For every voiced region, epochs are extracted using the zero-frequency filtering method with a window size for trend removal derived adaptively from the signal. (Explained in detail in Section 3.1.)

3. The call level and bout level features described in Section 3.2 are extracted for every epoch in the voiced region.

4. If the previous segment is identified as laughter, the 'value thresholds' of the features of the segment are modified using a) feature confidence of the previous segment (c_{-1}), b) percent difference in the durations between current and previous segments (δd) and c) time interval between the two segments.

5. If a voiced segment has more epochs satisfying the 'modified value threshold' than determined by the 'fraction threshold' for atleast 4 features, then that segment is considered as a laughter segment.

6. Finally, the obtained laugh segments are passed through a post-processing block for detecting the boundaries of complete laughter bouts.

4.2 Data

The proposed algorithm for laughter detection is tested on two different datasets a) AMI Corpus which is a clean data and b) TV Show data which is a noisy data.

AMI Corpus: The AMI Meeting Corpus is a multi-modal data set consisting of 100 hours of meeting recordings[[43]]. The data is recorded in 70 different meetings. The system is tested on some part of the data which includes 10 meetings in ES subset, 6 meetings in IS subset and 8 meetings in TS subset. The data has about 5485 laughter bouts with an average of 3.96 calls per bout (21720 calls) uttered by 41 female and 55 male speakers.

TV Show data: The data is a TV broadcast data with each episode typically about 30 minutes of informal interview with one or more celebrities. It contains spontaneous laughter in between naturally occurring speech. The data has different kinds of noise like low amplitude background music and noise, multi-speaker utterances etc. The data is around 3 hours containing 106 laughter bouts with an average of 4.13 calls per bout (437 calls) uttered by 6 male and 3 female speakers.

4.3 Results

The laughter segments are manually labeled with start time and end time by listening to the data. The manually labeled laughter segments has the time stamps of the complete laughter bout, but not the individual calls. For obtaining the MDR (Missed Detection Rate) and FAR (False Alarm Rate) on voiced segments, these laughter regions are automatically segmented into voiced and non-voiced regions. This gives the start and end times of the voiced regions (calls) in a laughter. MDR and FAR are calculated based on these time stamps. If there is an intersection between the labeled time stamps and the

hypothesized time stamps, then that segment is considered as correct one.

The post-processing step is performed after obtaining the laugh segments. This resulted in a considerable reduction in the FAR as can be seen from Table 4.2. Table 4.2 shows the MDR and FAR at segment level and at bout level detection of laughter before and after post-processing. We can see from the results that FAR has reduced by a significant amount after the post-processing step in both the cases. The best performance of 10.6% EER (Equal Error Rate) is obtained on the AMI corpus and and an EER of 16.5% is obtained on the TV show data. The results are on par with the previous best of 8.9% EER, but cannot be directly compared since they were evaluated on a different dataset.

Table 4.2: Results of laughter spotting on AMI corpus and TV Show data.

	AMI Corpus		TV Show	
	MDR (%)	FAR (%)	MDR (%)	FAR (%)
Segment level	6.1	12.9	9.1	17.1
Segment level after Post-processing	6.1	7.2	9.2	10.2
Bout level	2.1	14.5	5.2	21.1
Bout level after Post-processing	2.1	8.5	5.4	11.1

4.4 Summary

This chapter describes the proposed method for detecting laughter in continuous speech. Two databases (one of which is a clean corpus and the other a noisy data) are used for evaluating the method. The performance of the method on the two datasets is discussed but however could not be directly compared with the existing techniques because of different data sets used.

Chapter 5

Synthesis of laughter

The ultimate goal of the speech synthesis systems is to synthesize long exchanges of human-machine dialogues in more natural way. There are many components that play key role in improving the naturalness. Some of them include imparting emotion to speech, intonation variations, non-lexical cues (throat clearing, tongue clicks, lip smacks, laughter etc.) etc.(Shrikanth Narayanan, 2006).

The goal of this chapter is to synthesize laughter by modeling and incorporating the feature variations explained in previous chapters. An analysis on the significance of each of these features in improving the naturalness of the synthesized laughter is also done from the results of the subjective perceptual evaluation performed on the synthesized laughter.

Section 5.1 describes how the features are modeled. In Section 5.2, the modification and incorporation of the features is explained. Section 5.3 describes the laughter synthesis procedure. In Section 5.4, the perceptual significance of the features is discussed and Section 5.5 discusses the results of the perceptual evaluation on the synthesized laughter.

5.1 Modeling the features

The source and system characteristics of laughter signals at call level are analyzed using features like pitch period (T_0), strength of excitation (α), spectral energy and each of them

is modeled separately as explained below.

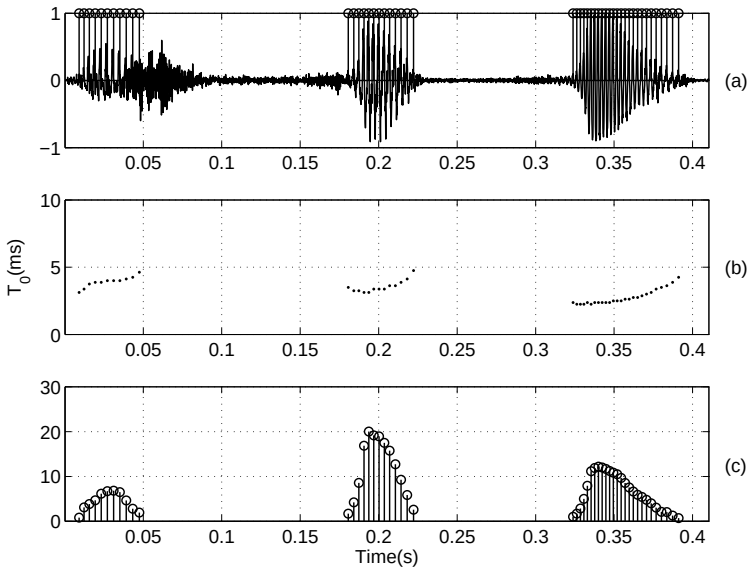

Fig. 5.1: (a) A segment of laughter signal. (b) Pitch period derived from the epoch locations. (c) Strength of excitation at the epochs.

5.1.1 Pitch Period

As mentioned earlier, the general pattern that is observed in the pitch period within a call is that it starts with some value, decreases slightly to some minimum, and then increases non-linearly to a high value. Many such pitch contours are analyzed and it is observed that the polynomial x^2 is a best fit for majority of natural laugh signals. The higher the slope of this rise in the pitch contour (t'), the more intense the laughter is. This slope of pitch period contour typically falls as the calls progress. The rate at which it falls (t'') is assumed to be linear. The pitch period contour of a call is obtained from the following equation

$$T_0[i] = t_0 + t'[c] * i^2, \quad -pn_e \leq i \leq (1-p)n_e \tag{5.1}$$

$$t'[c] = t'_0 - t'' * (n - c), \quad 1 \leq c \leq n \tag{5.2}$$

where

c is the call number,

n is the total number of calls,

i is the epoch number within the call,

p is the point in the call where minimum pitch period occurs,

t' is the slope of the pitch period contour,

t_0 is the mean pitch period of speaker obtained from input signal,

n_e is the number of epochs in the call and

t'' is the rate at which t' varies across calls.

Note that n, p, t' are the user inputs and all others are either obtained from these three values or fixed in the program.

5.1.2 Excitation Strength

The excitation strength at the epochs follows a similar pattern as pitch period contour. As in the case of the pitch period, the strength of excitation at epochs also changes rapidly. It increases non-linearly to some maximum value and then decreases almost at the same rate. The slope of strength of excitation contour (s') typically falls as the calls progress. The rate at which it falls (s'') is assumed to be linear as in case of pitch period. A similar model as pitch period is used here. The pitch contour of a call is obtained from the following equation

$$\alpha[i] = s_0 + s'[c] * i^2, \quad -n_e/2 \leq i \leq n_e/2 \tag{5.3}$$

$$s'[c] = s'_0 - s'' * (n - c), \quad 1 \leq c \leq n \tag{5.4}$$

where c is the call number,

n is the total number of calls,

i is the epoch number within the call,

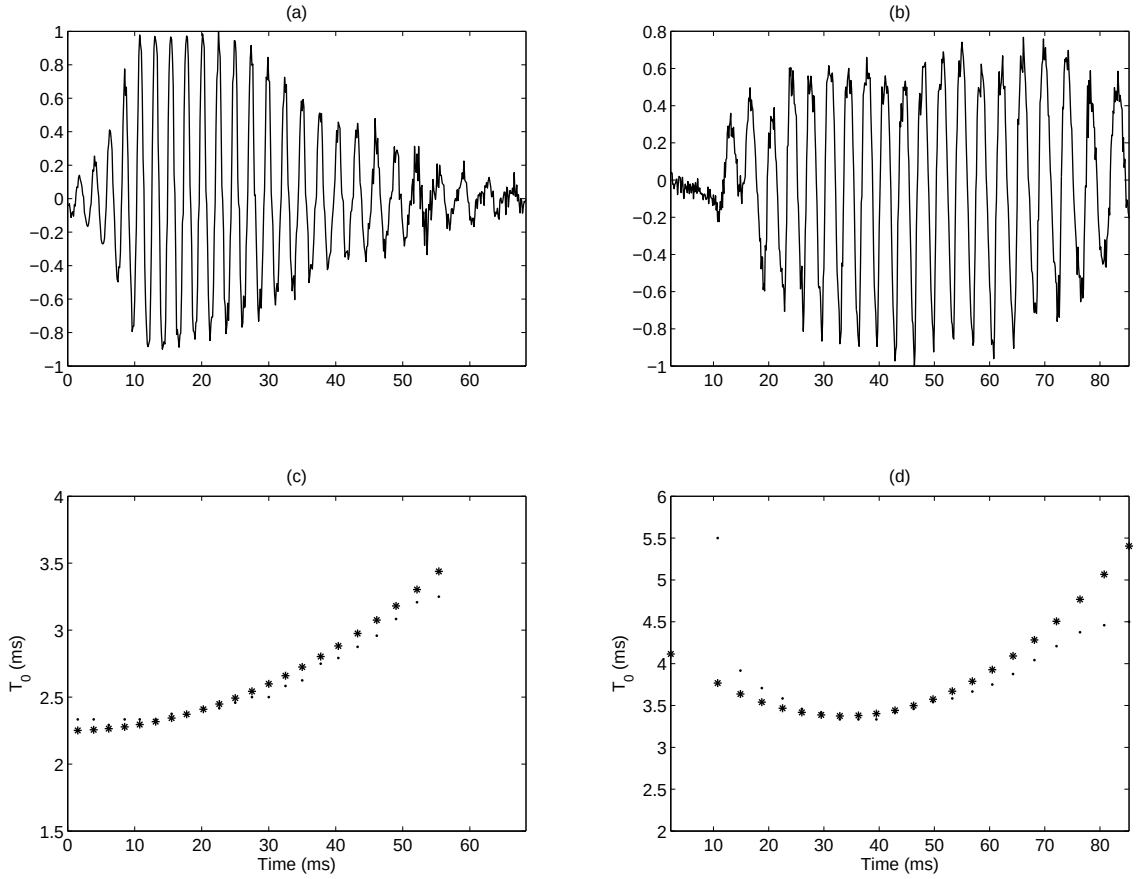

Fig. 5.2: Illustration on comparison of original and modeled pitch period contours. Two laugh calls are shown in (a) and (b) with their corresponding pitch period contours in (c) and (d), respectively. In (c) and (d), the actual pitch period contour is showed in thin dots (.) and modeled one is shown in thick dots (*)

s' is the slope of the pitch period contour,

s_0 is the mean pitch period of speaker obtained from input signal,

n_e is the number of epochs in the call and

s'' is the rate at which s' varies across calls.

Note that n, s' are the user inputs and all others are either obtained from these three values or fixed in the program.

5.1.3 Call and Inter-call Duration

As we have already seen, the laughter bout consists of voiced calls in between silence or non-voiced regions. Bachorowski and Owren [4] reported that there is specific pattern followed by the call durations and intercall durations for different call numbers based on their observations. The durations of these voiced regions (calls) and the non-voiced regions almost remain same through out the laughter bout. This pattern which is unique to laughter cannot be seen in case of normal speech. Duration of the inter-call interval in a laughter bout typically increases as the calls progress. There is no such pattern observed in case of call duration. It is either increasing or decreasing depending on the speaker and type of laughter.

From the frequency distribution of the duration of the calls in Fig. 3.12, we can see that mean is around 130 ms and the variance is 40 ms. So, the duration of the first call (d_1) is generated from a gaussian distribution with the observed mean and variance as follows

$$d_1 = 130 + 40 * randn(). \tag{5.5}$$

The variation of duration of calls within a bout is assumed to be linear. So, the durations of rest of the calls are determined as follows

$$d_i = d_{i-1} - d' * (n - i), 2 \leq i \leq n, \tag{5.6}$$

where d_i is the duration of the i_{th} call, n is the total number of calls and d' is the rate at which call duration varies across calls.

5.1.4 Frication

Because of high amount of airflow and constant glottal leakage, there will be turbulence generated with vocal folds as a result of which glottal fricative /h/ is produced. In most of the cases it is predominantly observed in the intercall interval. The volume velocity of air typically decreases as we go from left to right with in a call as a result of which

the amount of breathiness also fall down with in a call. Also the amount of breathiness decreases as the calls progress.

5.2 Incorporating feature variation

The derived models have to be finally incorporated in a synthesis framework for synthesizing laughter. Modification or incorporation of each of the modeled features are explained below.

5.2.1 Pitch period and Duration modification

Pitch period and duration modification is done in a pitch synchronous way. It involves the following steps:

1. The signal is passed through zero-frequency resonator for deriving the epoch locations. Pitch period is obtained by taking the difference of the epoch locations.
2. 10th order pitch synchronous linear prediction analysis is performed on the signal to separate it into source (residual) and system (LP coefficients). So, there will be a residual and lp coefficients associated with every epoch location.
3. Desired pitch period contour is generated either from the older contour or a completely different one can be taken (Pitch period modification).
4. New pitch period contour is generated by resampling the existing one with duration modification factor (Duration modification).
5. New epoch locations are derived from the obtained pitch period contour.
6. New set of residual and lp coefficients are generated for obtained epoch sequence from the nearest original epochs. The residual at each epoch is then resampled by pitch modification factor (obtained from original pitch period and obtained pitch period) at that epoch.

7. For every epoch in the epoch sequence, residual is passed through the corresponding LP filter to obtain the new signal. These signals are finally concatenated to obtain the desired signal.

5.2.2 Strength of Excitation modification

Strength of excitation is an estimate of the excitation strength at the epochs derived from the zero-frequency filtered signal. Since the speech signal cannot be reconstructed back from the zero-frequency filtered signal, modification of strength of excitation cannot be done on the filtered signal. So, we hypothesized that the strength of excitation can be modified in the residual signal. In order to find out the relation between α and the amplitude of peaks in the residual signal, if there is any, a small experiment has been conducted. A sequence of impulses with varying durations between consecutive ones (pitch period) and different amplitudes are generated and passed through all-pole resonator with lp coefficients corresponding to different vowels. The output signals are passed through zero-frequency resonator and strength of excitation values are calculated. The obtained α values are compared with the amplitudes of the impulses, which is the approximate residual. It can be clearly observed in the figure that the impulse strengths (residual) and the strength of excitation obtained from zero-frequency filtered signal have a linear relation between them.

So, the amplitudes of the samples in the vicinity of the epochs in the residual are modified according to the desired strength of excitation contour in an attempt to modify the strength of excitation.

5.2.3 Frication Incorporation

Frication or breathiness is incorporated by modifying the residual. 10th order linear prediction analysis is performed on the input signal to separate it into residual and system (LP coefficients). Random white gaussian noise of length equal to the length of the residual signal and total energy equal to the one-fifth of the total energy of the signal is generated.

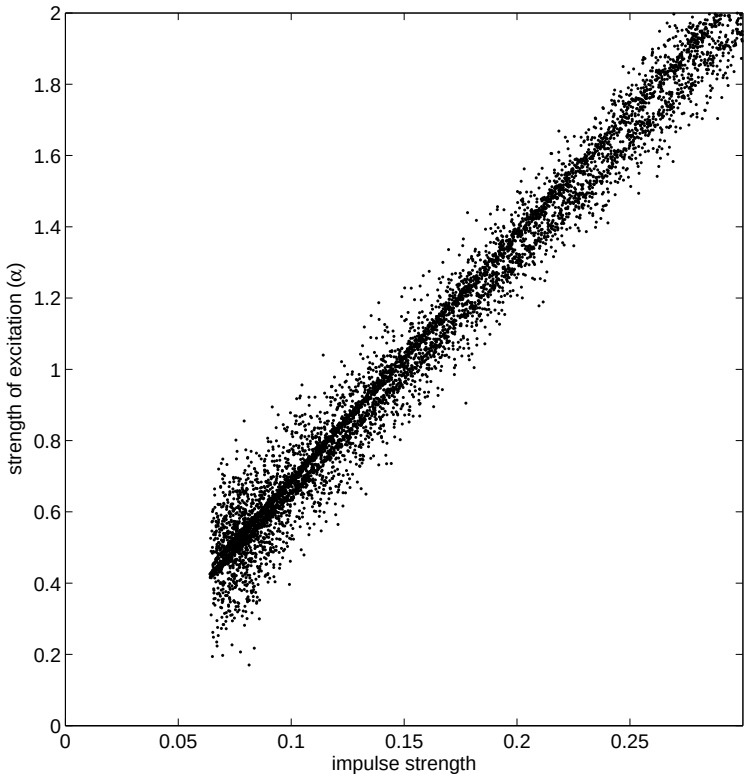

Fig. 5.3: Scatter plot demonstrating the correlation between impulse strength and strength of excitation obtained from zero-frequency filtered signal.

It is then passed through a resonator with central frequency at 4000 Hz and bandwidth of 1000 Hz. The resulting signal is first multiplied with a hanning window of equal length and then with a linearly falling line. The obtained signal is then added to the original residual and then passed through the LP filter for generating the breathy voiced signal.

5.3 Laughter Synthesis Procedure

Laughter (ha-ha or hi-hi) is synthesized by modifying the above features of vowel /a/ or /i/ uttered by a speaker. The process involves only modifying the source while the system almost remains the same.

The proposed method for laughter synthesis consists of the following steps:

1. The input signal (/a/ or /i/) is first passed through a zero-frequency filter for deriving the epoch locations. Obtained epoch locations are corrected using hilbert envelope

based method and new epoch set of epoch locations are obtained. Pitch period is obtained by taking the difference of the epoch locations.

2. The signal is resampled to 8000 Hz and a 10th order pitch synchronous linear prediction analysis is performed on the signal to separate it into source (residual) and system (LP coefficients). This gives a residual and a 10 element lpcc vector associated with each epoch.

3. For every call in the laughter bout, the values n, p, t', s' are obtained as user given inputs. The value of t_0, s_0 (derived from the pitch period contour and strength of excitation contour of the original vowel respectively), $t'[c]$, $s'[c]$ (derived from equation 5.2 and 5.4 respectively) are derived and pitch period contour, strength of excitation contour, duration of the call, duration of the following inter-call interval are calculated from the pre-assumed models.

4. New residual is obtained after modifying the pitch period and duration, strength of excitation and incorporating breathiness effect as explained in the previous section.

5. The obtained residual is then passed through the LP filter of the vowel to synthesize the call. The following intercall interval is also generated and concatenated to the call.

6. Steps 3,4,5 are repeated for all the calls in the laughter and finally concatenated to synthesize the laughter.

The complete laughter synthesis system described in Fig. 5.4 is implemented in MATLAB (http:// www.mathworks.com).

The synthesis procedure is mostly concentrated on synthesizing calls and not much care is taken on synthesizing bouts. It is also important to note that the voiced calls of real laughter are not truly speech vowel-like sounds, while we used the same (lpcc's of system) for synthesizing laugh calls.

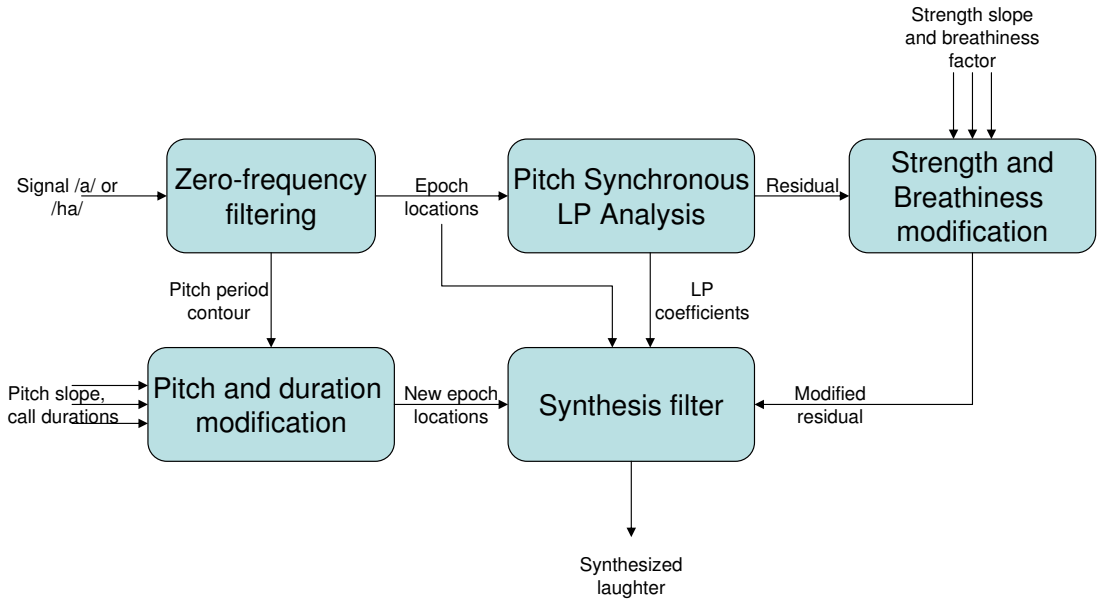

Fig. 5.4: Block diagram of the synthesis system.

5.4 Perceptual significance of features

This study aims to analyze the perceptual significance of the features described above. An experiment based on analysis by synthesis approach is conducted. As a part of the experiment, some original laugh signals are taken and the proposed features are modified. For each original clip, modifications are performed for all possible feature combinations, thus generating 15 ($2^4 - 1$) different clips. These clips are played randomly to 20 subjects who are told to score them for naturalness and acceptability. Acceptability is a measure of how close the sample is sounding to laughter and naturalness is a measure of how natural it is. The features modified are T_0, strength of excitation, amount of breathiness and call and inter-call durations. The modification is performed as given below:

The laugh signal is first segmented by making voiced non-voiced decision and calls and intercalls are extracted. For every laugh syllable (call plus inter-call), the following changes are performed:

Table 5.1: Perceptual evaluation scores obtained for the modified versions of an original laugh signal.

Sample	T_0	α	breathiness	call, inter-call durations	Naturalness		Acceptability	
					mean	variance	mean	variance
1	0	0	0	0	4.2	0.36	4.33	0.22
2	0	0	0	1	3.9	0.22	4.01	0.21
3	0	0	1	0	3.6	0.21	3.69	0.22
4	0	0	1	1	3.4	0.31	3.52	0.29
5	0	1	0	0	3.99	0.29	4.09	0.32
6	0	1	0	1	3.61	0.09	3.79	0.22
7	0	1	1	0	3.24	0.16	3.44	0.23
8	0	1	1	1	3.01	0.26	3.19	0.29
9	1	0	0	0	2.9	0.24	3.01	0.23
10	1	0	0	1	2.69	0.19	2.99	0.21
11	1	0	1	0	2.41	0.21	2.43	0.09
12	1	0	1	1	2.02	0.21	2.19	0.11
13	1	1	0	0	2.31	0.14	2.44	0.19
14	1	1	0	1	2.11	0.11	2.21	0.36
15	1	1	1	0	1.69	0.36	1.79	0.24
16	1	1	1	1	1.26	0.26	1.69	0.14

1. Rising pitch period contour is replaced with a flat contour.

2. Strength of excitation has been reduced.

3. Breathiness is reduced by decreasing the relative amplitudes of the samples in the non-epoch regions (samples which are more than 1 ms away from the epoch) and intercall intervals.

4. Desired call and inter-call durations have been generated randomly from the distribution of voiced unvoiced regions of normal speech.

Table 5.1 gives the results of the study. The table gives the naturalness and acceptability scores of some laughter clips for all possible modifications (including the original sample). In the Table , a ‘0’ in a feature column indicates that the feature is not modified and a ‘1’ indicates that it is modified in the given sample. We can see from the table that sample1, which is the original version (all 0’s), has the high Naturalness and Acceptability scores as expected and sample 16, in which all the feature are modified (all 1’s) has a

very less score (1.26), which means that these four features could characterize laughter to a maximum extent.

Table 5.2: Results of the experiment on perceptual significance of features.

	modified		not-modified	
	Naturalness (mean)	Acceptability (mean)	Naturalness (mean)	Acceptability (mean)
T_0	2.1	2.3	3.6	3.7
α	2.9	3.1	3.1	3.2
breathiness	2.4	2.5	3.3	3.4
call intercall durations	2.8	3.0	3.3	3.3

For finding the significance of each individual feature, the mean of samples when a particular feature is modified (1's) is compared with mean of samples when it is not modified (0's). Table 5.2 shows these comparisons for all features. The more the difference between the two, the more significant the feature is. We can see from that Table that difference is maximum in case of T_0 and minimum in case of α with breathiness taking the second position and call, inter-call durations taking the third position. The reason for α showing the minimum difference may be attributed to the level of our understanding on real excitation strength. It is also possible that it α doesn't play major role in isolated laughter, but becomes significant in laughter occurring in between speech, because unlike other features, it is a highly relative measure.

We can also see that variance of pitch period is less, which means that the decision is less speaker specific and the variance of breathiness is more showing that its perception is more speaker dependent.

5.5 Experiment for subjective evaluation

Subjective evaluation tests are performed on 28 naive volunteers. The volunteers are presented with 25 laughter-only clips of which 17 clips are synthesized offline using the technique presented here and the remaining 8 clips are isolated laughter taken from the AMI Corpus. The number of calls in the synthesized laughter, its duration, the F0 changes, the strength of excitation changes in each sample are given in Table 5.3. The 25 clips

are randomly played and not grouped in any particular order. The tests are performed in a typical quiet office environment on a computer terminal. Each volunteer had to listen and score each sample for naturalness and acceptability according to their preference on a scale of 1-5: 1-Very Poor, 2-Poor, 3-Average, 4-Good, 5-Excellent.

Table 5.3: Results of the laughter synthesis system.

Sample	slope of T_0	slope of α	breathiness	No. calls	Avg. Call Duration (ms)	Gender	Naturalness		Acceptability	
							mean	variance	mean	variance
1	0	0	0	4	124	M	1.21	0.24	1.24	0.21
2	0.1	0.1	0.1	4	101	M	1.69	0.16	1.87	0.24
3	0.2	0.1	0.1	4	115	F	2.41	0.09	2.59	0.11
4	0.3	0.1	0.1	5	86	M	3.23	0.31	3.62	0.24
5	0.4	0.1	0.1	3	81	M	3.11	0.34	3.31	0.21
6	0.4	0.2	0.1	4	96	F	3.33	0.21	3.46	0.23
7	0.4	0.3	0.1	5	75	M	3.49	0.26	3.64	0.31
8	0.4	0.4	0.1	4	130	F	3.41	0.09	3.63	0.16
9	0.4	0.4	0.2	6	125	F	3.55	0.11	3.61	0.19
10	0.4	0.4	0.3	5	117	M	3.69	0.15	3.74	0.19
11	0.4	0.4	0.4	6	90	F	3.61	0.29	3.76	0.31
12	0.4	0.4	0.3	7	95	M	3.59	0.25	3.79	0.24
13	0.4	0.4	0.3	8	125	F	3.22	0.12	3.31	0.39
14	0.4	0.4	0.3	5	89	F	3.49	0.22	3.56	0.21
15	0.4	0.4	0.3	5	121	M	3.61	0.16	3.69	0.11
16	0.4	0.4	0.3	6	88	M	3.53	0.22	3.63	0.16
17	0.4	0.4	0.3	4	98	M	3.66	0.21	3.77	0.17
18	–	–	–	5	121	M	4.26	0.12	4.61	0.11
19	–	–	–	4	84	M	4.44	0.16	4.72	0.06
20	–	–	–	6	55	F	4.39	0.21	4.77	0.09
21	–	–	–	4	96	M	4.42	0.09	4.54	0.13
22	–	–	–	5	30	F	4.09	0.14	4.51	0.19
23	–	–	–	5	145	F	4.21	0.11	4.64	0.09
24	–	–	–	4	89	F	4.54	0.10	4.61	0.11
25	–	–	–	6	134	M	4.39	0.19	4.59	0.07

5.6 Evaluation Results

For the analysis of the evaluations, we make the assumption that each laughter clip is an independent encounter by an individual subject. Thus, for N=25 subjects and 17 synthesized samples, we have a total of 28*17=425 samples ans for eight real laughter clips, we have 25*8=200 samples.

The 17 samples are synthesized by varying the input parameters like slope of pitch period contour, slope of α contour, amount of breathiness, number of calls etc. The mean and variance of the evaluation scores are listed in Table 5.3. The evaluation results are summarized below:

We can see from the Table that the mean scores increased with an increase in slope of T_0, slope of α and amount of breathiness. The increase is high in case of T_0, but low in case of α and amount of breathiness. But they are decreasing after some point in all the cases. This may be because of the unnaturalness introduced by the modification algorithm. Also notice that the scores decreased with an increase in the number of calls. This is because of the reason that we have used a simple linear model for deriving the call and intercall durations. The mean naturalness score obtained for the original samples (18 to 25) is 4.34 and the mean naturalness score obtained for the best synthesized samples (11 to 17) is 3.55. Note that we have only accounted for the differences in the source and not modified the system (used the system coefficients of the speech vowels). Better results may be obtained by modifying the system parameters also.

5.7 Summary

In this chapter, a method for synthesizing laughter, by modeling and incorporating the feature variations is described. Pitch period (T_0), strength of excitation (α), breathiness and call intercall durations of speech vowels are modified for synthesizing laughter calls. An experiment was conducted for estimating the perceptual significance of features. The experiment indicated that pitch period contour is the most significant factor, followed by breathiness, call intercall duration and strength of excitation. Perceptual evaluation was

also conducted on the synthesized laughter and the results were discussed. The quality of the synthesis can be improved further by also modifying the system coefficients along with the source.

Chapter 6

Discussions and Conclusion

Laugh signals have been mostly analyzed in the literature using traditional spectral features like formants, MFCC's, LPCC's etc. The phonetics and the prosodic characteristics of laughter has also been studied thoroughly. However, it is hypothesized in this work that the source plays a significant role in the production of laughter and hence the analysis is focused on understanding the source variations during laughter. The source features used for the analysis are (a) pitch period, (b) strength of excitation, (c) breathiness, (d) loudness and their variants. Pitch period and strength of excitation at epoch are obtained from adaptive zero-frequency analysis of the signal. Breathiness and loudness are derived from the Hilbert envelope of the linear prediction residual.

Distinct patterns are observed in the pitch period contour and strength of excitation contours of laugh signals. The general pattern that is observed in the pitch period within a call is that it starts with some value, decreases slightly to some minimum, and then increases rapidly to a high value. Similarly, the strength of excitation increases non-linearly to some maximum value and then decreases almost at the same rate. Also because of high amount of air flow through the vocal tract, laugh signals are typically accompanied by some amount of breathiness. Breathiness is shown to result in high non-deterministic component and less loudness in the signal. Using these features, a method was proposed for detecting laughter in spontaneous speech and the performance was tested on two different data sets.

A method was also proposed for automatically synthesizing laughter by modeling the features. An experiment was conducted for estimating the perceptual significance of features. The experiment indicated that pitch period contour was the most significant factor, followed by breathiness, call intercall duration and strength of excitation. Perceptual evaluation was conducted on the synthesized laughter and the results were discussed.

6.1 Contributions of the Work

1. Improved an already existing zero-frequency based epoch extraction method for handling the cases of large pitch period variations, by adaptively choosing the window for trend removal.

2. Proposed features based on pitch period, strength of excitation, breathiness and loudness for analyzing laugh signals.

3. Proposed a new feature based on hilbert envelope for detecting breathiness in a signal.

4. An algorithm for detecting laughter in continuous speech was proposed and it was tested on two different data sets.

5. Modeled the feature variations and synthesized laughter using these developed models.

6. Performed an experiment for deriving the perceptual significance of features.

6.2 Scope for future Work

1. A more detailed study can be done by considering various types of laughter.

2. The performance of the proposed laughter spotting algorithm can be studied on data with different noise levels and collected with different channels.

3. The performance of the system can be improved by taking the joint probability of all the features instead of deciding on each feature individually.

4. During laughter synthesis, only the source was modified while the system remained same. A more natural synthesis can be obtained by modifying the system along with the source.

5. It was showed earlier that the speaker has very less control on the articulators while producing laughter. So, it can be hypothesized that laughter cannot be mimicked like normal speech. A study can be made on analyzing the speaker specific characteristics of laughter which can be used for speaker verification etc.

References

[1] Trouvain, J, "Segmenting phonetic units in laughter," in *Proc. 15th ICPhS*, Barcelona, 2003, pp. 2793–2796.

[2] Wallace Chafe, "The Phonetics of Laughter - A Linguistic Approach," in *Interdisciplinary Workshop on The Phonetics of Laughter*, Saarbrucken, Aug. 4-5 2007.

[3] Grammer, K. and Eibl-Eibesfeldt, I., "The ritualization of laughter," in *Naturlichkeit der Sprache und der Kultur: Acta colloquii*, vol. edited by W. Koch Brockmeyer, Bochum, Germany, , 1990, p. 192214.

[4] Bachorowski J., Smoski M and Owren M., "The acoustic features of human laughter," *J. Acoust. Soc. Am.*, vol. 111, pp. 1582–1597, 2001.

[5] K. Sri Rama Murthy and B. Yegnanarayana., "Epoch Extraction from Speech Signals," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[6] Plato, "Laughter," in *Laws*, vol. 11, 1981, pp. 935–936.

[7] Aristotle, "Laughter," in *Nicomachean Ethics*, vol. 4, p. 1128.

[8] Thomas Hobbes, "Laughter," *Human Nature*, vol. 6, p. 46, 1840.

[9] Albert Rapp, "The Origins of Wit and Humor," New York, 1951, p. 21.

[10] Voltaire, *Preface to L'Enfant Prodigue*, 1829.

[11] Immanuel Kant tr. by J. H. Bernard, "Kritik of Judgment, London," , p. 223, 1892.

[12] Arthur Schopenhauer tr. by R. B. Haldane and J. Kemp, "The World as Will and Idea, London," , vol. 1, p. 76, 1964.

[13] Spencer. H, "On the physiology of laughter," *Essays on Education*, 1911.

[14] Dewey. J, "The theory of emotion," *The Psychological Review*, vol. 1, pp. 553–569, 1894.

[15] Freud. S tr. by J. Strachey, "Jokes and Their Relation to the Unconscious," *Penguin*, 1976.

[16] John Morreall, "A NEW THEORY OF LAUGHTER," 1981.

[17] Filippelli M, Pellegrino R, Iandelli I, Misuri G, Rodarte JR, Duranti R, Brusasco V and Scano G, "Respiratory dynamics during laughter," *J Appl Physiol*, vol. 4, no. 90, pp. 1441–6, April 2001.

[18] Citardi MJ, Yanagisawa E and Estill J., "Videoendoscopic analysis of laryngeal function during laughter," *Ann Otol Rhinol Laryngol.*, vol. 7, no. 105, pp. 545–9, July 1996.

[19] SC Gandevia, JE Butler, JL Taylor, A Anand and A Paintal, "No laughing matter," *The Lancet*, vol. 354, no. 9195, p. 2086, Dec 1999.

[20] S Overeem, GJ Lammers MD and JG van Dijk MD, "Weak with laughter," *The Lancet*, vol. 354, no. 9181, p. 838, Sep 1999.

[21] Luschei, E.S., Ramig, L.O., Finnegan, E.M., Baker, K.K. and Smith, M.E., "Patterns of laryngeal electromyography and the activity of the respiratory system during spontaneous laughter," no. 96, 2006, pp. 442–450.

[22] R. Provine, "Laughter," in *American Scientist*, 1996, pp. 38–45.

[23] C. Bickley and S. Hannicutt, "Acoustic analysis of laughter," in *Proceedings of the International Conference on Spoken Language Processing*, 1992.

[24] R. Provine, "Laughter: A Scientific Investigation," Penguin, New York, 2000.

[25] Menezes Caroline and Yosuke Igarashi, "The speech laugh spectrum," in *Proceedings of the 6th International Seminar on Speech Production (ISSP)*, Dec. 13-15 2006, pp. 157–524.

[26] R. Cai, L. Lu, H. Zhang and L. Cai, "Highlight sound effects detection in audio stream," *IEEE ICME*, 2003.

[27] Truong K.P. and Van Leeuwen D.A., "Automatic detection of laughter," in *Proc. of the INTERSPEECH 2005*, Lisbon, Portugal, 2005, pp. 485–488.

[28] L. Kennedy and D. Ellis, "Laughter Detection in Meetings," in *Proc. ICASSP Meeting Recognition Workshop*, Montreal, Canada, 2004.

[29] A. Carter, "Automatic acoustic laughter detection," MS Thesis, Keele University, 2000, submitted.

[30] Khiet P. Truong and David A. van Leeuwen, "Evaluating laughter segmentation in meetings with acoustic and acoustic-phonetic features," in *Speech Communication*, Workshop on the Phonetics of Laughter, 2007.

[31] Sundaram, Shiva and Narayanan, Shrikanth, "Automatic acoustic synthesis of human-like laughter," *J. Acoust. Soc. Am.*, vol. 121, p. 527, 2007.

[32] Toshiaki Haga, Masaaki Honda and Katsuhiko Shirai, "Acoustic analysis and synthesis of laughter," *J. Acoust. Soc. Am.*, vol. 120, no. 5, p. 3374, Nov 2006.

[33] Trouvain, J. and Schroeder, M., "How (Not) to Add Laughter to Synthetic Speech," in *Proceedings of the Workshop on Affective Dialogue Systems (ADS)*, Kloster Irsee, 2004, pp. 229–232.

[34] E. Lasarcyk and J. Trouvain, "Imitating conversational laughter with an articulatory speech synthesis," in *Proceedings of the Interdisciplinary Workshop on The Phonetics of Laughter*, Saarbruken, Germany, 2007, pp. 43–48.

[35] Eva Lasarcyk and Juergen Trouvain, "Spread lips + raised larynx + higher f0 = smiled speech? - an articulatory synthesis approach," in *ISSP*, Strasbourg, France, 2008, pp. 345–348.

[36] B. Yegnanarayana and K. S. R. 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.

[37] K. Sri Rama Murty, B. Yegnanarayana, Anand Joseph M., "Characterization of Glottal Activity from Speech Signals," *IEEE signal processing letters*, vol. 16, no. 6, June 2009.

[38] Kenneth N Stevens, *Acoustic Phonetics*. Cambridge: MIT Press, 1998.

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

[40] Ananthapadmanabha, T.V. and Yegnanarayana, B., "Epoch Extraction from linear prediction residual for identification of closed glottis interval," *IEEE Transactions on Speech and Audio Processing*, vol. 27, pp. 309–319, 1979.

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

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

[43] Carletta, J., "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus," *Language Resources and Evaluation Journal*, vol. 41(2), pp. 181–190, 2007.

List of Publications

Conferences

1. Sudheer Kumar K., Sri Harish Reddy M. , K. Sri Rama Murty and B. Yegnanarayana, Analysis of Laugh Signals for Detecting in Continuous Speech, in Proc. INTERSPEECH 2009, Brighton, UK, pp. 1591-1594.