Analysis for synthesis of nonverbal elements of speech communication based on excitation source information

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science (by research)
inElectronics and Communications Engineering

by

SATHYA ADITHYA THATI
200731009
sathya.adithya@research.iiit.ac.in

SPEECH AND VISION LAB
Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
December 2012
Copyright © Sathya Adithya Thati, 2012
All Rights Reserved
It is certified that the work contained in this thesis, titled “Analysis for synthesis of nonverbal elements of speech communication based on excitation source information” by SATHYA ADITHYA THATI, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Adviser: Prof. B. YEGNANARAYANA
To my PARENTS and TEACHERS
Acknowledgments

First and foremost, I would like to thank the Almighty and my Master for providing an inspiring guide, excellent facilities, and extremely encouraging research environment. I am grateful to Him for making me what I am now.

I would like to express sincere-most gratitude to my guide, Prof. B. Yegnanarayana, for having accepted me as his student. He has been continually motivating me and instilling research aptitude in me. Without his constant encouragement and guidance, I would not have achieved what I have now. His dedication and discipline have inspired me and enriched my growth as a student, researcher, and above all, a person.

I need to thank Dr. Kishore Prahallad for he has always been a motivator for hard work. I need to express gratitude to Prof. Peri Bhaskararao for sharing his immense knowledge with us and to have spent his invaluable time for enriching our knowledge. Thanks to Dr. Suryakanth Gangashetty for providing excellent infrastructure and research environment in the lab.

I would like to thank my colleagues (friends) and senior members Anand sir, Dhanu sir, Guru sir and Chetana ma’am for sharing their knowledge (experiences) on various aspects of life (both technical and non-technical). I would like to mention my dear friends Gomathi, Baji, RSP, Ronanki, Sudarsana, Vishala, Aneefa, Gangamohan and Nivedita for being there with me through good and bad, during success and failure, and during work and fun. They have provided encouragement during periods of distress. I thank all my past and present labmates and friends Sudheer, Rambabu, Karthik, Naresh, Anand swarup, Gautam, Basil, Apoorv, Abhijeet, Sivanand, Vasanth, Bhargav, Santosh, Padmini and Sreethar for being there and creating a friendly atmosphere in the lab. I still relish all the wonderful moments we had during our visits to IIT Guwahati and IISc. Bangalore.

I thank my batchmates Srikanth Vaddepally, Srikanth S, Gowtham Raghunath, Sunil, Naga Kartheek, Subbu, Varun, Anil, Gattu, and Kasikanth for sharing some wonderful moments together.

Special thanks to my special friends Raval and Mudita for constant encouragement and immense moral support extended during my research work.

Needless to mention the enormous support and endless love received from my family (parents, sister, uncle, grandparents). I owe my accomplishments to my parents.

Finally, I would like to dedicate this thesis to my parents, Balakrishna and Krishna Priya, and to my guide, Prof. B. Yegnanarayana.
Abstract

Speech communication is major medium of communication among human beings. Nonverbal elements of speech help in communicating the paralinguistic information such as emotions, attitudes, intentions, etc. along with message which is conveyed by the lexical part. They play a major role in conveying the unspoken message in the speech signal.

Voice quality is one such nonverbal element which helps in communicating the paralinguistic information. Voice quality (phonation type) also serves a linguistic function in various languages across the world. It also plays a role in assessment of various voice disorders. Laughter is a nonverbal vocalization which is produced and used by human beings in speech communication. It appears in natural conversation in everyday speech and provides naturalness to conversational speech. Synthesizing laughter helps in improving the expressiveness of speech synthesis.

In this work, analysis and synthesis of breathy voice and laughter has been performed. Analysis has been performed based on excitation characteristics, contrary to general spectral and spectrographic methods.

Features such as instantaneous fundamental frequency, strength of excitation, spectral tilt, periodic to aperiodic energy ratio and perceived loudness measure have been computed for analysis of breathy voice. Comparisons have been made between values of these parameters derived from the breathy and modal voiced speech segments. Classification experiments have been performed to discriminate breathy voice and modal voice from each other using periodic to aperiodic energy ratio and loudness measure, which proved to be discriminating the voice qualities successfully. Also, some approaches based on modifying the excitation source have been employed to synthesize breathy voice.

Laugh signals have been analyzed to study the pattern and structure of the signal and the source features derived from it. The contours and range of instantaneous fundamental frequency and strength of excitation have been studied and modeled to synthesize laugh signal from a vowel segment. Other features such as presence of frication like noise within laughter, duration and intensity have also been analyzed. Perceptual significance of these features have been studied. Suitable modifications have been made in source characteristics derived from a vowel segment to synthesize a laugh signal. Subjective evaluation has been conducted to gauge the quality and the level of acceptance of the synthesized laugh signals. Scores indicate that synthesized laughter was in the acceptable range without a compromise in the quality of synthesis.
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Introduction</td>
</tr>
<tr>
<td>1.1</td>
<td>Motivation</td>
</tr>
<tr>
<td>1.2</td>
<td>Objective and scope of the thesis</td>
</tr>
<tr>
<td>1.3</td>
<td>Organization of the thesis</td>
</tr>
<tr>
<td>2</td>
<td>Review of breathy voice and laughter</td>
</tr>
<tr>
<td>2.1</td>
<td>Breathy voice</td>
</tr>
<tr>
<td>2.1.1</td>
<td>Introduction to breathy voice</td>
</tr>
<tr>
<td>2.1.2</td>
<td>Production mechanism of breathy phonation</td>
</tr>
<tr>
<td>2.1.3</td>
<td>Functions of breathy voice</td>
</tr>
<tr>
<td>2.1.4</td>
<td>Previous studies</td>
</tr>
<tr>
<td>2.2</td>
<td>Laughter</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Introduction to laughter</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Production mechanism of laughter</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Previous studies</td>
</tr>
<tr>
<td>3</td>
<td>Analysis and synthesis of breathy voice</td>
</tr>
<tr>
<td>3.1</td>
<td>Data collection</td>
</tr>
<tr>
<td>3.2</td>
<td>Parameters/features for breathy voice</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Zero-frequency filtering technique</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Periodic-aperiodic energy computation</td>
</tr>
<tr>
<td>3.2.3</td>
<td>Formant extraction method</td>
</tr>
<tr>
<td>3.2.4</td>
<td>Measure of loudness</td>
</tr>
<tr>
<td>3.3</td>
<td>Results of Analysis</td>
</tr>
<tr>
<td>3.4</td>
<td>Classification experiments</td>
</tr>
<tr>
<td>3.5</td>
<td>Synthesis of breathy voice</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Increasing aperiodic component</td>
</tr>
<tr>
<td>3.5.2</td>
<td>Enhancing regions after the epoch in the glottal period</td>
</tr>
<tr>
<td>3.5.3</td>
<td>Adding frication like noise</td>
</tr>
<tr>
<td>3.5.4</td>
<td>Steps involved</td>
</tr>
<tr>
<td>3.6</td>
<td>Results of synthesis</td>
</tr>
<tr>
<td>3.7</td>
<td>Summary</td>
</tr>
</tbody>
</table>
4 Analysis and synthesis of laughter ................................................. 26
  4.1 Method to extract instantaneous fundamental frequency and strength of excitation at
      epochs ................................................................................ 26
  4.2 Analysis of laugh signals for synthesis ........................................... 28
      4.2.1 Pitch period .................................................................. 28
      4.2.2 Strength of excitation ...................................................... 29
      4.2.3 Duration ...................................................................... 30
      4.2.4 Frication ...................................................................... 30
  4.3 Synthesis of Laughter ................................................................. 31
      4.3.1 Incorporation of feature variations .................................... 31
        4.3.1.1 Pitch period modification ........................................... 31
        4.3.1.2 Modifying strength of excitation ................................. 31
        4.3.1.3 Incorporation of frication .......................................... 32
      4.3.2 Steps in the synthesis of laughter ........................................ 32
  4.4 Experiments ............................................................................. 35
      4.4.1 Perceptual significance of features .................................... 35
      4.4.2 Speaker identification from laughter .................................. 37
      4.4.3 Understanding the significance of source and system parameters 38
  4.5 Results of synthesis ................................................................. 38
  4.6 Summary ................................................................................. 39

5 Summary and conclusions ............................................................... 40
  5.1 Summary of the work ................................................................. 40
  5.2 Major contributions of the thesis ................................................. 41
  5.3 Scope for future work ............................................................... 41
  5.4 Journals ................................................................................. 42
  5.5 Conferences ........................................................................... 42

Bibliography .................................................................................. 43
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Illustration of source-filter model of speech production (color online)</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Illustration of cross sectional view of a human speech production system</td>
<td>2</td>
</tr>
<tr>
<td>2.1</td>
<td>Laryngeal parameters in the articulatory description of phonation types [1]</td>
<td>5</td>
</tr>
<tr>
<td>2.2</td>
<td>Glottal configurations for various phonation types: (a) Glottal Stop, (b) Creak, (c) Creaky voice, (d) Modal voice, (e) Breathy voice, (f) Whisper, (g) Voicelessness</td>
<td>6</td>
</tr>
<tr>
<td>3.1</td>
<td>Modal voiced syllable (/bi/)’s (a) Waveform, (b) Instantaneous fundamental frequency and (c) Strength of Excitation</td>
<td>14</td>
</tr>
<tr>
<td>3.2</td>
<td>Breathy voiced syllable (/bø/)’s (a) Waveform, (b) Instantaneous fundamental frequency and (c) Strength of Excitation</td>
<td>15</td>
</tr>
<tr>
<td>3.3</td>
<td>PAP for modal voice (/bi/) : (a) Waveform, (b) Periodic energy, (c) Aperiodic energy and (d) PAP ratio</td>
<td>16</td>
</tr>
<tr>
<td>3.4</td>
<td>PAP for breathy voice (/bø/) : (a) Waveform, (b) Periodic energy, (c) Aperiodic energy and (d) PAP ratio</td>
<td>17</td>
</tr>
<tr>
<td>3.5</td>
<td>Spectral Tilt of modal (/bo/) and breathy (/bø/) voices : (a) A1-A2 for modal voice, (b) A1-A2 for breathy voice, (c) A1-A3 for modal voice and (d) A1-A3 for breathy voice</td>
<td>18</td>
</tr>
<tr>
<td>3.6</td>
<td>Breathy voiced syllable (/bø/)’s (a) Waveform, (b) LP residual and (c) Hilbert envelope of LP residual</td>
<td>19</td>
</tr>
<tr>
<td>3.7</td>
<td>Modal voiced syllable (/bi/)’s : (a) Waveform, (b) LP residual, (c) Hilbert envelope of LP residual</td>
<td>20</td>
</tr>
<tr>
<td>3.8</td>
<td>Illustration of increased aperiodic component: (a) Modal voiced speech signal, (b) Periodic component of modal voiced speech signal, (c) Aperiodic component of modal voiced speech signal, and (d) Modified speech signal after modifying aperiodicity</td>
<td>22</td>
</tr>
<tr>
<td>3.9</td>
<td>Illustration of a modified LP residual: (a) Modal voiced speech signal, (b) LP residual of modal voiced speech signal, (c) Hilbert envelope of LP residual, (d) Modified LP residual and (e) Hilbert envelope of modified residual</td>
<td>25</td>
</tr>
<tr>
<td>4.1</td>
<td>(a) A segment of speech signal. (b) zero-frequency filtered signal using a window length of 3 ms for trend removal. (c) Voiced/nonvoiced decision based on ZFF. (d) Filtered signal obtained with adaptive window length for trend removal. (e) Strength of excitation (SoE). (f) Pitch period ($T_0$) obtained from epoch locations</td>
<td>27</td>
</tr>
<tr>
<td>4.2</td>
<td>(a) Spectrogram of laugh signal (b) A segment of laugh signal. (c) Pitch period derived from the epoch locations. (d) Strength of excitation (SoE) at the epochs</td>
<td>29</td>
</tr>
</tbody>
</table>
4.3 Illustration 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 shown using dashed line and modeled one is shown using dotted line (Color online). .................................................. 30
4.4 Block diagram of the laughter synthesis system (Color online). ............................ 32
4.5 Illustration of synthesized laugh signal: (a) Desired strength of excitation (SoE) contour (b) Desired pitch period ($T_0$) contour (c) Synthesized laugh signal. (d) Spectrogram of the synthesized laugh signal. ................................................................. 35
### List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Table showing the H1, H2 and H1-H2 values for breathy and modal sounds at different instances (in dB).</td>
<td>16</td>
</tr>
<tr>
<td>3.2</td>
<td>Table showing the mean loudness values for breathy and modal vowels.</td>
<td>18</td>
</tr>
<tr>
<td>3.3</td>
<td>Table showing the mean values of the parameters for breathy and modal sounds.</td>
<td>19</td>
</tr>
<tr>
<td>4.1</td>
<td>Parameters and preferred range of values for laughter synthesis.</td>
<td>34</td>
</tr>
<tr>
<td>4.2</td>
<td>Perceptual evaluation scores obtained for the modified versions of an original laugh signal.</td>
<td>36</td>
</tr>
<tr>
<td>4.3</td>
<td>Results of the experiment on perceptual significance of different features.</td>
<td>37</td>
</tr>
<tr>
<td>4.4</td>
<td>Results showing the difference in perceptual significance experiment scores.</td>
<td>37</td>
</tr>
<tr>
<td>4.5</td>
<td>Performance of laughter synthesis system in terms of MOS.</td>
<td>39</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Speech is the major medium of communication among human beings. The physiology behind the production of speech is complex and is highly sophisticated. The movement of articulators in different ways and the co-articulation effects result in producing different kinds of sounds in different ways.

Production of speech in a human being can be represented as a source-filter model, i.e., as a combination of a sound source (vocal folds) and a linear acoustic filter (vocal tract). The source excites the vocal tract system to produce speech sounds. While the shape of the vocal tract system characterizes what sound to produce, the excitation source controls how it is produced. The excitation source plays a major role in controlling the quality of voice, emotion in the speech uttered, and the pitch. The text part of the message in speech is shaped by the movement of the vocal tract system, whereas the intentions and other paralinguistic characteristics associated with the message are conveyed by the excitation source.

Figure 1.1 shows a representation of a source-filter model of speech production.

![Source-Filter Model](image)

Figure 1.1 Illustration of source-filter model of speech production (color online)

During human speech production, air is released from lungs, and due to higher sub glottal pressure, it is released from vocal folds and passed through the vocal tract system to come out of the mouth. The variations in air pressure are perceived as speech. The air passing through the larynx gets modulated by
the vocal folds before passing through the oral/nasal cavity, which act as a tube. Figure 1.2 shows the cross sectional view of a human speech production system.

During speech production, the vocal folds vibrate in one of the several possible modes of vibration, which vary according to how closely together the folds are held [2]. The manner in which the vocal folds vibrate is referred to as phonation. These modes of vibration forms a continuum, but can be categorized into five major phonation types, namely, breathy, slack, modal, stiff and creaky, with breathy phonation being the most open setting of the vocal fold vibration, and creaky phonation being the most constricted setting of vibration [2]. The difference in the production of speech of different voice qualities lies in the way the vocal tract system is excited. Laryngeal configuration differs for each phonation type (voice quality). The excitation source plays a prominent role in the production of speech for deciding the voice quality. Breathiness is an aspect of voice quality that is difficult to analyze and synthesize, especially since its periodic and noise components are typically overlapping in frequency. The decomposition and manipulation of these two components is of importance in a variety of speech applications such as text-to-speech synthesis, speech encoding, and clinical assessment of disordered voices.

All sounds produced by the speech production mechanism are not speech (lexical). There are certain nonverbal vocalizations produced and used by a human being in speech communication. Examples of nonverbal vocalizations include laughs, clicks, coughing, whistling, screams, etc. Some of them play a major role in speech communication and convey the unspoken message in the speech signal. Each of them play their respective roles in conveying different kinds of paralinguistic information. They all appear in natural conversation in everyday speech. They convey information about the environment/scenario in which the speech is produced. They provide naturalness to conversational speech.
Laughter is a common phenomenon in natural conversation. It plays a major role in day-to-day conversations. Like all nonverbal elements of speech in communication, laughter also has a unique pattern/structure to it and involves highly varying and complex production mechanism. The source of excitation plays an important role in the production of laughter, as air is released with high sub-glottal pressure.

1.1 Motivation

There is a need to understand what kind of information nonverbal elements carry, and how to include that information in synthesized speech for naturalness. Thus there is a need for analysis and synthesis of nonverbal sounds.

We need to develop signal processing methods/techniques to analyze and synthesize such signals. Characterization of the excitation source features has great potential for use in speech analysis, synthesis, and diagnosis of voice disorders.

Analysis and synthesis of two different types of nonverbal means of speech communication are considered in this work. They are breathy and laughter sounds. Since production mechanism for both the types of sounds are not normal, additional vocal effort is needed to produce them. Excitation source is different from normal for these sounds, and it plays a major role in understanding these sounds. Thus analysis of excitation source information is necessary. Also, analysis is necessary for identifying prominent features to be retained or modified for synthesis. Major emphasis in this work is to explore and exploit excitation source characteristics of such non-normal events in speech, as much of the previous work in this direction was through spectral and spectrographic analysis.

1.2 Objective and scope of the thesis

One kind of voice quality and one kind of nonverbal vocalization produced in speech are chosen for analysis, and to synthesize. Although this thesis considers analysis and synthesis of breathy voice (voice quality/phonation type) and laughter (nonverbal vocalization), the approaches are more general in nature. The analysis is performed based on excitation characteristics.

The difference in the production of a breathy voice and a modal voice lies in the way the vocal tract system is excited. Breathy voice consists of a more open laryngeal configuration compared to modal voice. Thus there is variation in the way the vocal tract system is excited.

There is a rapid movement in the excitation source articulators during the process of laughter production. This rapid movement causes rapid variations in the excitation features as opposed to typical excitation process that takes place. These rapid variations in the features are observed in laugh signals and their patterns are analyzed. By making suitable changes in these features for a vowel segment of speech, laugh signals are synthesized.
1.3 Organization of the thesis

The thesis is organized into five chapters. The first chapter gives the basic description and background of the work being addressed.

Chapter 2 gives a brief review on the breathy voice and laughter. Background information and literature survey of breathy voice and laughter are discussed in this chapter. Previous research work done on these sounds are summarized in this chapter.

Chapter 3 gives the analysis of breathy voiced speech, and describes signal processing methods employed to extract features used to characterize breathy voice. Features such as instantaneous fundamental frequency ($F_0$), Strength of Excitation (SoE) are extracted using zero-frequency filtering technique. Ratio of periodic to aperiodic energies have been computed by iteratively decomposing speech signal into periodic and aperiodic components. An objective measure for perceived loudness is used to measure the abruptness of the glottal closure in a vibration cycle. Spectral tilt is also used for both breathy and modal voices. Techniques to add breathiness to a modal voice are discussed in this chapter.

Chapter 4 gives the analysis of laugh signals and the procedure for synthesizing laughter. Features such as pitch period, strength of excitation and their contours, durations, etc. are analyzed in laugh signals. These features are modified and incorporated within the desired parametric range for a speech vowel to synthesize a laughter bout. Experiments are conducted to evaluate the perceptual significance of features used in the laughter synthesis system.

Chapter 5 summarizes the contributions of the work and highlights some issues arising out of the studies made. Also discussed in the chapter are some possible extensions of the studies reported in this thesis.
Chapter 2

Review of breathy voice and laughter

2.1 Breathy voice

2.1.1 Introduction to breathy voice

During production of speech, the variations in the manner of vibration result in different nonmodal phonations, which in turn result in respective voice qualities such as breathy voice, creaky voice, etc. The term ‘phonation’ refers to the manner of vibration of vocal folds. Modal phonation is the typical phonation practised by humans in normal conditions. While modal voice is produced by regular vibrations of the vocal folds at any frequency within the speaker’s normal range, in breathy voice vocal folds vibrate without appreciable contact, with arytenoid cartilages farther apart than in modal voice and with higher rate of airflow than in modal voice [2].

![Figure 2.1 Laryngeal parameters in the articulatory description of phonation types [1]](image_url)
2.1.2 Production mechanism of breathy phonation

Human beings can produce speech sounds with not only regular voicing vibrations at a range of different pitch frequencies, but also with a variety of voice source characteristics reflecting different voice qualities.

The voice is controlled by different types of muscular tensions, namely, adductive tension, medial compression and longitudinal tension [3]. Adductive tension is controlled by interarytenoid muscles and draws the arytenoids together. Medial compression is controlled by the lateral cricoarytenoid muscles and keeps the ligamental glottis closed. Similarly, longitudinal tension is mediated primarily by muscles of the vocal folds and the cricothyroid muscles. The contraction of the cricoarytenoid muscles can also increase the longitudinal tension by tilting the arytenoid cartilages backwards. Figure 2.1 shows the laryngeal parameters in the articulatory description of phonation types [4]. Figure 2.2 shows the glottal configurations of various phonation types.

Breathy voice can be produced by maintaining an open glottis for most of the vibration cycle, and by vocal folds closing more slowly than for modal phonation. Breathiness is characterized by low adductive tension and moderate to high medial compression. A triangular opening between the arytenoid cartilages produces a breathy voiced phonation. It is characterized by vocal folds having little longitudinal tension. This results in some turbulent air flow through the glottis. Thus we have the auditory impression of “voice mixed in with breath” [5].

Breathy voice and whisper are different. They differ in the manner of production. In breathy voice the vocal muscle tension is low and there is voicing, whereas in whisper the vocal muscle tension is high and there is no voicing involved. This can be observed from the glottal configurations, shown in Figure 2.2 (e) and (f), for breathy voice and whisper, respectively.

From an articulatory perspective, breathy voice is a different type of phonation from aspiration. However, breathy voiced and aspirated stops are acoustically similar in that in both cases there is an audible period of breathiness following the stop.
2.1.3 Functions of breathy voice

Breathy voice serves a linguistic function in various languages across the world such as Persian, Hindi, Gujarati, Marathi, Jalapa Mazatec, Tagalog, French, Italian, Nepali, Khmer, etc [5, 6, 7, 8, 9, 10]. It serves a contrastive property of vowels in various languages and of consonants in some languages. Breathy phonation has been consistently associated with lowered tone in many languages [11]. The most prominent and consistent cue to the aspirated affricates in Nepali language is breathy voice of the following vowel. Preliminary study of the Nepali affricates /tsh/ and /dzh/ revealed that both are distinguished from their nonaspirated counterparts /ts/ and /dz/ by breathy voice on the following vowel [8]. Perceptual effect of glottal fricatives have been studied in Persian [10]. Gujarati is one of the languages known for distinguishing breathy and modal phonation in both consonants and vowels, as in the words: bAr meaning ‘twelve’; bAr meaning ‘burden’; and b¨Ar meaning ‘outside’ – where b is a breathy voiced consonant and ¨A is a breathy voiced vowel. The superscript (¨) is used as a notation to represent the breathy variant of a vowel or a consonant in this draft.

Breathy voice is also employed to convey paralinguistic information, such as intentions, attitudes, emotions, etc. For example, breathiness has been associated with “sadness” [12]. It is used in expressing “disappointment” in Japanese [13, 14]. Breathiness and whispery voices were reported to be present in laughs (both funny and forced), surprise, embarrassment, politeness, gentleness (tenderness) and in other such paralinguistic information/elements present in speech [15]. Relationships between phonation types and paralinguistic information are reported in [12, 16]. Although prosodic features, like \( F_0 \), power, duration, have important roles in carrying paralinguistic information, variations in voice qualities (VQ) are commonly observed, mainly in expressive speech utterances. It is used in improving expressiveness in speech synthesis [17, 18].

Breathy voice is found in pathological speech as in Parkinson’s disease, dysphonia and dysarthria [9, 19, 20, 21], where it is believed to be due to glottal leakage of air. When the part of the brain that control speech production is damaged, the link from the brain to the muscles of speech is affected, which may result in vocal folds being uncoordinated or immobile. If the vocal folds cannot come together properly, then air can escape between them causing croaky (hoarse) or breathy speech. It can also be an inherent (natural) voice quality for some human beings as well. Controlling noise parameters may allow clinicians to modify disordered (breathy) voices and estimate improvements in speech acoustics after patients undergo voice therapy or surgery [22].

2.1.4 Previous studies

Acoustic measures of breathy voice are not often explicitly described in the literature. Analysis of breathy voice was mostly influenced by the spectrum analysis and spectrographic methods. Some source features were measured through spectrum analysis. This involved computation of features such as fundamental frequency (\( F_0 \)), formant frequencies, acoustic intensity, periodicity, additive noise and spectral tilt [6, 23]. Open quotient (OQ), i.e., duration of open phase in total pitch cycle, was observed
to be higher in breathy speech. Since, there is little closed phase, or less closed phase abruption, it leads to a steeper spectral slope.

Breathiness is thought to be due to incomplete and nonsimultaneous glottal closure during the ‘closed’ phase of the phonatory cycle [19, 24, 25, 26, 27, 28]. Breathy glottal source signals obtained through inverse filtering typically show more symmetrical opening and closing phases with little or no complete closed phase [29, 30, 31]. The near-sinusoidal shape of breathy glottal waveforms is responsible for a relatively high amplitude of the first harmonic (H1) and relatively weak upper harmonics [19, 26, 28, 30, 31]. The differences in the amplitudes of the first two harmonics (H1-H2) was seen to correlate with open quotient (the percentage of glottal vibration cycle for which the glottis is open) [32]. Enhanced H1 amplitude in the spectra of breathy voice signals has been observed by a number of investigators [26, 29, 30, 31, 33, 34, 35].

The relatively more symmetrical or near-sinusoidal shape of breathy glottal waveform does not only boost the lower harmonics, but also is responsible for a decrease in the amplitude of the harmonics in the higher frequency region, or degree of spectral tilt. The more symmetrical the glottal pulse, the steeper the spectral tilt [36]. The difference in the amplitude of first harmonic and the amplitudes of the first three formant frequencies (H1-A1 (an indicative measure of bandwidth of first formant), H1-A2, H1-A3) [5, 7, 37] has been used to measure spectral tilt. Normalized Amplitude Quotient (NAQ) of the glottal waveform and its derivative waveform [38] characterizes the spectral slope properties of the breathy voice.

When a portion of the air stream from the lungs passes through a persistent and relatively narrow glottal chink during the production of breathy vowels, this results in the generation of noise [26, 28, 39]. The spectrum becomes dominated by dense aspiration noise, particularly at high frequencies where noise may actually replace harmonic excitation of the third and higher formants [40, 41]. To isolate and estimate the relative strength of noise components of samples, Klatt and Klatt [26] used a bandpass filter centered at $F_3$. Another method for calculating a spectral harmonics-to-noise ratio (HNR) in speech signals was proposed by Krom [42]. This harmonics-to-noise ratio algorithm used a comb-filter defined in the cepstral domain to separate the harmonics from the noise. Sensitivity of Krom’s HNR to both noise and jitter made it a valid method for determining the amount of spectral noise.

Several other features to reflect the effects of aspiration noise were also proposed. Less periodic signals such as those often produced in breathy phonation have a spectrum with less definite harmonics, resulting in a cepstrum with a low peak at the pitch period. This method is unreliable though, when there are rapid pitch changes and where vocal folds of a modal vowel happen to be vibrating irregularly. Cepstral peak prominence (CPP), a measure of amplitude of the cepstral peak corresponding to the fundamental period, normalized for overall signal amplitude [28], Glottal to noise excitation ratio (GNE) [43, 44], and Harmonics to noise ratio (HNR) reflect the presence of aspiration noise components in breathy voice. A synchronization measure between the amplitude envelopes of the first and third formant frequency band signals (F1F3syn) is reported in [15], and Normalized breathiness power measure (NBP), which is calculated based on F1F3syn, was used to characterize the amount of breathi-
ness present in a signal [45]. Jitter, shimmer and higher order statistics (HOS) properties (like skewness and kurtosis of the data samples) were computed in [46]. A few new indexes like harmonic energy of residue, harmonic to signal ratio, and number of voiced frames in a segment were also used in [20] for characterizing breathiness.

Some attempts to synthesize breathy vowels have been made [22, 26, 28, 47]. Most of them deal with the addition of aspiration noise. In [47], a combination of lowpass-filtered pulses and synchronous highpass-filtered noise burst of equal energy was used as a source signal in a simple source-filter model. Quatieri et.al. stressed on the need for decomposition and manipulation of periodic and noise components, which is a difficult task to undertake, as they are typically overlapping in frequency. Envelope shaping has been applied to the noise source that was derived from the inverse-filtered noise component [22].

The difference in the production of a breathy voice and a modal voice lies in the way the vocal tract system is excited. Breathy voice consists of a more open laryngeal configuration compared to modal voice. Thus there is variation in the way the vocal tract system is excited. The excitation source thus plays a prominent role in production of speech for deciding the voice quality. Analysis based on excitation source attempts to take into account the timing information of the glottal activity. In this work, features based on source characteristics, derived using robust signal processing methods are proposed for analysis of breathy voices.

2.2 Laughter

2.2.1 Introduction to laughter

In natural human conversation, nonverbal vocalization plays a key role in expressing emotions. Laughter is one such vocalization that is mostly used to express joyous mood. It induces a positive emotive state on listeners. To a lesser extent, laughter is also used in other emotional contexts such as sarcasm, humiliation, etc, making it an important indicator of emotion/mood.

Laughter is categorized into 3 basic types: voiced “song-like” laughter, “snort-like” laughter with perceptually salient nasal turbulence, and “grunt-like” laughter with laryngeal and oral-cavity frication [48]. Although only about 30% of the analyzed laughs are predominantly voiced, they induce significantly more positive emotional responses in listeners than unvoiced laughs. Trouvain has segmented laughter at different levels (phrasal, syllabic, segmental, phonation and respiration) for understanding the structure of a typical laugh [49]. An instance of laughter is referred to as an ‘episode’. The segment of the laughter episode produced between two inhalation gaps is known as ‘bout’ or ‘laughter bout’. An entire laugh can have several bouts separated by inhalation [49]. The discrete acoustic events that together constitute a bout are called ‘calls’ [48]. Each call of a voiced laughter consists of voiced part followed by an unvoiced/silence part (inter-call interval). Each laughter bout contains several calls. Provine concluded that laughter is usually a series of short syllables repeated approximately every 210
ms [50]. Different acoustic descriptions of laughter have been used in the literature for different studies [48, 49, 51].

The main sound feature of a laughter is aspiration /h/. Laughter sounds as a sequence of syllables which are consonants followed by vowels (open mouthed laughter) or vocalic nasals (closed mouthed laughter) [52]. They are typically perceived as ha-ha-ha or hi-ha-ha sequence in the case of open-mouthed laughter. Here laughter causes jaw lowering, resulting in an /a/-colored sound for all vowel categories [53]. It sounds like a sequence of breathy CV syllables (/hV/) as in ha-ha-ha or heh-heh [51]. Bachorowski et.al. found that the vowel-like laughs generally contained central vowel sounds [48]. Ekman et.al. also mentioned that the laughter “vowel” is the central vowel schwa or /e/ [54].

2.2.2 Production mechanism of laughter

Speech production is a controlled process that is guided by a set of rules. The movement of articulators is defined by the sequence of subword units to be uttered. But unlike speech, there are no rules guiding the process of production of laughter. Laughter is typically produced by a series of sudden bursts of air, released by the lungs, keeping the vocal tract almost steady. Lungs and vocal folds (source of excitation) play a major role in laughter production. Due to high air pressure built up in the lungs, there is larger than normal air flow per unit time through the vocal tract. This results in rapid vibration of the vocal folds. Since vocal folds cannot maintain/sustain that unusual high pitch frequency, their vibration tends to decrease to reach the normal pitch frequency. There is also turbulence generated at the vocal folds which results in the signal being breathy (noisy), when compared to normal speech [53]. All this is for the production of a call. The process of call production repeats itself with certain inter-call variations to produce a bout.

2.2.3 Previous studies

Laughter has been analyzed using both source and system characteristics of production. Since laughter is produced by the human speech production mechanism, the laugh signal is also analyzed like a speech signal in terms of the acoustic features of speech production. Typically, the acoustic analysis of laughter is carried out using duration, fundamental frequency of voiced excitation ($F_0$), and spectral features [48, 50, 55]. Conventional methods of analysis were used to derive the features of the glottal vibration by Bachorowski [48] and Bickley [51]. Mostly spectrum-based features like harmonics, spectral tilt and formants were used to analyze laughter [51, 53, 56]. The importance of the acoustic structure of human laughter was discussed by Todt et.al.[55, 57]. Observations were also made on the number of calls per bout and number of bouts in a laughter episode. The problem of extracting rapidly varying instantaneous fundamental frequency ($F_0$) was addressed by Sudheer [58]. Sudheer et.al. measured the following features: (a) rapid changes in the instantaneous fundamental frequency ($F_0$) within a call, (b) strength of excitation (SoE) within each glottal cycle and its relation to $F_0$, and (c) temporal variability of $F_0$ and SoE across calls within a bout. These features were used for spotting laughter in continuous
speech [58]. Analysis at subsegmental level (< pitch period) captures the physiological characteristics of the excitation source.

Some attempts to synthesize laughter have also been made. There have been attempts to insert available laughter samples in speech for simulating natural conversation [59]. There have also been attempts to model laughter [60, 61, 62]. To insert laughter into conversational speech, laugh samples from a corpus were selected and incorporated in concatenative synthesis [59]. Trouvain and Shröeder superimposed the duration and $F_0$ of natural laughter samples onto recordings of diphones ('hehe') to generate laughter [60]. Results showed that careful control of the laugh intensity is required for better perception. An attempt to synthesize laughter has been made by Shiva Sundaram and Shrikanth Narayanan, using the principle of a damped simple harmonic motion of a mass-spring model to capture the overall temporal behaviour of laughter at episode level. The voicing pattern of laughter was seen as an oscillatory behaviour, and was observed in most laughter bouts. The behaviour of alternate voiced and unvoiced segments was modeled with equations that described the simple harmonic motion of a mass attached to the end of a spring [61]. An articulatory speech synthesizer was used by Lasarcyk and Trouvain to model laughter. A real laugh was taken from a spontaneous speech database, and synthetic versions of it were created. Features like breathing noises were also approximated, as they do not normally occur in speech. It was reported that synthesis taking into account the variations in durational patterns, intensity and $F_0$ contours resulted in better score for perception of naturalness [62].

In this work, a method is proposed for synthesis of laughter making use of the characteristics of the excitation source.
Chapter 3

Analysis and synthesis of breathy voice

In this chapter, methods to extract features which characterize breathy voice have been described. Features are analyzed and the values of the parameters obtained are compared with those of modal voice. Classification experiments are performed to investigate the significance of certain features. Some approaches to synthesize breathy voice are then mentioned.

3.1 Data collection

Initial data for the analysis of breathy voice was collected at a sampling frequency of 48kHz in a quiet room, and was recorded by an expert phonetician. Data was also recorded from 10 native Gujarati speakers (8 male and 2 female). The participants include those who have stayed most of the time in Gujarat, and those who haven’t been there but are natives of that language.

Syllables containing the breathy voice phonation in the vowel part were recorded by an expert phonetician. Some words containing the breathy phonation and their corresponding contrasting modal phonated words were recorded by the native Gujarati speakers. To ensure uniform prosodic effect, and to help the speakers to speak naturally, meaningful declarative carrier words and sentences were used.

3.2 Parameters/features for breathy voice

Following are the techniques used to compute the parameters for the analysis of the breathy voiced signals. These techniques are robust because they attempt to capture the acoustic properties of the actual speech production mechanism.

3.2.1 Zero-frequency filtering technique

A method is proposed for extraction of the instantaneous $F_0$, epoch extraction and strength of impulse-like excitation at epochs [4, 63]. 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 of impulse at the epochs.

A zero-frequency filtered (ZFF) signal is derived as follows:

(a) The speech signal $s[n]$ is differenced to remove unwanted very low frequency components.

$$x[n] = s[n] - s[n - 1]$$  \hfill (3.1)

(b) The differenced speech signal is passed through a cascade of zero-frequency resonators (digital resonators having poles at zero frequency) given by the following equation:

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

where $a_1 = -4$, $a_2 = 6$, $a_3 = -4$ and $a_4 = 1$

(c) The trend in $y_0[n]$ is removed by subtracting the mean computed over a window at each sample. The resulting signal $y[n]$ is the zero-frequency filtered signal, given by

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

where $(2N + 1)$ is the size of the window which is in the range of 1 to 1.5 times the average pitch period in samples.

The negative to positive zero crossing instants in the resulting zero frequency filtered (ZFF) output are called epochs. The slopes of the ZFF signal at epochs give the relative strengths of the impulse-like excitation (SoE) around epochs. The reciprocal of the interval between successive epochs gives the instantaneous fundamental frequency ($F_0$).

It is observed that the $F_0$ of a speaker is lowered during breathy phonation when compared to $F_0$ during modal voice as shown in Figures 3.1 and 3.2. The decrease in the overall fundamental frequency values can be attributed to the fact that during the breathy voice, there is a gap for air to flow through the vocal folds and this results in the slower vibrations of the vocal folds. Breathy phonation starts with lower $F_0$ and increases steeply in a short duration. The rise is as high as 20% in less than a period of 10 milliseconds.

We also observe from Figure 3.1 that there is a sudden rise in the SoE from the stop consonant to the vowel in the modal voice signal as anticipated, whereas we observe from Figure 3.2 that the transition in the SoE for breathy voice is gradual. This is because there is no abruptness in the glottal closure mechanism of a breathy phonation.
Figure 3.1 Modal voiced syllable (/bi/)’s (a) Waveform, (b) Instantaneous fundamental frequency and (c) Strength of Excitation.

3.2.2 Periodic-aperiodic energy computation

Breathy voiced speech consists of increased spectral noise, particularly at higher frequencies. This is due to persistent leakage of air through the glottis during breathy phonation. The ratio of the periodic and aperiodic energies (PAP) is used as a measure to reflect this property.

This approach to calculate PAP involves iterative decomposition of speech into periodic and aperiodic components as proposed in [64]. The method is summarized in the following steps:

(a) Linear prediction (LP) analysis is performed to compute the LP residual

(b) The LP residual is divided into frames of size 32 ms with a frame shift of 4 ms. It is checked for voiced and unvoiced frames.

(c) Cepstrum is computed using 512 point FFT and Hamming window. The peak in cepstrum relating to harmonics in spectrum is identified by using pitch information obtained by ZFF method (Section 3.2.1).

(d) The harmonic log spectrum is computed by making all the coefficients in cepstrum, except the 9 samples around the peak corresponding to the pitch period to zero, and IDFT is taken.

(e) The spectrum of the LP residual frame is computed. Samples from the spectrum are now divided into periodic and aperiodic parts.
Figure 3.2 Breathy voiced syllable (/b¨ı/)’s (a) Waveform, (b) Instantaneous fundamental frequency and (c) Strength of Excitation.

(f) An iterative algorithm is used to compute the aperiodic component of the residual. Periodic component is obtained by subtracting the aperiodic component from the residual of the speech signal.

(g) Periodic and aperiodic components of the speech signal are synthesized by exciting the all pole filter (LP synthesis) with the periodic and aperiodic components of the residual as excitation, respectively.

The ratio of energy of the periodic and aperiodic components \( \frac{E_p}{E_{ap}} \) computed over each of the frames in the voiced regions of the utterances is analyzed. Since the intensity of noise (aperiodicity) is higher at higher frequencies in breathy voice speech signal, we observe that the aperiodic energy is significantly high in breathy signals than that in modal signals, resulting in lesser PAP value for breathy vowels. Figures 3.3 and 3.4 show the speech signal, periodic energy, aperiodic energy and the ratio of periodic to aperiodic energies for modal and breathy voiced syllables respectively.

### 3.2.3 Formant extraction method

Spectral tilt is a measure of the degree to which intensity drops off as frequency increases. It is one of the acoustic parameters used to differentiate breathy phonation type from other phonation types. It is generally quantified by comparing the amplitudes of the first harmonic to that of higher frequency harmonics, which could be the second harmonic or the formant frequencies. Spectral tilt is observed to be more for breathy vowels, which means that there is larger fall off in energy at higher frequencies.
in the signal. The values of the measures used to define the spectral tilt (H1-H2, A1-A2 and A1-A3) are higher for breathy vowels compared to that for its modal counterpart. Locations of formants are computed using the group delay based method given in [65].

The computed H1-H2 values are shown in Table 3.1. Figure 3.5 shows plots of A1-A2 and A1-A3 of modal and breathy speech signals. It can be observed from the figures that spectral tilt is higher for breathy voice.

### Table 3.1 Table showing the H1, H2 and H1-H2 values for breathy and modal sounds at different instances (in dB).

<table>
<thead>
<tr>
<th>Breathy voice (/b¨a/)</th>
<th>Modal voice (/ba/)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>H2</td>
</tr>
<tr>
<td>-17</td>
<td>-30.7</td>
</tr>
<tr>
<td>-20.3</td>
<td>-29.3</td>
</tr>
<tr>
<td>-16.6</td>
<td>-28.9</td>
</tr>
</tbody>
</table>

### 3.2.4 Measure of loudness

Perceived loudness of speech is related to the abruptness of the glottal closure. In a breathy voice, the glottal closure is not so abrupt compared to modal voice, and hence the perceived loudness of breathy
speech is less compared to the perceived loudness of modal speech. This can be used as a measure to compare different voice qualities.

An objective measure ($\eta$) of perceived loudness based on the abruptness of glottal closure derived from the speech signal is discussed in [66]. The abruptness of the glottal closure derived from the EGG signal was shown to be high for loud speech compared to soft and normal speech. When the glottal closure is abrupt, the Hilbert envelope of the LP residual of the speech signal will have sharper peaks at the epochs. This sharpness of peaks for a modal voice can be observed in Figure 3.7.

Figure 3.6 illustrates the sharpness (bluntness) of the peaks in Hilbert envelope of an LP residual of a breathy voiced speech. The sharpness of the peaks in Figure 3.6 can be compared with the sharpness of peaks in Figure 3.7, and it can be observed that the peaks are not as sharp as they typically are for modal voiced speech.

The sharpness of the peaks in the Hilbert envelope at the epochs is derived by computing the ratio

$$\eta = \frac{\sigma}{\mu}$$  \hspace{1cm} (3.4)

Here $\mu$ denotes the mean, and $\sigma$ denotes the standard deviation of the samples of the Hilbert envelope of the LP residual in a short interval (2 ms) around the epochs.

Table 3.2 shows the means of the loudness values calculated for breathy and modal vowels.
Figure 3.5 Spectral Tilt of modal (/bo/) and breathy (/bʊ/) voices: (a) A1-A2 for modal voice, (b) A1-A2 for breathy voice, (c) A1-A3 for modal voice and (d) A1-A3 for breathy voice.

Table 3.2 Table showing the mean loudness values for breathy and modal vowels.

<table>
<thead>
<tr>
<th>Vowel</th>
<th>Breathy</th>
<th>Modal</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>0.51</td>
<td>0.70</td>
</tr>
<tr>
<td>e</td>
<td>0.53</td>
<td>0.72</td>
</tr>
<tr>
<td>i</td>
<td>0.53</td>
<td>0.69</td>
</tr>
<tr>
<td>o</td>
<td>0.53</td>
<td>0.72</td>
</tr>
<tr>
<td>u</td>
<td>0.52</td>
<td>0.67</td>
</tr>
</tbody>
</table>

3.3 Results of Analysis

Duration of the stop consonant preceding the breathy vowel is observed to be lesser than that for the modal voice. This is due to lower vowel onset time for the breathy speech. The reason attributed to this is that the speaker knows before hand that the succeeding phone is breathy, and thus his production mechanism is preset to that of a breathy voice. In this configuration, it is difficult for the speaker to utter the consonant for a longer duration. This initial setting of the production mechanism for the breathy sounds may also be one of the reasons behind the gradual transition of SoE in Figure 3.2.

The mean values of the various parameters computed for the breathy vowels and its modal counterpart are given in Table 3.3. In the case of a naturally breathy voice, the contrast between the breathy and modal voice is lesser than the contrast observed for a naturally non-breathy voice.
Figure 3.6 Breathy voiced syllable (/bë/)’s (a) Waveform, (b) LP residual and (c) Hilbert envelope of LP residual.

Table 3.3 Table showing the mean values of the parameters for breathy and modal sounds.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Breathy</th>
<th>Modal</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F_0$ (Hz)</td>
<td>114</td>
<td>123</td>
</tr>
<tr>
<td>SoE</td>
<td>0.0118</td>
<td>0.0101</td>
</tr>
<tr>
<td>PAP</td>
<td>9.93</td>
<td>19.00</td>
</tr>
<tr>
<td>Loudness</td>
<td>0.52</td>
<td>0.70</td>
</tr>
<tr>
<td>A1-A2 (dB)</td>
<td>2.61</td>
<td>2.21</td>
</tr>
<tr>
<td>A1-A3 (dB)</td>
<td>3.25</td>
<td>2.84</td>
</tr>
</tbody>
</table>

Both conventional features such as spectral tilt and new features like $PAP$, $\eta$, $SoE$ are used to describe the acoustic characteristics of breathy voice quality. These features can be used to spot breathiness in a speech signal. We observe that breathy voice is perceived to be less loud than the modal voice, and it is measured by the loudness measure. Due to higher amount of aperiodicity attached with breathy phonation, we see that the PAP ratio is less for breathy voice quality. The average $F_0$ is less and the strength of excitation is more for breathy voice. The spectral tilt is higher for breathy voice as confirmed by the measures of A1-A2 and A1-A3.
Figure 3.7 Modal voiced syllable (/be/)’s: (a) Waveform, (b) LP residual, (c) Hilbert envelope of LP residual.

3.4 Classification experiments

A few classification experiments are performed to evaluate the significance of some of the features used in the analysis. Features such as PAP and \( \eta \) are used to classify samples of breathy speech from modal ones and vice versa. Significant differences in the values of these parameters derived from breathy and modal voices have been used as cues to discriminate breathy voice from modal voice.

These experiments were performed using 105 test samples including both breathy and modal voices. (80 breathy, 25 modal)

Accuracy is the number of times a sample was identified correctly, i.e., breathy sample as ‘breathy’ and modal sample as ‘modal’.

False Alarm Rate is the number of times a test sample from the other class was identified as a sample from this class, i.e., modal sample as ‘breathy’ and breathy sample as ‘modal’.

Missed Detection Rate is the number of times a sample was not classified as belonging to that class, i.e., breathy sample not detected as ‘breathy’ and modal sample not detected as ‘modal’.

- Discriminating breathy vowels from modal vowels

(a) Using PAP:
Accuracy: 88.75%
False Alarm Rate: 11.25%
Missed Detection Rate: 11.25%

(b) Using $\eta$:
Accuracy: 95.18%
False Alarm Rate: 4.82%
Missed Detection Rate: 1.25%

• Discriminating modal vowels from breathy vowels

(a) Using PAP:
Accuracy: 64%
False Alarm Rate: 36%
Missed Detection Rate: 36%

(b) Using $\eta$:
Accuracy: 95.45%
False Alarm Rate: 4.55%
Missed Detection Rate: 16%

Overall accuracy in classification using PAP: 82.85%
Overall accuracy in classification using $\eta$: 95.23%

Best performance (accuracy) in classification is obtained using optimal threshold values for respective parameters.

### 3.5 Synthesis of breathy voice

An attempt to synthesize breathy voice is made by incorporating breathiness in a modal voice. Three different approaches are attempted to achieve this task. Major modifications are made on the source component of the modal speech signal to synthesize breathy speech signal. The following are the three approaches to incorporate breathiness in a modal voiced speech:

(a) Modifying the proportion of aperiodicity

(b) Enhancing regions after the epoch in glottal periods of LP residual

(c) Adding frication like noise
3.5.1 Increasing aperiodic component

A modal speech signal is taken and its periodic and aperiodic components are derived using the iterative decomposition method applied on the LP residual of the signal, as explained in Section 3.2.2. Since there is presence of higher aperiodicity in breathy voiced speech signals because of glottal leakage of air, an attempt to increase the aperiodicity is made so as to incorporate the effect of breathiness in a modal signal. This is done by increasing the proportion of aperiodic component of the residual of modal speech signal. After increasing the relative proportion of aperiodic component, it is combined/added with/to the periodic component to form a new/modified/desired breathy residual. It is then passed through an all-pole filter with LP coefficients as filter coefficients to obtain a new breathy speech signal. Figure 3.8 illustrates the modified speech signal after modifying its aperiodic component.

![Figure 3.8 Illustration of increased aperiodic component: (a) Modal voiced speech signal, (b) Periodic component of modal voiced speech signal, (c) Aperiodic component of modal voiced speech signal, and (d) Modified speech signal after modifying aperiodicity](image-url)
3.5.2 Enhancing regions after the epoch in the glottal period

It can be observed from Figure 3.6 and Figure 3.7 that regions between epoch locations (non-epoch regions) in the LP residual and Hilbert envelope of LP residual of breathy voiced signal are more noisy (higher in amplitude or higher variance) than similar regions of the LP residual and Hilbert envelope of LP residual, of modal voiced signal. This reflects the lesser abruptness during the glottal closure of production of breathy voice. Changes are made in the LP residual of the modal speech signal to modify the source characteristics of the signal.

A modal speech signal of a vowel/syllable is taken and LP analysis is performed to extract LP coefficients and LP residual. Hilbert envelope of the LP residual is computed. All the samples in the residual, except for a few samples around the peaks of Hilbert envelope of LP residual, are scaled up by a factor so as to decrease the relative dominance of the peak in the residual signal. Figure 3.9 illustrates the modified residual of a modal speech signal and its Hilbert envelope. The significance of the peaks is lesser in the modified residual compared to the original residual. This residual can be passed through LP filter to obtain a modified (breathy) speech signal.

3.5.3 Adding frication like noise

To generate the effect of aspiration noise, frication-like noise is added to the modal signal to make it sound breathy. White Gaussian noise is generated and is passed through a resonator with a center frequency at 2500 Hz and a bandwidth of 500 Hz. The resulting filtered noise is added to the LP residual of the modal speech signal in desired proportion to obtain a residual for desired breathy voiced speech signal. This residual could then be used to synthesize a breathy voiced speech.

All the above described methods can be jointly employed to create a better perception of breathiness in a signal.

3.5.4 Steps involved

1. A modal voiced speech signal is taken and its periodic and aperiodic components are derived using the method suggested in Section 3.2.2.

2. Speech signal is synthesized with higher aperiodicity, as explained in Section 3.5.1

3. LP analysis is performed on the modal speech signal to derive LP residual and LP coefficients.

4. Non-epoch regions in the LP residual are enhanced using the process explained in Section 3.5.2.

5. This residual is further modified by adding friction like noise to it as described in Section 3.5.3

6. This modified residual is passed through an all-pole filter to synthesize a speech signal.

7. Signals obtained in Step 2 and Step 6 are added to obtain desired breathy voiced speech signal.
3.6 Results of synthesis

In the attempt to synthesize breathy voice, noise-like features are successfully introduced into the modal voice. Informal listening suggests that breathiness is incorporated in the modal speech to a considerable extent. There is slight roughness added to the voice by the addition of noise-like features. Roughness or hoarseness has been generally associated with breathy voice. The most important factor to consider while adding noise-like aperiodic component/stream is to mix it well with the periodic component of the signal, so as to not sound like a separate stream perceptually. This constraint is well satisfied as the noise-like characteristics introduced are spread across the frequency domain rather than just into a particular frequency bin.

3.7 Summary

In this chapter, breathy voiced speech was analyzed and features such as $F_0$, SoE, PAP, Spectral tilt, and $\eta$ have been derived from the signal. Comparisons have been made between the values of parameters derived from the breathy and modal voiced speech segments. It is observed that the difference between the values of the parameters ‘PAP’ and ‘$\eta$’ is considerably higher. PAP and $\eta$ were able to discriminate one voice quality from the other with convincing results. Some approaches to incorporate breathiness in a modal voice have been described.
Figure 3.9 Illustration of a modified LP residual: (a) Modal voiced speech signal, (b) LP residual of modal voiced speech signal, (c) Hilbert envelope of LP residual, (d) Modified LP residual and (e) Hilbert envelope of modified residual
Chapter 4

Analysis and synthesis of laughter

Analysis of signals of natural laughter is needed to understand the characteristics of laughter at both call-level and bout-level. This will help to bring the synthesized laugh closer to a natural laugh, both at segmental and suprasegmental levels. In this chapter, a slightly modified version of the zero-frequency filtered technique, described in Chapter 3, is used to capture the rapidly varying features of laugh signals. The analysis involves the understanding of the patterns of various features in a call as well as across the calls. A segment of a vowel is then modified to follow such patterns in the process of synthesizing laugh signals. Experiments have been performed to assess the significance of various features in the synthesis of laughter.

4.1 Method to extract instantaneous fundamental frequency and strength of excitation at epochs

A method was proposed [4, 63] for extraction of the instantaneous $F_0$, epochs and strength of impulse-like excitation at epochs. 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 method involves passing the differenced speech signal through a cascade of two ideal digital resonators, each located at 0 Hz. The trend in the output is removed by subtracting the local mean at each sample, computed over a window length in the range of about 1 to 2 pitch periods. The negative to positive zero crossing instants in the resulting zero frequency filtered (ZFF) output are called epochs. The slopes of the ZFF signal at epochs give the relative strengths of the impulse-like excitation (SoE) around epochs. The reciprocal of the interval between successive epochs gives the instantaneous fundamental frequency ($F_0$).

This method described in Chapter 3 does not capture the rapid variations of $F_0$ that appear in the calls of a laughter episode. To capture the rapid variations of a laugh signal, the method was modified...
using the following steps to derive the epochs and their strengths from the zero frequency filtered (ZFF) signals [58].

1. Pass the signal through the zero-frequency resonator with a window length of 3 ms for trend removal. The ZFF signal has high energy in the regions of voiced speech and laughter, and low energy in the nonvoiced and silence regions.

2. Voiced and nonvoiced segments of the signal are determined using the ZFF signal. Samples of normalized ZFF signal are squared and their running mean over a window of 10 ms is calculated to estimate the envelope of the signal. It is then normalized by using the following equation

\[ s_2 = 1 - e^{-10s_1} \] (4.1)
where \( s_1 \) is the estimated envelope, and \( s_2 \) is the normalized envelope. The set of samples in \( s_2 \) having a value above the threshold of 0.3 are marked as voiced regions in the signal. The value 10 in Eq. 4.1 and the threshold 0.3 are determined based on study on large amount of speech data.

3. After finding the voiced segments, the signal in each voiced region is passed separately through a zero-frequency resonator and window length for trend removal is derived from that segment. The location of the maximum peak in the autocorrelation function of the segment is used to determine the window length for trend removal in that region. Due to rapid changes in the pitch period values, the window size for trend removal is chosen adaptively for each segment.

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 (slope) gives the strength of excitation.

The results at various stages to obtain the pitch contour and strength of excitation from a segment of speech signal are shown graphically in Figure 4.1. Speech signal and the ZFF signal obtained in the first step are plotted in Fig. 4.1(a) and Fig. 4.1(b) respectively. Fig. 4.1(c) illustrates the voicing and nonvoicing decision on the speech signal using the ZFF signal. Fig. 4.1(d) shows the filtered signal obtained after passing the voiced segments through a zero-frequency resonator with an adaptive window length. Fig. 4.1(e) and Fig. 4.1(f) show the contours of the strength of excitation and pitch period obtained for the segment of speech signal.

4.2 Analysis of laugh signals for synthesis

Analysis of laugh signals is done in terms of the excitation characteristics of the production mechanism to determine the features needed to synthesize laughter [58]. Features that are taken into consideration are: (a) rapid changes in \( F_0 \) within calls of a laughter bout, (b) strength of excitation at each epoch, (c) durations of different calls in a bout, and (d) breathiness/fricative segments in the laugh signal. Following are the main features that are modified to generate laugh signals.

4.2.1 Pitch period

Fundamental frequency of laughter is observed to be significantly higher than that for normal speech. As described earlier, during laughter production, there will be more airflow through the vocal tract (high sub-glottal pressure). This results in faster vibration of vocal folds, and hence reduction in the pitch period. Also, there is a raising pattern observed in the pitch period contour of a call. The general pattern that is observed in the pitch period contour within a call is that, it starts with some value, decreases slightly, and then increases nonlinearly to a high value, with vocal folds tending to reach the normal pitch frequency. This is because it is not normal for the vocal folds to maintain the initial high fundamental frequency (\( F_0 \)). It is also observed that a quadratic approximation seems to fit the pitch period contour well for majority of the laugh signals. The higher this slope, the more intense is the laughter. With the
progress of the calls, the slope of the pitch period contour also tends to fall. The rate with which it falls is assumed to be linear. Figure 4.2(b) shows the pattern of pitch period contour for a segment of laugh signal. We can observe from the figure that the pitch period values change nonlinearly. Figure 4.3 shows the actual pitch period contour and the modeled pitch period contour plotted together.

### 4.2.2 Strength of excitation

Similar to pitch period, the strength of excitation at epochs also changes rapidly. It increases nonlinearly and then decreases almost in a similar fashion. The slope of the strength of excitation contour typically falls with the progress of the calls. Figures 4.2(c) and 4.2(d) illustrate the general trend of the contours of the pitch period and strength of excitation for a segment of laugh signal. We can observe the pattern of the nonlinear increase and decrease in the strengths for the laugh signals. Note also the somewhat inverse relation in the variation of $T_0$ and SoE contours.
Figure 4.3 Illustration 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 shown using dashed line and modeled one is shown using dotted line (Color online).

4.2.3 Duration

The gap between two calls of a laughter is referred to as the intercall gap. The duration of the intercall gap is called intercall duration (ICD), and the duration of the call is called call duration (CD). Call durations are typically observed to be in the range of 0.08 to 0.20 seconds. For synthesis, any value in that range could be used for the call duration. Intercall durations are generally in the range of 0.5 to 1.5 times the call duration. The ratio of the duration of unvoiced to voiced segments in a laugh signal was reported to be greater than 1 Bickley and Hunnicut [51]. Intercall duration in a laughter bout was observed to increase with the progress of the calls. This was also confirmed by Kipper and Todt [57], where the duration of call was reported to decrease and the duration of the interval increases within a bout. In general no pattern was observed for call durations. The call durations vary depending on the speaker and the kind of laughter.

4.2.4 Frication

Because of high amount of airflow, there is turbulence generated within vocal folds, as a result of which glottal fricative /h/ (aspiration) is produced. It is predominantly observed in the intercall interval in most of the cases. The volume velocity of the air typically decreases from left to right within a call, as a result of which the amount of breathiness also falls within a call. The air flow during the open phase of the glottis is very high. This results in a strong turbulent noise source at the glottis [53]. As the calls progress, the amount of breathiness also decreases.
4.3 Synthesis of Laughter

In this work, laughter is synthesized by modifying the features mentioned in Section 4.2, for a vowel (preferably /a/) uttered by a speaker. The process involves modifying the characteristics of the source without changing the characteristics of the system. The following are the main stages involved in generating a laugh signal.

4.3.1 Incorporation of feature variations

4.3.1.1 Pitch period modification

Pitch period of the input vowel signal is modified using the method discussed in [67]. The input speech signal of a vowel is passed through the zero-frequency resonator for deriving the epoch locations as described in Section 4.1. The interval between the epoch locations gives the pitch period. A 10th order pitch synchronous linear prediction analysis is used to separate the source (LP residual) and system (LP coefficients) components. The LP residual and LP coefficients are associated with every epoch location. The desired pitch period contour for laughter is generated from the specification for the pitch period modification. The original pitch period contour of the vowel segment is modified, so that it follows a quadratic polynomial. New epoch locations are derived from the modified pitch period contour. The LP residual and LP coefficients are copied for each epoch in this new epoch sequence from the corresponding nearest epochs of the original signal. The residual at each epoch of the new epoch sequence is resampled by the pitch modification factor at that epoch. The new residual signal is used to excite the corresponding all-pole filter to obtain a signal with the desired prosody.

4.3.1.2 Modifying strength of excitation

Strength of excitation is an estimate of the strength of the impulse at the epoch. In order to find the relation between the strength of excitation and the amplitude of the peaks in the residual signal in each cycle, the following experiment was conducted [4]. A sequence of impulses with varying durations between consecutive impulses and with different amplitudes are generated. The sequence is passed through an all-pole filter with LP coefficients corresponding to different vowels. The output signals are passed through zero-frequency resonator, and the values of the strength of excitation are obtained. The resulting strength of excitation values are compared with the amplitudes of the impulses. There is an approximate linear relation observed between them. Hence the amplitudes of the samples in an epoch interval in the residual are modified by multiplying them with the scaling factor corresponding to the desired SoE contour. An inversion of quadratic approximation is assumed for the desired SoE contour, since the general trend of the SoE contour in each call duration is to first increase, and then decrease nonlinearly.
4.3.1.3 Incorporation of frication

Frication or breathiness is incorporated in the signal by further modifying the residual. To generate frication, white Gaussian noise equal to the length of the residual signal is generated. The noise samples are scaled to obtain energy equal to desired amount (i.e., 5% to 20% of the energy of the signal) of frication. The desired amount depends on the call number in the bout. The noise samples are passed through a resonator with a center frequency at 2500 Hz and a bandwidth of 500 Hz. The sequence is then multiplied with a weighing function \( w(n) = 1 - n/L \), where \( n \) is the sample number and \( L \) is the total number of samples in the signal, so as to obtain a linearly decreasing effect of frication. The resulting noise samples are then added to the residual samples to obtain the residual signal, which is then passed through the LP (all-pole) filter to synthesize laugh signal.

4.3.2 Steps in the synthesis of laughter

The block diagram of the synthesis system (Figure 4.4) shows the steps involved in the synthesis of laugh signal.

Figure 4.4 Block diagram of the laughter synthesis system (Color online).
1. The input signal (speech vowel /a/) is passed through a zero-frequency resonator for deriving the epoch locations. Pitch period is obtained by computing the interval between successive epoch locations.

2. A 10th order linear prediction analysis is also performed on the signal to derive the source (LP residual) and system (LP coefficients) components. The LP residual between epochs and the LPCs are associated for each epoch.

3. A segment of the signal corresponding to the length of a call duration is chosen.

4. For synthesizing the call in a bout, pitch period contour and strength of excitation contour are determined as described in Section 4.2, according to the desired prosody modification, described in Section 4.3.1.

5. The strength of excitation of the residual is modified as described in Section 4.3.1.2.

6. A new residual sequence is obtained after modifying the pitch period of the residual as explained in Section 4.3.1.1.

7. Frication is then incorporated in the resulting residual as explained in Section 4.3.1.3.

8. The residual signal is then used to excite the LP filter of the vowel to synthesize the call.

9. Random noise with very low amplitude (about 0.1% of the energy of the call) is generated and passed through a resonator with center frequency at 2500 Hz and a bandwidth of 500 Hz to synthesize the signal in the intercall duration.

10. The above steps are repeated for synthesizing different calls in the laughter, and to finally obtain a laughter bout.

11. Multiple bouts are synthesized, each with different number of calls and with different values for control parameters.

Figure 4.5 shows a synthesized laugh signal along with the desired SoE and $T_0$ contours that are used to generate it. Fig. 4.5(a) shows the desired SoE contour, which follows the inverse of a quadratic polynomial. The contour is generated by using the following equation:

$$y[n] = 1 - \frac{(n - \frac{4L}{T})^2}{(\frac{4L}{T})^2}$$  \hspace{1cm} (4.2)

where $n$ is the sample number and $L$ is the length of the signal in number of samples. The SoE contour value at each epoch location is used to multiply the LP residual signal in the following epoch interval. Fig. 4.5(b) shows the desired pitch period contour. The LP residual is modified to incorporate the desired pitch period ($T_0$) contour. The contour is generated by using the following equation:

$$y[n] = T_{min} + \frac{(n - \frac{L}{T})^2}{(\frac{2L}{T})^2} * (T_{max} - T_{min})$$  \hspace{1cm} (4.3)
Table 4.1 Parameters and preferred range of values for laughter synthesis.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Preferred range of values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of bouts</td>
<td>1-3</td>
</tr>
<tr>
<td>Number of calls in each bout</td>
<td>4-7 (depends on bout number)</td>
</tr>
<tr>
<td>Duration of each call</td>
<td>50-250 ms (depends on call number)</td>
</tr>
<tr>
<td>Duration of each intercall</td>
<td>50-250 ms (0.5 to 1.5 times of call duration)</td>
</tr>
<tr>
<td>Maximum (T_0) of each call</td>
<td>5-8 ms (for male)</td>
</tr>
<tr>
<td></td>
<td>4-6 ms (for female)</td>
</tr>
<tr>
<td>Minimum (T_0) of each call</td>
<td>3-4 ms (for male)</td>
</tr>
<tr>
<td></td>
<td>1-2 ms (for female)</td>
</tr>
<tr>
<td>Amount of frication in each call</td>
<td>5% to 20%</td>
</tr>
<tr>
<td>(in terms of signal energy)</td>
<td></td>
</tr>
<tr>
<td>Intensity ratio of first call to last call</td>
<td>1 to 100 (&lt; 1 ⇒ increasing intensity)</td>
</tr>
</tbody>
</table>

where \(n\) is the sample number and \(L\) is the length of the signal in number of samples. \(T_{min}\) and \(T_{max}\) are the minimum and maximum \(T_0\) values of the desired contour. The contour is normalized so that the maximum and minimum values of the contour correspond to desired maximum and minimum \(T_0\) values.

Fig. 4.5(c) shows the synthesized laugh signal, and Fig. 4.5(d) shows its spectrogram. The \(T_0\) of the first call ranges from 3.5 ms to 7.5 ms. The \(T_0\) of the last call ranges from 5.5 ms to 8 ms. The minimum \(T_0\) is increased as the calls progress. Also, the call duration decreases with calls. The call duration for the first call is chosen as 0.165 seconds. Durations of the remaining calls are decreased gradually. The first intercall duration (ICD) is chosen to be same as the duration of the first call, and the ICD is increased progressively. After the calls are generated, intensity of the calls is decreased as desired.

The laughter synthesis system is a flexible system, where the parameters to generate laughter can be controlled by the user. The parameters that can be manually set by the user along with their range are given in Table 4.1. Although any value chosen in the mentioned range will work, an improper combination of values could result in poor quality of the synthesized laughter. Following are a few examples of the many subtle and important interdependencies among the parameters that need to be taken into account to avoid generating poor quality laugh signal.

- Long bouts are associated with higher values of mean \(F_0\) for calls.
- Calls are longer in duration when they are less in number.
- Intercall duration depends on the call number.

34
There are several such interdependencies which need to be taken into account in order to produce natural sounding synthetic laughter.

![Figure 4.5 Illustration of synthesized laugh signal: (a) Desired strength of excitation (SoE) contour (b) Desired pitch period ($T_0$) contour (c) Synthesized laugh signal. (d) Spectrogram of the synthesized laugh signal.](image)

### 4.4 Experiments

#### 4.4.1 Perceptual significance of features

An experiment based on analysis-by-synthesis approach was conducted to determine the perceptual significance of the features described in Section 4.2. For this experiment, original laugh signals are taken, and the following features are modified: $T_0$ ($= 1/F_0$), SoE, amount of breathiness, and call and intercall durations. For each original sample, modifications are made for different combinations
Table 4.2 Perceptual evaluation scores obtained for the modified versions of an original laugh signal.

<table>
<thead>
<tr>
<th>Sample</th>
<th>$T_0$</th>
<th>SoE</th>
<th>Breathiness</th>
<th>Call, Intercall durations</th>
<th>Naturalness mean</th>
<th>Naturalness variance</th>
<th>Acceptability mean</th>
<th>Acceptability variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.2</td>
<td>0.36</td>
<td>4.35</td>
<td>0.22</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3.9</td>
<td>0.22</td>
<td>4.05</td>
<td>0.21</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3.6</td>
<td>0.21</td>
<td>3.7</td>
<td>0.22</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>3.4</td>
<td>0.31</td>
<td>3.5</td>
<td>0.29</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>4.0</td>
<td>0.29</td>
<td>4.1</td>
<td>0.32</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>3.65</td>
<td>0.09</td>
<td>3.8</td>
<td>0.22</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>3.25</td>
<td>0.16</td>
<td>3.45</td>
<td>0.23</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3.0</td>
<td>0.26</td>
<td>3.2</td>
<td>0.29</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.9</td>
<td>0.24</td>
<td>3.0</td>
<td>0.23</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2.7</td>
<td>0.19</td>
<td>2.95</td>
<td>0.21</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2.4</td>
<td>0.21</td>
<td>2.45</td>
<td>0.09</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2.05</td>
<td>0.21</td>
<td>2.2</td>
<td>0.11</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2.35</td>
<td>0.14</td>
<td>2.45</td>
<td>0.19</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>2.1</td>
<td>0.11</td>
<td>2.2</td>
<td>0.36</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1.7</td>
<td>0.36</td>
<td>1.8</td>
<td>0.24</td>
</tr>
<tr>
<td>16</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.25</td>
<td>0.26</td>
<td>1.7</td>
<td>0.14</td>
</tr>
</tbody>
</table>

of features, thus generating 15 ($= 2^4 - 1$) different samples. These samples were played randomly to 20 subjects who were asked to score the samples for naturalness and acceptability according to their preference on a scale of 1 to 5: 1-Very Poor, 2-Poor, 3-Average, 4-Good, 5-Excellent. Acceptability is a measure of how close the sample is sounding to a laughter, and naturalness is a measure of how natural it is.

The proposed modification of the original laugh signal is performed as follows: The laugh signal is segmented using the voiced/nonvoiced decision to extract the calls and intercalls. For every laugh syllable (call plus intercall), the following changes are made:

1. $T_0$: Rising pitch period contour is replaced with a flat contour.
2. SoE: Strength of excitation is reduced.
3. Breathiness: Breathiness is reduced by decreasing the relative amplitudes of the samples which are more than 1 ms away from the epoch within each pitch period and in the intercall intervals.
4. Duration: Call and intercall durations are changed randomly using different proportions of the voiced and unvoiced regions.

Table 4.2 gives the results of the experiment. The table gives mean opinion scores (MOS) for naturalness and acceptability of some laughter samples for modifications of different features (including the original sample). In Table 4.2, a ‘0’ in the 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 the 1st sample,
Table 4.3 Results of the experiment on perceptual significance of different features.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Naturalness (mean)</th>
<th>Acceptability (mean)</th>
<th>Naturalness (mean)</th>
<th>Acceptability (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T&lt;sub&gt;0&lt;/sub&gt;</td>
<td>2.18</td>
<td>2.34</td>
<td>3.63</td>
<td>3.77</td>
</tr>
<tr>
<td>SoE</td>
<td>2.66</td>
<td>2.84</td>
<td>3.14</td>
<td>3.28</td>
</tr>
<tr>
<td>Breathiness</td>
<td>2.58</td>
<td>2.75</td>
<td>3.23</td>
<td>3.36</td>
</tr>
<tr>
<td>CD &amp; ICD</td>
<td>2.76</td>
<td>2.95</td>
<td>3.05</td>
<td>3.16</td>
</tr>
</tbody>
</table>

which is the original version (all 0s), has high naturalness and acceptability scores as expected, and 16<sup>th</sup> sample, in which all the features are modified (all 1s) has very low score (1.26) in the table, which shows the significance of these four features in characterizing laughter.

For finding the relative significance of individual features, the mean of scores when a particular feature is modified (1’s) is compared with the mean of scores when it is not modified (0’s). Table 4.3 shows these comparisons for the features, and Table 4.4 shows the difference between the values in modified and unmodified cases for the features. The difference in the values for a feature indicates the significance of that feature in the laughter. The drop in the score of perceptual evaluation indicates the significance of the missing feature. The more the difference between the two values, the more significant the feature is. We can see from Table 4.4 that the difference is maximum in case of T<sub>0</sub>, and minimum in case of call and intercall durations.

Table 4.4 Results showing the difference in perceptual significance experiment scores.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Difference in Naturalness (mean)</th>
<th>Difference in Acceptability (mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T&lt;sub&gt;0&lt;/sub&gt;</td>
<td>1.45</td>
<td>1.43</td>
</tr>
<tr>
<td>SoE</td>
<td>0.48</td>
<td>0.44</td>
</tr>
<tr>
<td>Breathiness</td>
<td>0.65</td>
<td>0.61</td>
</tr>
<tr>
<td>CD &amp; ICD</td>
<td>0.29</td>
<td>0.21</td>
</tr>
</tbody>
</table>

4.4.2 Speaker identification from laughter

An experiment was conducted to explore the possibility of identifying a person from laughter. As part of the experiment, subjective tests were conducted. Subjects were presented with 10 isolated laughter samples comprising of both synthesized and natural laughter of speakers (people who laughed) whom the subjects personally have known and heard, and natural laughter samples of a few celebrities (famous people known to all the subjects). Identity of the speakers was hidden from the subjects, and they were asked to identify the speakers from the laughter samples. Speech produced by the same speaker was also presented to the subjects. Subjective evaluation tests were performed with 7 subjects. None of the subjects were able to identify any of the speakers from their isolated laughter samples, though they were
able to identify all the speakers from the respective speech samples. This implies that laughter does not
carry significant speaker-specific information.

The results are not really surprising. We can identify a person from his/her laughter, only if the
laughter contains any special/unique attention-gathering characteristics, or if we have had enough expo-
sure to the person’s laughter in isolation. It is even more difficult to identify a person from synthesized
laughter, probably because a person’s natural way/style of laughter might not match with the prosodic
and intonation features of the synthesized laughter.

### 4.4.3 Understanding the significance of source and system parameters

An experiment was conducted to understand the significance of source and system parameters on the
synthesis of laughter. As part of the experiment, signals of vowel sounds are taken from two different
speakers as input to the synthesis system. Source parameters (residual) of input speech signal of the first
speaker are extracted and modified accordingly as described in Section 4.1 and Section 4.3.1. System
parameters (LP coefficients) are extracted from the vowel signal of the second speaker. The modified
residual of the first speaker is passed through the LP filter whose system coefficients correspond to sys-
tem parameters of the second speaker to synthesize laughter. The same process is repeated, when source
parameters of the second speaker and system parameters of the first speaker are used for synthesis. The
experiment was conducted taking 5 different pairs of speakers as input.

The synthesized samples are good enough to be acceptable as laughter. The synthesized samples
from both the methods sound different. It is observed in some cases that the synthesized signal is more
similar to the laughter of the speaker whose source parameters are used. Though there is no way to
quantify this observation, the differences in the voice qualities of the speakers could be observed with
confidence.

### 4.5 Results of synthesis

Subjective evaluation tests were performed to evaluate the synthesized laughter using 8 subjects.
The subjects were presented with 9 laughter-only samples, which were synthesized offline using the
proposed method. All the samples were randomly played, and were not grouped in any particular order.
The tests were performed in a quiet office environment. Each subject had to listen and score each sample
for naturalness and acceptability according to their preference on a scale of 1 to 5. The samples were
synthesized by varying the input parameters like the ratio of the first call to last call intensity, ratio
of intercall duration (ICD) to call duration (CD) of first call, presence of breathiness and number of
calls. The mean opinion scores (MOS) for a few selected values of the parameters are tabulated in Table
4.5. The selected values of the parameters are only for illustration, and hence they are not exhaustive.
Presence of breathiness is indicated by ‘1’ in the table.
Table 4.5 Performance of laughter synthesis system in terms of MOS.

<table>
<thead>
<tr>
<th>Sample</th>
<th>First to last call intensity ratio</th>
<th>ICD to CD ratio</th>
<th>Breathiness</th>
<th>No. of calls</th>
<th>Gender</th>
<th>Naturalness (mean score)</th>
<th>Acceptability (mean score)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>0.5</td>
<td>0</td>
<td>7</td>
<td>M</td>
<td>2.88</td>
<td>2.75</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>M</td>
<td>2.13</td>
<td>2.25</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>M</td>
<td>2.5</td>
<td>2.63</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>0.5</td>
<td>1</td>
<td>5</td>
<td>M</td>
<td>3</td>
<td>2.88</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>F</td>
<td>3</td>
<td>2.63</td>
</tr>
<tr>
<td>6</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>M</td>
<td>3.25</td>
<td>3.13</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>F</td>
<td>3.13</td>
<td>2.5</td>
</tr>
<tr>
<td>8</td>
<td>10</td>
<td>1.5</td>
<td>1</td>
<td>5</td>
<td>M</td>
<td>3.38</td>
<td>3.38</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>1.5</td>
<td>1</td>
<td>8</td>
<td>F</td>
<td>3.25</td>
<td>3.13</td>
</tr>
</tbody>
</table>

Scores in the table indicate that both acceptability as laughter and quality of synthesis of most of the synthesized samples lie in the range of Average to Good in the continuum of evaluation scale. It is also observed that maintaining a ratio of 10 for intensity for the first to last calls produced better results. It produced sufficient effect of dampening of the intensity levels. Ratio of ICD to CD did not have a major impact on the perception of laughter. It is observed from the table that better scores are obtained for samples which included breathiness. Synthesized male laughter samples were accepted with better score as laughter than for the female ones. Though there is no specific range for the number of calls present in a laughter, synthesized laughter samples were more acceptable when the number of calls in a single bout ranged from 4 to 7. Laughter is modeled such that $F_0$ values do not exceed 300 Hz in the case of synthesized male laughter, and 400 Hz in the case of a synthesized female laughter. The maximum and minimum values of $F_0$ decrease as the calls progress. The breathiness/frication that is incorporated should lie within a certain limit/threshold, as mentioned in Table 4.1, beyond which it leads to an output that sounds noisy.

4.6 Summary

In this chapter, analysis of laugh signals for synthesis was performed. Source features derived from a vowel segment (/a/) were modified suitably to synthesize a laugh signal. Experiments have been conducted to find the perceptual significance of features and the role of source and system characteristics in the perception of laughter. Finally, results of the laughter synthesis system have been presented.
Chapter 5

Summary and conclusions

5.1 Summary of the work

In this dissertation, analysis and synthesis of two nonverbal means of communications (one voice quality and one nonverbal vocalization) have been addressed. Major focus has been on deriving excitation source characteristics and to be able to modify them to produce a sound with desired effects. Analysis has been performed to extract excitation source information of the voice quality, and certain source features have been explored to characterize breathy voice. Features such as instantaneous fundamental frequency ($F_0$), strength of excitation ($SoE$), periodic to aperiodic energy ratio ($PAP$) and perceived loudness measure ($\eta$) have been derived from the speech signal using robust signal processing methods and analyzed. Comparisons between values obtained for various features is made for breathy and modal voiced syllables. Classification experiments have been conducted to discriminate breathy vowels from modal ones using PAP and $\eta$. Both of them give good discrimination, and the ‘$\eta$’ in particular is highly successful in classifying breathy voice. Attempts to synthesize (characterize) breathy voice have then been made by modifying the source characteristics of a modal voiced speech.

Laughter signals have been analyzed to study the pattern and structure of the signal and the source features extracted from it. The contours and ranges of $F_0$ and contours of SoE have been studied and modeled for synthesis. A vowel (/a/ or /i/) segment is taken and its excitation features (pitch contour) and excitation source signal (LP residual) are derived. Suitable modifications in source characteristics are made on the LP residual by modifying the pitch contour, duration, SoE, and by incorporating frication-like noise. The modified residual signal is used to excite the LP filter to synthesize laughter. Experiments have been performed to assess the significance of various features, role of source and system characteristics, and to identify speaker from laughter. Subjective evaluation has been performed on the results of the synthesis system. Results of the laughter synthesis system have been discussed. Synthesized laugh samples have been considered as acceptable. Suitable (preferable) range of values of parameters are given for synthesizing an acceptable laughter.
5.2 Major contributions of the thesis

The major contributions of the thesis are:

- Analysis of breathy voice is done based on excitation characteristics of production mechanism.
- Measure of significance of perceived loudness measure ($\eta$) and periodic-to-aperiodic energy ratio ($PAP$) has been done by using them for classification.
- Analysis of excitation source information and patterns followed by the source parameters in laughter have been studied.
- A method is proposed to synthesize laughter from speech vowel by modifying its source characteristics.
- Perceptual significance of various features in laughter have been studied by conducting experiments.

5.3 Scope for future work

- In this work, only breathy voice has been chosen for analysis and synthesis. Similarly, other voice qualities such as creaky voice, whispery voice, slack voice, etc. could be chosen to be explored.
- In this work, only laughter has been analyzed and synthesized. Similarly, other paralinguistic entities (nonverbal events) could be analyzed and synthesized.
- System characteristics have not been modified in the process of synthesizing laughter. This can be done for further improvement in laughter synthesis.
Publications

The work done during my Masters has been disseminated to the following journals and conferences.

5.4 Journals


5.5 Conferences


Bibliography


