2009-12-21

Emotion Recognition Using Glottal and Prosodic Features

Alexander Iliev Iliev
University of Miami, ailiev@miami.edu

UNIVERSITY OF MIAMI

EMOTION RECOGNITION USING GLOTTAL AND PROSODIC FEATURES

By

Alexander Iliev Iliev

A DISSERTATION

Submitted to the Faculty
of the University of Miami
in partial fulfillment of the requirements for
the degree of Doctor of Philosophy

Coral Gables, Florida

December 2009

UNIVERSITY OF MIAMI

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

EMOTION RECOGNITION USING GLOTTAL AND PROSODIC FEATURES

Alexander Iliev Iliev

Approved:

Michael Scordilis, Ph.D. Research Associate Professor of Electrical and Computer Engineering	Terri A. Scandura, Ph.D. Dean of the Graduate School
Miroslav Kubat, Ph.D. Associate Professor of Electrical and Computer Engineering	Mohamed Abdel-Mottaleb, Ph.D. Professor of Electrical and Computer Engineering
Subramanian Ramakrishnan, Ph.D. Associate Professor of Mathematics	Mei-Ling Shyu, Ph.D. Associate Professor of Electrical and Computer Engineering

ILIEV, ALEXANDER ILIEV (Ph.D., Electrical and Computer Engineering)
Emotion Recognition Using (December 2009)
Glottal and Prosodic Features.

Abstract of a dissertation at the University of Miami.

Dissertation supervised by Professor Michael Scordilis.
No. of pages in text. (148)

Emotion conveys the psychological state of a person. It is expressed by a variety of physiological changes, such as changes in blood pressure, heart beat rate, degree of sweating, and can be manifested in shaking, changes in skin coloration, facial expression, and the acoustics of speech. This research focuses on the recognition of emotion conveyed in speech. There were three main objectives of this study. One was to examine the role played by the glottal source signal in the expression of emotional speech. The second was to investigate whether it can provide improved robustness in real-world situations and in noisy environments. This was achieved through testing in clear and various noisy conditions. Finally, the performance of glottal features was compared to diverse existing and newly introduced emotional feature domains. A novel glottal symmetry feature is proposed and automatically extracted from speech. The effectiveness of several inverse filtering methods in extracting the glottal signal from speech has been examined. Other than the glottal symmetry, two additional feature classes were tested for emotion recognition domains. They are the: Tonal and Break Indices (ToBI) of American English intonation, and Mel Frequency Cepstral Coefficients (MFCC) of the glottal signal. Three corpora were specifically designed for the task. The first two investigated the four emotions: *Happy*, *Angry*, *Sad*, and *Neutral*, and the third added *Fear* and *Surprise* in a six emotions recognition task.

This work shows that the glottal signal carries valuable emotional information and using it for emotion recognition has many advantages over other conventional methods. For clean speech, in a four emotion recognition task using classical prosodic features achieved 89.67% recognition, ToBI combined with classical features, reached 84.75% recognition, while using glottal symmetry alone achieved 98.74%. For a six emotions task these three methods achieved 79.62%, 90.39% and 85.37% recognition rates, respectively. Using the glottal signal also provided greater classifier robustness under noisy conditions and distortion caused by lowpass filtering. Specifically, for additive white Gaussian noise at SNR = 10 dB in the six emotion task the classical features and the classical with ToBI both failed to provide successful results; speech MFCC's achieved a recognition rate of 41.43% and glottal symmetry reached 59.29%. This work has shown that the glottal signal, and the glottal symmetry in particular, provides high class separation for both the four and six emotion cases. It is confidently surpassing the performance of all other features included in this investigation in noisy speech conditions and in most clean signal conditions.

Dedication

To my Parents

In recognition of your diligent efforts in guidance and support through all the years of my studies, for understanding the significance of following my dreams, for showing me the right way in times when no path seemed possible, for standing next to me in vital moments of my life that shaped me into the man I am today, I will be forever grateful.

With lots of love, I sincerely thank you!

"Try not to become a man of success, but rather try to become a man of value."

Albert Einstein

Acknowledgments

First and foremost I would like to thank my adviser Dr. Michael Scordilis for his unremitting support, for his encouragement and confidence in my work, and for his persistent and relentless guidance in my professional advancement.

To all of my committee members: Dr. Abdel-Mottaleb, Dr. Mei-Ling Shyu, Dr. Miroslav Kubat and Dr. Subramanian Ramakrishnan, my gratitude for your astute advice and vision that led to considerable enhancement of my dissertation. Thank you for closely following my progress throughout the years at the University of Miami, and for being part of my most important professional effort by far.

I would also like to thank my friend Mark Mandica for the detailed illustration of the larynx and to express my gratitude to my sister, extended family, and friends for giving me support all these years.

Alexander I. Iliev

University of Miami

December 2009

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Glottal Symmetry on lowpass filtered speech:

Glottal Symmetry with SNR of 30dB:

Glottal Symmetry with SNR of 10dB:

MFCC on clean speech:

MFCC with SNR of10dB:

Classical prosodic features on clean speech:

Combined features (Classical with ToBI) on clean speech:

CHAPTER 1

Introduction

Interpersonal communication is greatly facilitated by the detection of emotion through visual and auditory clues, which are used to deduce the motive, intent and general psychological state of a person. Speech, due to the multilayered processes involved in its production is a main vehicle for emotional expression. In speech communication, emotion enhances the information contained in the intended spoken message. In intelligent computing, automated recognition of emotion in speech is a growing area of interest with applications that span from speech synthesis and security to the services industry, psychology and medicine.

Phonetic, prosodic, and linguistic features undergo transformations associated with emotional expression. In this context, acoustical analysis of speech aims at the robust extraction of relevant signal features which best describe the changes associated with a particular emotion. Speech analysis provides considerable advantages over other techniques because it is non intrusive and it can be acquired simply with a microphone even over the telephone, and it has therefore recently received significant attention.

Emotion recognition is an important field leading to the innovation and implementation of devices able to improve the quality of life in many key areas. Possible fields that may directly benefit from it could be: speech synthesis, surveillance, customer service, quality control, entertainment, education, forensics, and aiding people with disabilities.

1.1 Emotional States

Different emotional states greatly affect the natural speech production of the speaker. Key parameters such as fundamental pitch, energy, and speech rate hold valuable information to determine different emotions, thus specific speaker-dependant shaping of their patterns are known to exist. That is why, to determine specific emotion, it is important to target them individually in any state of given speech production.

There is no clear consensus on the number and type of emotions that can be expressed in speech. In the review by Ververidis and Kotropoulos (2006), a total of 29 different emotions have been represented in 64 emotional speech data collections. The number of emotions used in most of those cases varied between four and six, which form a basic or "prime" set, while other higher level emotions are less widely used with less agreement on their identity. In Eckman's work (1992) was proposed that fifteen basic emotions can be detected, with *neutral*, *happiness*/*joy* and *anger* being the most predominant, while there is no absolute agreement in regards to the number of basic emotions. He also introduced the notion of "emotion families", which specifies that each basic emotion in not a single affective state but rather a *family* of related states. This implies that emotions in a particular family share common characteristics, such as a

common way of expression and physiological activity. It is pointed out that in *anger* for instance there is not only one expression but more than 60 different *anger* expressions. In the work done by Cowie and Cornelius (2003) and by Cornelius (1996), the number of emotions also varies, but the best known key emotions are the so-called "Big-six": *anger*, *happiness*, *sadness*, *fear*, *surprise*, and *disgust*. The same set of emotions has also been used in Bhatti et al. (2004), where it was concluded that prosody is a major emotional agent. Other authors experimented with four emotional types namely: *happy*, *angry*, *sad*, and *neutral* (Noda et al., 2006).

Evidently there are a number of emotions that can be considered for the emotion recognition task, but how many and which ones to use clearly relates to the specifics of the problem at hand. The majority of studies contain from three (Noda et al., 2006) to six emotions (Cowie and Cornelius, 2003) as the basic set. In this work, *Happy*, *Angry*, *Sad*, *Neutral*, *Fear*, and *Surprise* were selected for analysis in a variety of scenarios.

As in other classification problems, selecting the most effective feature set in an emotion recognition task is a key factor for success in classification. While some researchers focus strictly on speech signal processing (Kwon et al., 2003), others employed facial recognition techniques along with speech analysis (Go et al., 2003).

In addition to the standard pitch information Kwon et al. used the log energy, formant, mel-band energy, and mel-frequency cepstral coefficients (MFCC), while including first and second derivatives (velocity and acceleration) of pitch and MFCCs. Once again it was experimentally established that pitch and energy were the most important parameters in emotion recognition. Using the SUSAS database of emotional and stressed speech they reported 96.3% success in bi-polar emotion recognition

(stress/neutral) and 70.1% in four-class speaking style classification using Gaussian Support Vector Machines (SVM). That study was extended using the speaker independent AIBO database of emotional children's speech reporting 42.3% success rate for a five-class emotion recognition task.

In another extensive emotion recognition investigation using only speech information Wang et al. (2004) extracted 55 potential features (25 prosodic, 24 MFCC and 6 formant frequencies). Because of similarities, not all features contributed to the system accuracy and in fact some had a diminishing effect. Furthermore, such large number of features added substantially to the computational complexity of the system. This in turn led to feature vector pruning using the SPSS software and the Mahalanobis distance as criterion function. Their findings clearly show the need for improvement, by finding a better feature domain while further pruning its members.

1.2 Signal Features for Emotion Recognition

There may be significant overlap between emotions in the signal feature space depending on the selected feature domain and the number and type of targeted emotions. Bosch (2003) has concluded that there is no clear distinction between members of three 'emotional groups': *neutral-sadness*, *anger-fear*, and *happiness-surprise* making their separation difficult. In that work, as well as in Iliev et al. (2007), intonation, expressed as temporal pitch variations, is recognized as the most effective means for utterance-based emotion decoding. In the latter, the intonation information was described using the Tonal and Break Indices or ToBI tone tier features (Beckman and Elam, 1997). The second most prominent prosodic feature is energy. Extensive statistical analysis on large amounts

of data in Iida et al. (2003), established once more the strong link between pitch and duration in the detection of emotion. In the work done by Gobl and Chasaide (2003), the correlation between voice overtones in various types of speech and pitch dynamics was combined with vocal tract features and used for emotion detection.

While most emotion recognition approaches focus entirely on the speech signal, facial recognition techniques have also been employed together with speech features. In a study by Go et al. (2003), facial expression recognition is performed via multi-resolution analysis based on discrete wavelets using feature vectors derived through linear discriminant analysis (LDA). Speech features were extracted from each wavelet analysis sub band. The recognition system used a multi-decision making scheme by merging the facial and spoken emotion recognition results. An improvement of the overall system was reported when using both modalities.

Finally, there are studies that also include textual content in conjunction to the speech signal (Chuang and Wu, 2004), where "emotion modification words" were manually defined. Textual content has also been included in addition to the acoustical signal in Chuang and Wu (2004). Results showed that while recognition rates of textual content-based systems cannot reach the success rates of acoustics-only systems, by integrating both types of information emotion recognition performance may be improved.

1.3 Phonetic Information

A number of diverse phonetic description techniques have been proposed aiding the recognition or synthesis of emotion (Beckman and Elam, 1997; Klasmeyer and Sendlneier, 1995; Iida et al., 1998). In Roach's work (2000), the connection between

phonetic labels and physically measurable parameters of emotional speech is exploited and used to aid automatic detection. Scherer (2003) points to the fact that speech parameters need to be carefully chosen and the phonetic information plays an important role. In particular, he suggests a categorization in the type of affective states, which are: emotion, mood, interpersonal stances, attitudes, and personality trails. All of these can be described by: intensity, duration, synchronization, event focus, appraisal elicitation, rapidity of change, and behavioral impact.

1.4 Prosodic Information

Short-time supra-segmental features and their statistics provide important emotion-related information (Ververidis and Kotropoulos 2006). Among those, pitch contours and their dependencies play a key role. The actual number of prosodic parameters used when quantifying emotion varies greatly. Some of the most common include pitch frequency F_0, energy, formant locations and other temporal and spectral features. Several statistical parameters have also been proposed, such as mean, range, variability, formant F_1 bandwidth, formant F_1 mean, formant F_2 mean, formant precision, formant frequency range, speech rate, transition time, and spectral noise. In addition, a system that standardizes pitch variations and their dependencies had been used for emotion recognition by Roach (2000) and Stibbard (2000).

In the Tonal and Break Indices (Beckman and Elam, 1997) structure, prosody is transcribed into four main tiers (tone, break, orthographic, and miscellaneous) by means of labeling, which hold important information on the phrase accent, rise or fall of the pitch, the type of boundaries in an utterance, the number and length of pauses, non-

speech event and more. Rules for automatic generation of ToBI based prosody markers are given by Jilka et al. (1999).

1.5 Linguistic Information

Emotional expression is speaker and language dependent. Although it is primarily conveyed by prosody and to a lesser extent by semantics, analysis of the broad linguistic content of an utterance can aid the identification of approval, attention, impatience or frustration in the speaker's message (Bosch 2003). Furthermore, emotional keywords or phrases can be effective vehicles in expressing and classifying emotion in speech.

Pre-selected "emotional keywords", each measured with a corresponding "emotion intensity values" can augment acoustic features. This way, the spoken emotion is represented by a combination of lexical and acoustic streams. Waveform analysis with and without the use of linguistic information was introduced in Chuang et al. (2004) and Kwon et al. (2003), respectively. In the former, acoustical features related to intonation, timbre, tempo, and rhythm are extracted. The combination of acoustical, lexical, and syntactical information had also been explored by Ananthakrishnan and Narayanan (2005) in an effort to enhance the prediction of an emotional state by including non-acoustic events. For that purpose a small vocabulary of tags was used. Adequately detecting the syllable accent and the utterance boundary type played a key role in the proposed method. Using that knowledge a classifier is built by assigning prosodic events to syllables from an unlabeled test data set which is described by acoustic speech parameters.

Schuller et al. (2004) combined acoustic and linguistic information for the task of detecting among seven different emotions in spontaneous speech. Belief networks were utilized to spot emotionally-significant words and combine them into phrases. Although acoustic information alone resulted in a 25.8% error rate and linguistic information alone resulted in 40.4% error rate, when the two types of information were combined in a multilayer perceptron classifier the error rate dropped to just 8%.

In another study, Lee and Narayanan (2005) investigated the usefulness of language and discourse information in improving the discrimination between negative and non-negative emotions in spoken dialogs from a call center application. For this purpose, they introduced the information-theoretic notion of "emotional salience" with which they automatically calculated how much information a word provides about a given emotional category. When acoustic and linguistic information was combined emotion classification improved by 40.7% for male and 36.4% for female speakers over the implementation which used acoustic information alone.

1.6 The Role Of The Glottal Signal

The goal of this work is to study the possible contribution the glottal waveform plays in expressing different emotional states and whether glottal-based are effective in spoken emotion recognition. In general, the speaking style of a person can be revealed by visually examining the laryngograph signal, which can usually be obtained by an electroglottograph. There are several studies supporting this case for speech under stress (see Laukkanen et al., 1996 and references therein). Variations of the glottis have been studied in emotion-related disorders, such as clinical depression in (Moore et al., 2003)

where the spectral tilt and the bias of the glottal frequency response were key features used to derive inter- and intra- sentence statistics. The glottal spectral slope has also been used in Zhou et al. (1999). In that study features derived from the nonlinear Teager energy operator (TEO) were used for stress classification. As suggested, TEO-based features better represent the nonlinear airflow structure of speech production under various stress conditions. In a related study (Zhou et al., 2001), the same authors proposed new features including TEO-decomposed FM variation (TEO-FM-Var), normalized TEO autocorrelation envelope area (TEO-Auto-Env) and critical band-based TEO autocorrelation envelope area (TEO-CB-Auto-Env). Another noteworthy approach using glottal information was undertaken by (Ling et al., 2005). There, voice is considered to be the output of an Liljencrants-Fant (LF) source model (Fant, 1986) whose glottal formant parameters and spectral tilt can be measured. In that case, the glottal frequency characteristics are claimed to be more likely to be preserved in the spectrum of the speech, as opposed to obtaining the glottal waveform via inverse filtering. The normalized amplitude quotient of glottal flow, proposed by Matti and Paavo (2004), reveals significant differences between emotions expressed in continuous speech.

The role of glottal waveform control in expressing emotional speech has also received attention in speech synthesis and voice transformation. Cabral and Oliveira (2006) examined the relationship between emotions and glottal parameters and they proposed a system which simulates emotions in neutral speech by changing glottal source parameters and prosody. The role of glottal amplitude quotient in conveying paralinguistic information in speech was also investigated by Mokhtari and Campbell

(2003) in the context of automatically annotating large speech databases for use in concatenative speech synthesis.

In this work, the effectiveness of using the glottal signal in classifying emotion in speech was investigated. Glottal features include the glottal symmetry, defined as the ratio of closing to opening phase durations, and MFCCs of various orders. The effectiveness of using features based on the speech output was contrasted with that of glottal features alone as well as combinations of both glottal and speech features. Six different classifiers were used in this study including: the Bayesian classifier (BC), k-nearest neighbor (k-NN), Gaussian mixture model (GMM), C4.5 classifier, support vector machine (SVM), and the new optimum path forest search (OPF).

1.7 Tonal And Break Indices System – ToBI

In recent years ToBI became a standard prosodic transcription of intonation patterns, using American English. As discussed in (Beckman, 1997), it was created by speech scientist from many diverse areas such as linguistics, psychology, and electrical engineering, seeking a transcription system with common prosodic elements. So far, ToBI has mostly been used for speech synthesis applications, where the prosody of speech production remained in focus. The ToBI standard consists of several different sub domains that describe different information measures at different points in time. Those domains are referred to as tiers and the main four tiers are: *tone, break, orthographic, and miscellaneous*. Other tiers could be developed based on specific applications (Beckman, 1997). The first two tiers (*tone* and *break*) represent the core prosodic analysis, and so they are based on two acoustical parameters: pitch and energy. The *orthographic tier* just

like the *miscellaneous* is not part of the prosodic analysis, but it provides extra information for describing the production using common English orthography. Both of them are important, but the labels they provide are more useful for generating the F_0 contour (Black and Hunt, 1996), which is more applicable to synthesis rather than analysis. There were 16 ToBI features used in total, as described in Section 3.4. Methods of extraction are detailed in Section 4.3.

The main goal of this work was to determine if the glottal signal carries detectable emotional content as observed through the glottal symmetry. Another goal was also to investigate and establish more robust emotion recognition system than the ones currently available. For that reason, new features were added in addition to the eleven chosen in Wang (2004). This extended the pitch feature component by providing more affluent pitch information based on prosodic transcriptions using ToBI. In accordance with Wang (2004), using classical approach, in a given phrase the 11 features used are: *mean*, *median*, *standard deviation (STD)*, and *maximum of pitch* (1-4); *rising-falling of pitch ratio* (5); *maximum of falling range* (6); *mean*, *STD*, and *maximum of energy* (7-9); *average pause length* (10); and *speaking rate* (11). Full description of the extraction and application of all classical approach parameters is explained in details in Sections 3.3 and 4.2.

1.8 Mel Frequency Cepstral Coefficients – MFCC

The mel scale is widely used in music and speech signal processing. It has been introduced by *Stevens*, *Volkman* and *Newman* in 1937. The idea is directly associated with the nonlinear perception of sound in humans. It is a representation of equally spaced

pitches according to human perception. The higher the pitch the wider the frequency limen at which listeners can perceive pitch change. In short the mel-frequency scale is approximately linearly spaced below 1 kHz and logarithmically distributed for higher frequencies. To obtain the mel frequency cepstral coefficients (MFCCs) we convert back to time domain. Since the mel coefficients are real numbers we use the Discrete Cosine Transform (DCT) for the task. The reason why mel coefficients are so important is because they represent the local frequency changes in the signal, thus providing better understanding of the subtle changes occurring in the frequency domain. That is the reason why the mel cepstrum is also known as 'spectrum of the spectrum'.

1.9 Classification Overview

Identifying the feature extraction process is only one aspect of the classification system; the other is the type of classifier used. Emotion recognition has been implemented on a variety of classifiers including Maximum Likelihood Classifier (MLC), Neural Network (NN), k-nearest neighbor (k-NN), Fisher's Linear Discriminant Analysis (FLDA), and Gaussian Mixture Model (GMM). Some studies employ hidden Markov model (HMM) alone (Schuller et al., 2003), some use them in combination with support vector machine (SVM) (Lin and Wei, 2005). Other use Weighted Bayesian Classifier and Multi Layer Perception (MLP) (Jiang and Cai, 2004). Present are also classifiers using Neural Networks (NN) (Nicholson et al., 2000), (Razak et al., 2005), and also applying K-nearest neighbors (Petrushin, 2000) to the problem. In Kang et al. (2000), HMM, NN and MLB (Maximum-Likelihood Bayes) have been compared in an emotion classification task. In Wang (2004), the authors used 720 utterances for a six

emotion classification task (happiness, sadness, anger, fear, surprise, and disgust) with highest success rates reported at 67.22%, when using the FLDA classifier with 21 features, stepwise selected out of 55 initial features. Using analogical feature selection, k-NN showed 63.33% success. The remaining four classifiers performed better when 11 out of 25 features were selected. Success rates were 57.78% for NN, 55% for MLC, 53.33% for GMM(2) and 52.78% for GMM(3), where GMM(2) and GMM(3) denote the number of GM components used.

GMM is one of the most prominent statistical methods used for clustering and density estimation and as a result it has been involved in several emotion recognition tasks. In Schuller et al. (2003), a single state HMM's (GMM) was used to model each one of six emotions (one GMM per emotion). It was reported that there is no gain observed after up to four Gaussian components were used in each GMM, and the classification is based on Maximum Likelihood (ML). In Jiang (2004) used GMM as one of source likelihood scores which are combined together in the second phase to make a classification decision. In another investigation of human emotion recognition, Wang (2004) compared GMM, ML, NN and Fisher Linear Discrimination Analysis (FLDA). GMM has also been studied as a comparable method to other classifiers in Wang (2004) and Hung et al. (2004). In Chuang and Wu (2005) the GMM was trained for each emotion state based on a feature vector derived from acoustics of speech as well as textual content information. The acoustical information contained not only features like intonation, timbre, acoustics, tempo, and rhythm, but also special intonations, such as crying, trembling, and unvoiced speech. The combination of MFCC and pitch features was used as the feature vector to train the GMMs in the work of the emotion detection by

Neiberg et al. (2006). Also GMMs were built to identify emotions from the speech signals based on spectral features and prosodic features (Luengo et al., 2005).

For the emotion recognition task at hand, the k-NN classifiers have also been previously applied (Witten and Frank, 2005). Among other classification methods, the C4.5 decision tree is of particular interest. This algorithm is using the difference in information entropy or information gain. As described in Witten (2005), the default confidence value used for training was set to 25%. Every attribute of the dataset passed to the C4.5 classifier is used to make a decision and therefore prune the tree, thus partitioning the dataset. The normalized information gain after the dataset was divided is examined and the split with the highest information gain is retained. The process continues until termination criteria are met. For the distance measure the Euclidian or Manhattan distances are typically used.

Some of the other commonly used classifiers make assumptions that may not be suitable for the problem at hand. The ANN-MLP, for example, can address linear and piecewise linear problems, even some non linear situations, but it cannot handle non separable cases (Haykin, 1994). The SVM has been proposed to overcome this problem by assuming linearly separable classes in higher-dimensional feature space (Boser et al., 1992). However, its computational cost increases rapidly with the training set size and the number of included support vectors. As a binary classifier, multiple SVMs are required to solve a multi-class problem (Duan and Keerthi, 2005). Their approach suffers from slow convergence and high computational cost, because they first minimize the number of support vectors in several binary SVMs and then share these vectors among the

machines. However, in all SVM approaches the assumption of separability may also not be valid in any space of finite dimensionality (Collobert and Bengio, 2004).

Statistical classifiers, such as Bayesian systems, aim to find the decision boundaries assuming that the samples of the dataset have a probability density function (pdf) conditioned on the pattern classes (Jain et al., 2000). Thus, a decision rule can be used (Bayes or maximum likelihood, for instance) to determine decision boundaries. The main drawback of such recognition systems is the classification error present in the estimation of the pdf's.

The Optimum-Path Forest (OPF) classifier was recently proposed as an alternative approach to overcoming the problems highlighted. It is in fact, a simple, multi-class and parameter independent supervised pattern recognition technique that does not make any assumption about shape and can handle some degree of separability between classes (Papa et al., 2008). The OPF classifier interprets a training set as a complete graph, identifies prototypes in all classes, and computes an optimum-path forest rooted at them. The class of a sample in a tree is assumed to be the same at that of its root. A test sample is classified by identifying which tree would contain it. Because of these attractive features its classification effectiveness is tested in the emotional recognition task described in this work.

1.10 Research Objectives

The current research has three main goals:

a. To determine if the glottal signal carries emotional content and if it can be used as an emotion recognition factor in multi-class emotion detection;

b. To compare the performance of the glottal signal to already proposed feature domain signals; and
c. To establish the robustness of the new emotion recognition method in more realistic noisy scenarios.

The problem of emotion recognition in noisy environment has not been thoroughly investigated and literature is scarce on the matter, which gave the main motivation behind this work. Most research has dealt with recordings prepared in a specific well-controlled studio environment. This study addresses the problem of emotion recognition from different aspects as it applies different noisy conditions on the sound source in both training and testing conditions. It simultaneously investigates the problem in both controlled studio environments and "real world" noisy ambiance, thus bringing it closer to practical applications.

The research platform will include: clean, noisy, and severely degraded speech through filtering. Data will be arranged in three different sets for testing namely: utterance-based balanced speaker and text independent test; glottal symmetry-based balanced speaker and text independent test; and unbalanced speaker independent tests that included all emotions, all utterances, all speakers.

Six basic emotions are of interest: *anger, happiness, sadness, neutral, fear, and surprise*. As established by Bosch (2003) in Section 1.2 the six emotions of choice are more difficult to separate, which presents an additional challenge in this study. They will be tested in three different speech corpora. In the first two corpora, only the first four emotions were included for evaluation, while the third corpus will examine all six. Five feature domains will be examined for comparison, both novel and classical. The classical

prosodic feature domain is based on the extraction of *pitch, energy, velocity and acceleration of pitch, and length*. The ToBI feature domain will be using tone tier elements. The other three proposed domains will investigate glottal symmetry and MFCCs of the speech and glottal signals. The task is to include automatic feature extraction for all methods of interest, a task which can be far from trivial. An attempt will be made to set a system that addressed the issues of feature extraction and selection trough investigating the mutual information of the attribute set, while improving the recognition success rate of the supervised learning schemes. Once the feature attributes are finalized in each of the domains, they will be tested in the third corpus, which includes six emotions and it will used to demonstrates feature robustness under noisy conditions in a real-world, multi-speaker emotional exchange.

1.11 Organization of This Dissertation

The remainder of this work is organized as follows: Section 2 describes the production of voiced speech, while Section 3 deals with design of the system. Section 4 explains the extraction of features in more details, and Section 5 reveals how the final feature set was selected. Emotion modeling is included in Section 6 and all results are displayed and discussed in Section 7. Section 8 summarizes the work done in this dissertation and finally, conclusions and contributions of this research are included in Section 9.

CHAPTER 2

Production of Voiced Speech

There are several factors to be considered when discussing the production of voiced speech. The glottal flow, the vocal tract and the formation of formants are some of the most important properties that determine the quality of speech, which is why they are addressed in this chapter. Due to changes on any of these properties, voiced speech is shaped in ways that give us meaningful information not only about parts of the speech by constructing different phonemes, but also about age, gender, and emotion. Frequency range, for example, is different for male, female, and child speakers. While the literature reports slightly different measurements for the fundamental frequency or the pitch F_0, most would agree that for typical adult male speakers it varies roughly between 80Hz and 150Hz, while for female speakers that range is between 160Hz to over 240Hz, and for a child speaker it is usually between 180Hz and 300Hz.

2.1 The Glottal Flow

The source for the production of voiced sounds is the airflow passing through the glottis, modulated by the oscillation of the vocal folds. There is a great deal of variation

in glottal airflow and it results primarily from the speaker's control of the degree of tension on the vocal folds (Titze, 1994). The main parameters of the glottal signal are shown in Figure 2.1:

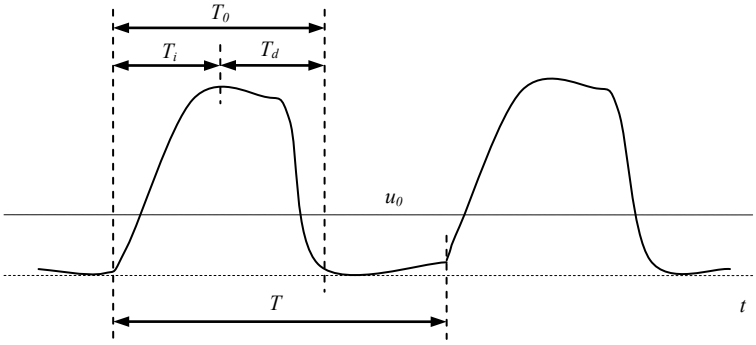


Figure 2.1: Glottal pulse and its main parameters.

where: T_i is the time increasing slope, T_d is the time decreasing slope, T_0 is the time duration of the flow, T is the time period of one glottal pulse, and u_0 is the average glottal flow. As given by Titze, two main dimensionalities can be found based on the parameters above:

$$\underline{Q_0} = \frac{T_0}{T} \qquad \text{and} \qquad \underline{Q_s} = \frac{T_i}{T_d} \tag{2.1}$$

where, Q_0 is the open quotient and Q_s is the skewing quotient. The shape of the glottal pulse is defined by the open quotient and skewing quotient and it represents the amount of power the glottal pulse holds. The two measures greatly affect the timbre of the speech as well, since the glottal signal is at the root of the quality of speech.

In American English, voiced phonemes include: vowels (*/a/ - father, /o/ - obey, e - /hate/, /c/ - all, /U/ - foot, /u/ - boot, /i/ - eve, /I/ - it*), semivowels (*/y/ - you, /w/ - we*), voiced plosive consonants (*/b/ - be, /d/ - day, /g/ - go*), voiced fricative consonants *(/v/ -*

vote, /D/ - then, /z/ - zoo, /Z/ - azure), nasals (*/m/ - me, /n/ - no, /G/ - sing*), and diphthongs (*/Y/ - hide*, */W/ - out*, */O/ - boy*, */JU/ - new*).

The source of voicing is the glottal signal, which is a low frequency signal as shown in Figure 2.2. From the power spectrum density estimated below, we see that most of the energy of the glottal pulse is located in the lower part of the frequency spectrum. The glottal signal bears important information about gender, age, and even emotion as will see later in this work. It is a quasi periodic pulse train with period $T_0 = \frac{1}{F_0}$, where F_0 is the pitch frequency, and typically it has only positive value and zero offset, although these two properties may vary in fluent speech.

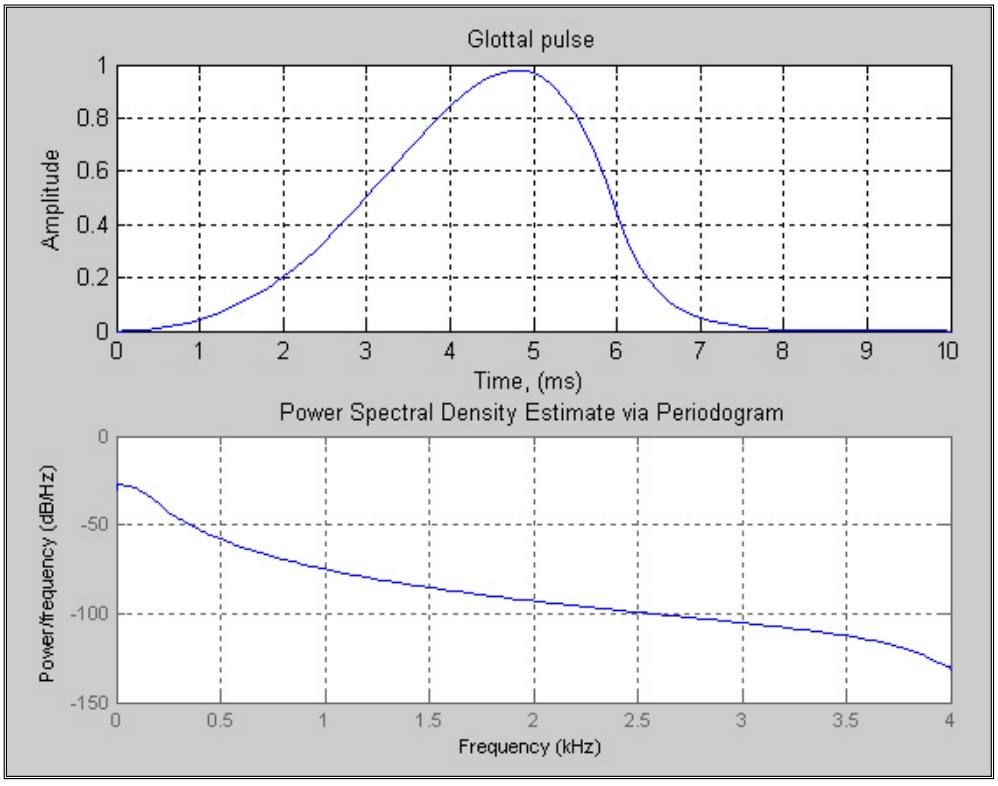


Figure 2.2: Glottal pulse and its estimated power spectral density.

2.2 The Vocal Tract

The vocal folds oscillate due to the air being expelled from the lungs and that in turn generates the glottal pulse signal. Once it has been produced, the signal passes through the vocal tract, which includes the pharynx, the oral cavity and also the nasal track. The spectrum of the signal becomes more colorful or harmonically enriched since the glottal signal passes through the acoustic cavities. For analytical simplicity, the entire acoustic system can be viewed as a set of concatenated tubes, beginning at the glottis and ending at the mouth where the sound energy is transmitted in the open space and is perceived as speech. Since the shape of the vocal tract changes as different sounds are articulated and it varies for each individual, it is natural to anticipate that some frequencies will be more favorably reinforced than others. It is therefore expected that those reflections have different characteristics, which defines the harmonic coloration or the timbre of the speech for a particular person. Changing the shape of the vocal tract changes its physical parameters, and in turn leads to the production of different phonemes. Detailed image of the larynx is depicted in Appendix B. For more information on how the shape of the vocal tract influences phoneme production, see Titze (1994).

2.3 The Nature of the Formants

The spectral peaks in the spectral envelope of the magnitude spectrum of speech are known as formants. Male speech exhibits one formant per 1000 Hz, on average, while female and child speech have about one formant per 1500 Hz and 2 kHz, respectively, on average (Rabiner and Schafer, 1978; Quatieri, 2002). Formant locations are important features for the phoneme perception in auditory analysis. Spectral analysis of the two

vowel sounds /a/ and /i/ for male speaker are shown in Figures 2.3 and 2.4. There typical locations of the first three formants, F_1, F_2 and F_3 for /a/, /u/, /aw/, /e/, /i/ are depicted in Table 2.1. It has been established that the first two formants F_1 and F_2 play an important

Table 2.1: Frequency centers for formants F_1, F_2 and F_3.

Vowel	Gender	F_1 (Hz)	F_2 (Hz)	F_3 (Hz)
a	*male*	730	1090	2440
	female	850	1220	2810
u	*male*	300	870	2240
	female	370	950	2670
aw	*male*	570	840	2410
	female	590	920	2710
e	*male*	530	1840	2480
	female	610	2330	2990
i	*male*	270	2290	3010
	female	310	2790	3310

role in vowel classification (Peterson and Barney, 1952). F_0 is reserved for the pitch frequency. For the vowel sound of /a/ we see that the first two formants are very close to

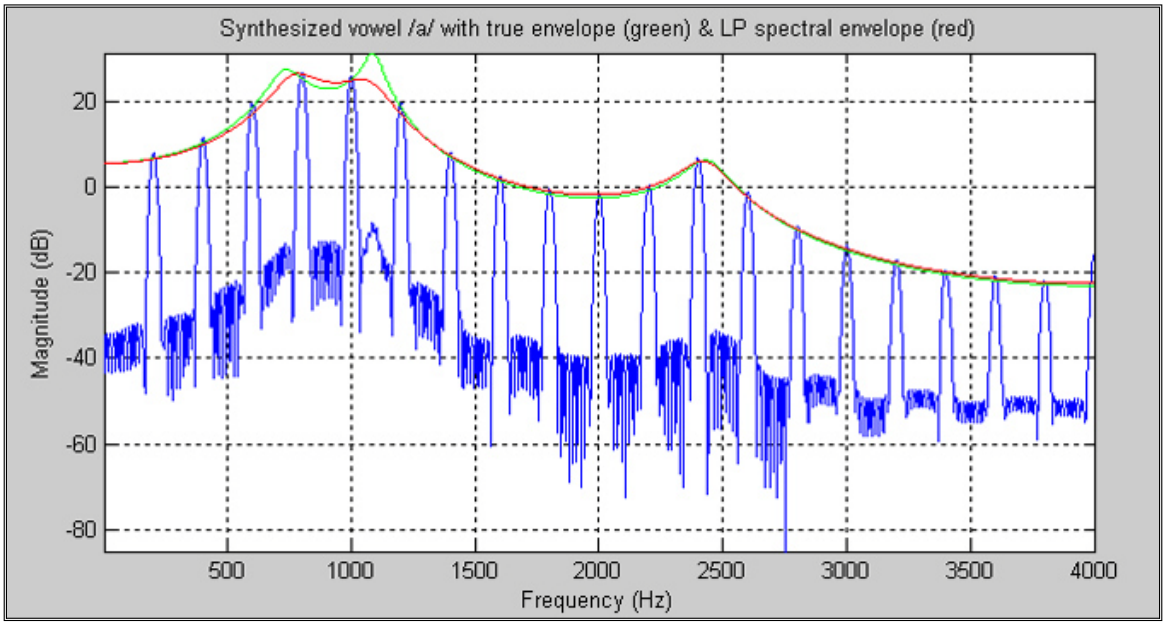


Figure 2.3: Frequency spectrum and Linear Prediction Coefficients for a synthesized vowel /a/.

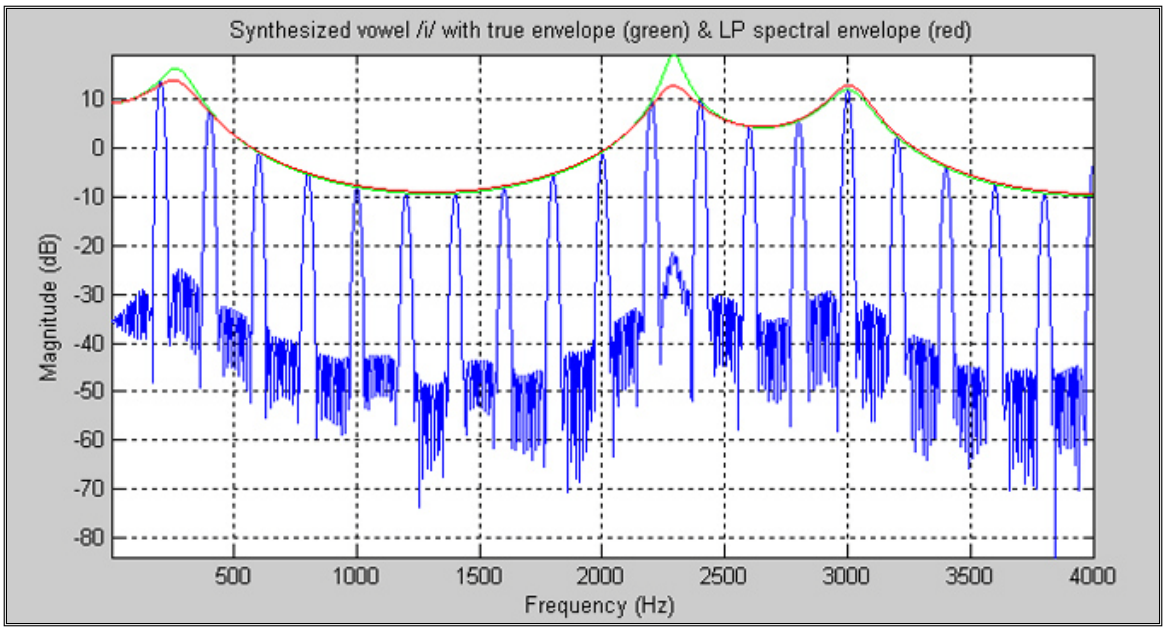


Figure 2.4: Frequency spectrum and Linear Prediction Coefficients for a synthesized vowel /i/.

one another. For the vowel sound of /i/ however F_0 and F_1 are a significant spectral distance apart. That difference reflects the articulation changes needed to produce these two vowel sounds. The frequency centers for the first two formants are given in Table 2.1.

The speech waveform and spectrogram for synthetic male vowels /a/ and /i/ are shown in Figures 2.5 and 2.6. One can clearly see the first and second formants in both vowels. A clear distinction should be made between harmonics with formants. Harmonics are multiples of the fundamental frequency F_0, while formants represent frequency spectral peaks. It is the reason why harmonics can coincide with formants, which leads to boost in the intensity of the later. Normally several harmonics overlap with one formant. This is known as *formant tuning* and as expected many singers are trained to take advantage of it to gain stronger voice intensity.

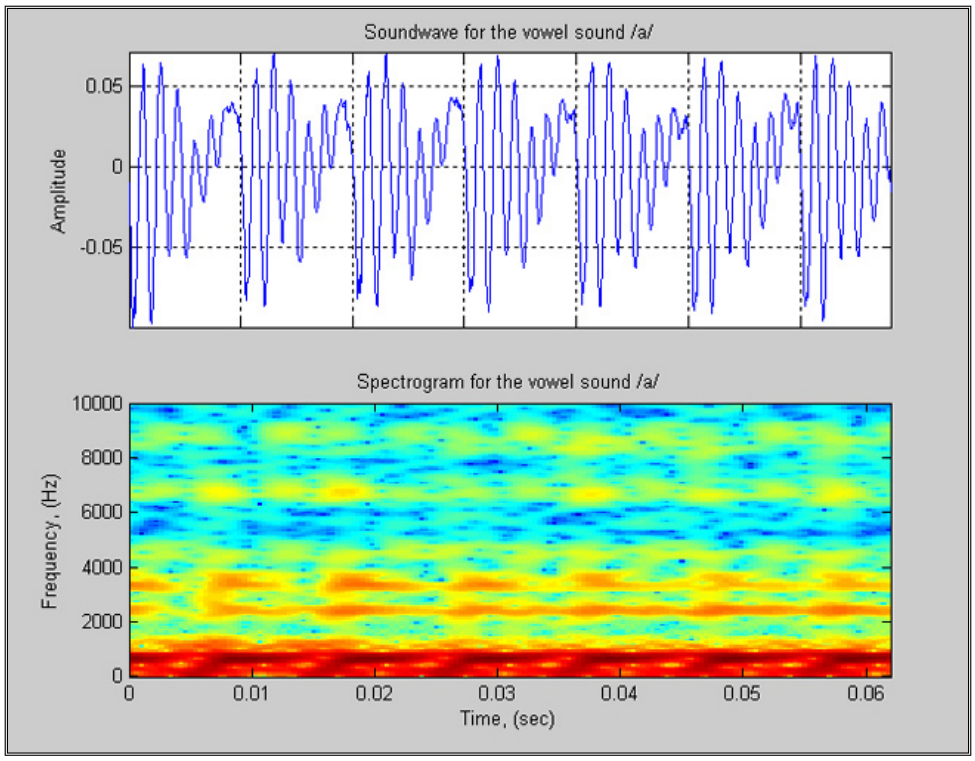


Figure 2.5: Speech waveform and spectrogram for spoken vowel /a/.

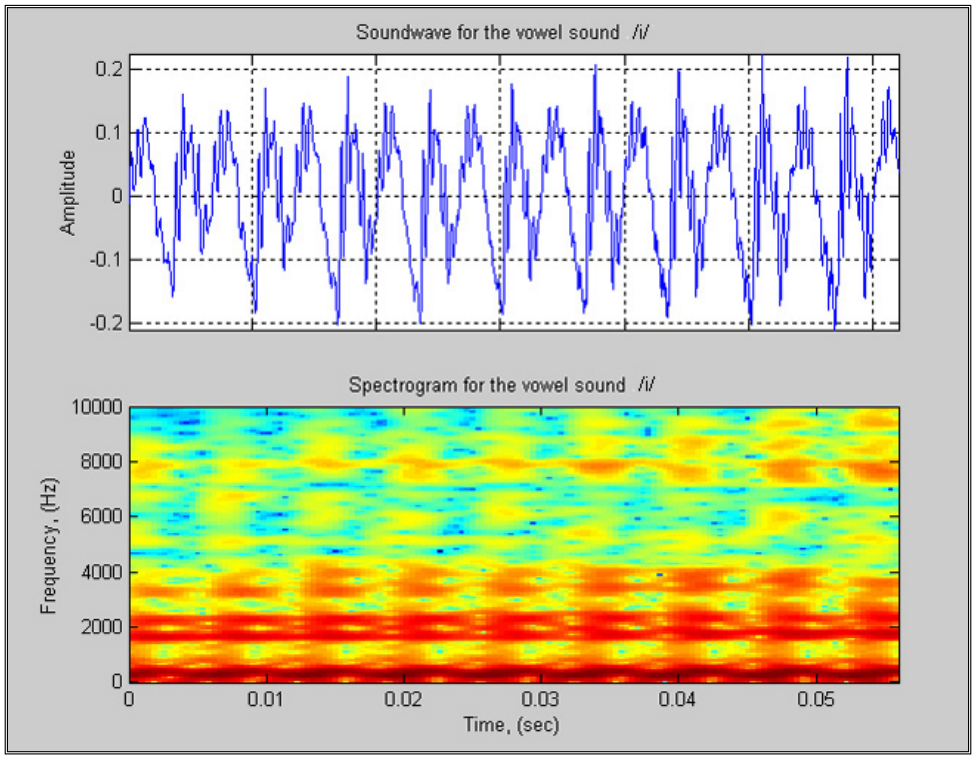


Figure 2.6: Speech waveform and spectrogram for spoken vowel /i/.

CHAPTER 3

Speech Database Preparation

3.1 Speech Corpora Design

Currently in the field of emotion recognition of speech there is no single specific dataset widely accepted as a research standard to be used for benchmarking. Most of the independent research in emotion recognition points to specifically prerecorded corpora, which only meet requirements pertinent to a specific study. Thus the available databases of emotional speech differ greatly in the number of speakers used, the number of emotions, the type of recording setup, type of speech; natural or acted, number of utterances, the purpose; recognition or synthesis, or by the language of choice. Ververidis and Kotropoulos (2006) summarized a total of 64 emotional speech data collections which included a total of 29 different emotions, but none of them met the specifications required for the current study. In particular, six emotional classes were not present in any but in only four corpora, three of which were in different languages. The one in English had very limited recording collection comprising of one speaker only. Creating a collection containing multi-speaker emotional recognition speech, which includes a good

variety of emotions, correct language of choice, or appropriate recording conditions, is very challenging and can be costly. The "Emotional Prosody Speech and Transcripts" dataset for example available at the Linguistic Data Consortium was a close match for the needs of this research. It still however provided very limited choice of speech examples, since all of the recorded utterances contained spoken dates and numbers only. On the other hand, the "Berlin Database" was a better match containing six emotions and good number of utterances, but it was recorded in German, thus remained out of the scope of this research. It was therefore decided that speech corpora must be created internally so that all the preset goals for this research could be addressed.

There were three corpora used in this study. The first speech corpus was based on a recording of acted speech recorded in the early 60's. That was the 100 minutes theatrical play "Waiting for Godot", written by Samuel Beckett in 1949 and released as a recording on April 3, 1961. This play has been voted as "the most significant English language play of the 20th century" (Berlin, N., 1999). The ages for the three male subjects were: *Subject 1* - Gogo / Estragon (Zero Mostel) age 46, *Subject 2* - Pozzo (Kurt Kasznar) age 47, *Subject 3* - Didi / Vladimir (Burgess Meredith) age 53, and *Subject 4* – Child (Luke Halpin) age 13. The test sets were split in two main groups: individual and combined. Manual labeling for four different emotions was performed on the original speech. To create the first corpus, the speech database was transcribed for four emotional classes: *happy*, *angry*, *sad*, and *neutral*. This corpus consists of three male speakers containing total of 2,252 emotional utterances (turns). Emotional states were detected in all three speakers. The play was originally recorded in analog and carried higher background noise levels. The speech was provided with sampling frequency of f_s =

22,050 Hz at single channel with 16 bits per sample linear quantization. For convenience and without loss of important information the corpus was downsampled to f_s = 8 kHz. The database analysis and labeling yielded the number of emotional utterances included in Table 3.1. In order to remove biases and to balance the corpus, 303 utterances from each emotion were used, which matched the smallest amount of utterances for *Happy*. For the development of the classification systems each set was randomly split in two parts: 80% for training and 20% for testing. The included utterance lengths varied in time between one and six seconds. In the labeling step, the beginning and end of each utterance for each emotion were time stamped. Instances with unclear or silent sections were removed. This corpus investigated only two feature domains, more specifically it dealt with investigating classical prosodic features and Tonal and Break Indices (ToBI).

Table 3.1: Emotional utterances for corpus 1.

Utterances per emotion		
1	ANGRY	403
2	HAPPY	303
3	SAD	318
4	NEUTRAL	1228

The second emotional speech corpus was collected in an anechoic chamber using condenser cardioid microphone RODE-NT2 with build-in HPF. Ten speakers; five male and five female in their early twenties, were fitted with neck impedance contacts. This way in addition to speech, the glottal signal was simultaneously recorded using the Glottal Enterprises EG2-PC laryngograph. The two recorded channels were sampled at 22,050 Hz with 16 bits linear quantization. The speakers were not professional actors and they did not receive special instructions about their speaking mode. Rather, the subjects

were only instructed to speak every sentence in one of the four emotional states of interest. They were provided with manuscript, which contained ten sentences - five short and five long as shown in Appendix C.

In the second corpus, the investigated features included the glottal symmetry and MFCC coefficients of various lengths, both on the glottal waveform and the corresponding speech signal. Spoken emotion features were presented for evaluation and tested on a new optimum path classifier (OPF) as well as on other previously established classification methods including: the Gaussian mixture model (GMM), support vector machine (SVM), k-nearest neighbor rule (k-NN), Bayesian classifier (BC), and C4.5. Experimental results from the clean speech indicate that best performance is obtained for the glottal-only features with SVM and OPF generally providing the highest recognition rates, while for GMM or the combination of glottal and speech features performance was relatively inferior. The top performing classifiers achieved perfect recognition rates for the case of 6^{th} order glottal MFCCs in clean speech. But when the same feature was tested in the third corpus in a noisy environment, it did not perform as well.

The voiced parts in the analysis were determined using the real part of the Hilbert transform and verified with the help of the Zero Crossing Rate (ZCR). Windows with length 64 ms were used on the voiced segments of each utterance to extract the glottal waveform via the inverse filtering method described in Wong et al. (1979). The glottal opening and closing phases of each period in that analysis window were automatically determined and later verified with the recorded glottal signal from the laryngograph. The tests performed in this setup were speaker and text dependent.

The third corpus used the same theatrical play as used in the first corpus, but this time the signal was manually transcribed for six rather than four emotions as described earlier. The additional two emotional classes were *fear* and *surprise*. Although they were both based on the same speech media, they differed not only by the number of emotions or the number of speakers used, but most importantly by the way the utterances were transcribed. In particular the lengths of the utterances for each emotional class were altered so that similar emotional instances were split, shortened, or combined together in the third corpus. This alone introduced number of changes to the speech media; hence it was treated as a separate corpus. The original recording included emotional speech from 5 speakers; 4 male speakers, and 1 child. They all spoke in random order. The small portion of the child's speech mainly contained *neutral* emotion. The number of combined emotions available from this corpus was 2,368 and their distribution is displayed in Table 3.2.

Table 3.2: Emotional utterances for corpus 3.

#	Emotion	Subject 1	Subject 2	Subject 3	Subject 4	Total
1	ANGRY	141	161	186	0	488
2	HAPPY	125	48	106	0	279
3	SAD	237	31	127	3	398
4	NEUTRAL	373	181	251	50	855
5	FEAR	41	29	55	2	127
6	SURPRISE	70	40	111	0	221

Extensive testing of all feature domain attributes from the first two corpora was performed, namely, Glottal Symmetry (GS), ToBI, MFCC, and classical prosodic features. Depending on the view point, there was a series of different category test performed on the corpus:

1. From signal quality stand point, tests included:
    - clean speech;
    - noisy speech with 10dB and 30dB SNR;
    - lowpass filtered (LPF) speech, which imposed sever degradation of the speech to the point of making it incomprehensible.
2. From data presenting and arrangement point of view, the tests included were:
    - utterance-based unbalanced, speaker dependent for each individual speaker;
    - utterance-based balanced, speaker and text independent with 100 utterances per emotion, and random number of speakers;
    - glottal symmetry-based balanced, speaker and text independent both per emotion and per speaker;
    - unbalanced speaker independent test including all emotions, all utterances, all speakers.

By testing corpora one and two first, reasonable information was collected to establish that emotional content is indeed available in all different feature domains employed. It also provided fine tuning of the amount and type of attributes to be used by further pruning. To meet the tougher requirements of a more uncontrolled, realistic environment, the use of the third corpus was necessary to extend the search for the most robust feature domain of all, and also to investigate class separability among six emotional instances. It has to be noted that the original recording for corpora 1 and 3 carried higher background noise levels than the one in corpus 2.

3.2 Glottal Symmetry

As already established in Section 2, the glottal signal is the source for voiced speech production. The shape of the glottal pulse has been well defined and several models specifying its phases from geometric point of view are available. As summarized by Hardcastle and Laver (1999) there are several widely adopted versions as proposed by Rosenberg (1971), Hedelin (1984), Fant (1979, 1982, 1985), Ananthapadmanabha (1984), and Ljungvistand and Fujisaki (1985). The model of choice adopted for use in this study is the one proposed by Fant, since it provides better analytical details of the typical shape of a glottal pulse. It is displayed in Figure 3.1, where U_0 is the peak volume velocity of the glottal pulse, which occurs at t_p, T_o is the opening phase of the pulse, T_c is its closing phase, FG is defined as the inverse of the glottal pulse width and it signifies the glottal frequency. These parameters can be expressed like:

$$T_c = \left(\frac{1}{FG}\right)\left[\frac{\cos^{-1}\left[\frac{(k-1)}{k}\right]}{2\pi}\right], \qquad T_o = \frac{1}{FG} - T_c, \tag{3.1}$$

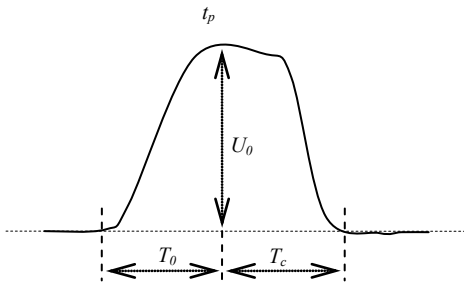


Figure 3.1: Glottal shape model as proposed by Fant (1979).

The glottal symmetry *GS* is given as the ratio between the closing phase over the opening phase: $S = \frac{T_c}{T_0}$. However in a more detailed version of the glottal pulse (Fant 1985), when the rate of change (its derivative) is considered, there is an extra constant T_a which is the time instant representing the return phase of the flow before complete glottal closure occurs. This is depicted in the Figure 3.2. In this work T_a is included in the closing phase T_c.

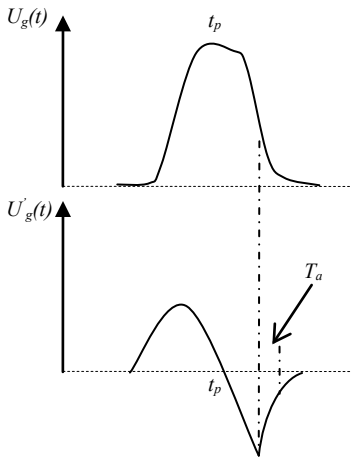


Figure 3.2: Glottal shape in terms of its differentiated glottal flow (Fant 1985).

Several other glottal parameters may be considered, including: opening quotient, closing quotient, the difference of opening-to-closing phase, their ratio, and the first derivatives of the ratio, and both phases. However, by observing and analyzing laryngograph plots under different emotions it appears that the shape of the opening to closing phases and their relationship is most susceptible to difference in speaking styles and emotions. On the other hand, in a real life scenario in the presence of noise, it is difficult to precisely determine the exact glottal closure and opening instances via inverse filtering techniques as it will be shown later in this work. Sometimes complete closing of the glottis, typically in female speakers, may not even occur. It was therefore more

practical to use the maximum and minimum waveform points to compute the closing phase **C** = $|\mathbf{A_2}\text{-}\mathbf{B_2}|$ and opening **O** = $|\mathbf{B_1}\text{-}\mathbf{A_2}|$ phases, as shown in Figure 3.3.The glottal symmetry defined as $GS = \frac{C}{O}$ proved to be a very effective feature in discriminating between speaking styles. The distribution of the glottal symmetry ratio for each emotion and for each speaker is shown in Figure 3.4, which was generated with the *Weka* program (Witten and Frank, 2005). There, the first five plots correspond to the male speakers, and the rest to the female speakers.

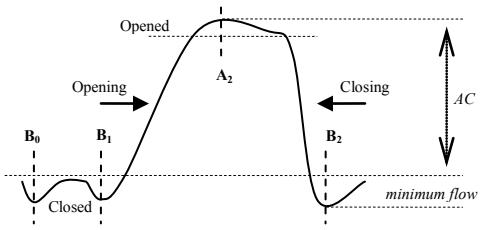


Figure 3.3: Glottal pulse with its four phases: Opening, Opened, Closing, and Closed.

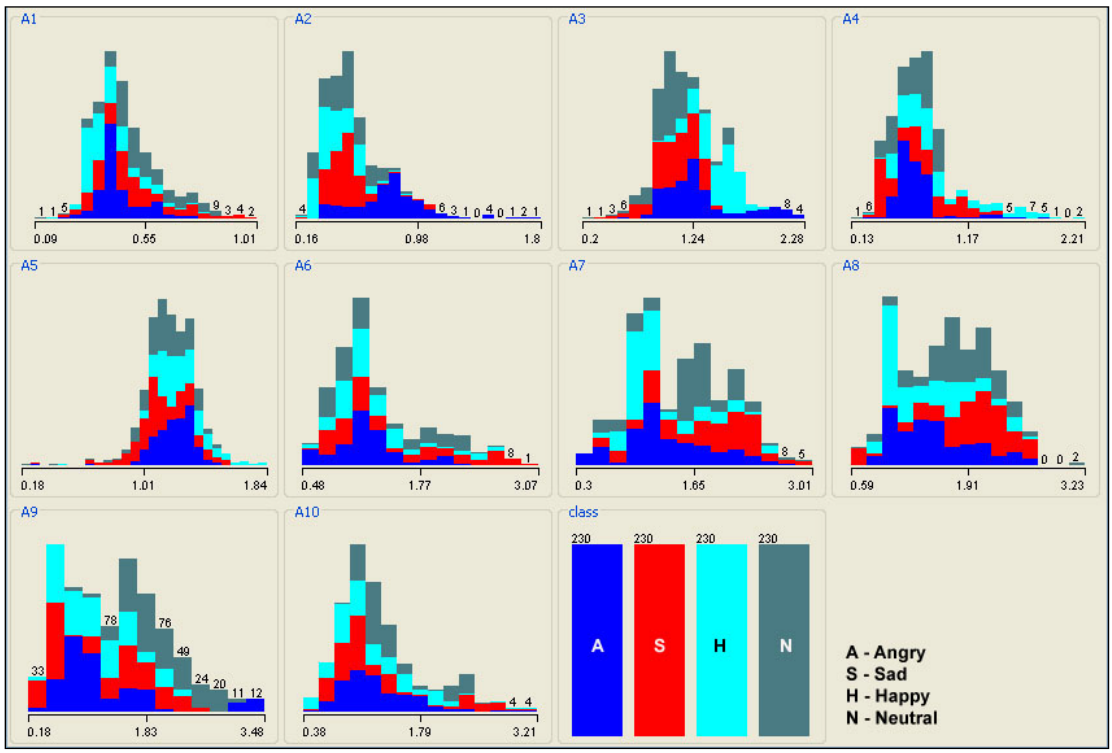


Figure 3.4: Glottal Symmetry data distribution per 10 subjects.

3.3 Realization of the Classical Approach

There are many different elements that can be considered when extracting features using classical techniques. There is evidence, suggesting that only the use of few combined features can contribute for more accurate classification Wang (2004), and yet the excessive use of many elements degrades system performance. In Wang (2004), extensive testing of many features was performed using Mel-Frequency Cepstral Coefficients (MFCC), prosodic and formant frequency features. Considering that one of the classification methods used in this work was GMM, and based on the data provided in Wang (2004), there were only used 11 prosodic features used for the realization of the classical approach. There were six pitch related and five energy related features all shown in Table 3.3.

Table 3.3: Classical Prosodic Features.

Classical Prosodic Features	
Pitch elements	
1	mean
2	median
3	standard deviation
4	maximum
5	rising-falling
6	maximum of falling range
Temporal Energy	
7	mean
8	standard deviation
9	maximum
	Durational features
10	average pause length
11	speaking rate: number/length of voiced segments

The feature vector can be represented as F_c:

$$\underline{F_c} = (p_1,..., p_6, e_1,..., e_3, d_1, d_2) \tag{3.2}$$

where, *p* represents pitch elements, *e* represents the energy features and d represent durational features from Table 3.3. The six pitch elements were extracted using the Simple Inverse Filter Tracking - SIFT algorithm (Markel, 1972), and they were: *average*, *median*, *standard deviation*, and *maximum of pitch*, as well as *rising-falling pitch ratio*, and *maximum of falling pitch range*. The extraction of the first four was based on computing the data from each emotion and it is easy to follow. The *rising-falling pitch ratio* was calculated following the pitch deviation for each sequence with no interruptions. Then, a count of all rise–and-falls was taken and finally the ratio was computed. The *maximum of falling pitch range* represented the maximum drop off in pitch for a sequence with no interruptions.

The energy and durational features were extracted in the time domain. They were: *mean energy*, *standard deviation*, *maximum energy*, *average pause length*, and *speaking rate*. The latter was defined as the ratio of the number of voiced segments to their total length. The *average pause length* was determined by using end-point detection. The extraction of the first three features is a straight forward calculation based on the data extracted from the particular state.

3.4 ToBI Elements

The only features of interest when studying the first corpus were prosodic and therefore the only useful ToBI tier was found to be the *tone* tier. In short, the *break* tier

describes the degree of junction that may occur between pair of words and it includes the silence after the last word of a given utterance. This way the break indices represent the prosodic groupings in an utterance. In the case of this research they were bound and detected by continuous voiced regions, thus forming the intermediary phrasing or (ip). The break indices can be determined only after all the words have been transcribed, which is done in the orthographic tier Beckman (1997). Orthographic transcription was not available for the corpus used. Since the *orthographic* and *break* tiers could not contribute to the focus of this work they were omitted from further investigation. Given that pitch represents the core of any prosodic analysis all extracted ToBI elements were based on pitch information from the *tone* tier alone.

There were 16 targeted features for extraction in the ToBI *tone* tier. According to the type of all ToBI features in the *tone* tier, they were allocated in five different categories shown in the Table 3.4.

Table 3.4: Targeted ToBI elements for extraction from the ToBI *tone* tier.

Sets of symbols:		**ToBI elements:**	
1. Pitch Accents:	Monotonal accents	L*	lowest pitch (monotonal PA)
		H*	highest pitch (monotonal PA)
	Bi-tonal accents	L*+H	scooped accent (bitonal PA)
		L*+!H	scooped accent with lower **H** (bitonal PA)
		L+H*	rising peak accent (bitonal PA)
		L+!H*	rising peak accent with lower topline **H*** (bitonal PA)
		H+!H*	downstepped accent (bitonal PA)
2. Boundary Tones		L%	final low boundary tone
		H%	final high boundary tone
		%H	initial mid-pitch boundary tone
3. Phrase Accents		L-	low phrase accent
		H-	high phrase accent
4. Phrase Accents and Boundary Tones Combinations		L-L%	very low point in the speakers range
		L-H%	low phrase accent followed by high boundary tone
		H-L%	high phrase accent slightly lowered at the boundary
		H-H%	extremely high in the speakers range (upstepping)
		!H-L%	downstepped high phrase accent slightly lowered at the end
5. Downstep		!H*	downstep - compression of pitch range or lowered topline

All ToBI features comprise the feature vector F_T, expressed like:

$$\underline{F_T} = (t_1,...,t_{16}) \tag{3.3}$$

where, t – represents all ToBI features found in Table 3.4. The combined feature vector will then look like:

$$\underline{F} = (p_1,...,p_6,en_1,...,e_3,d_1,d_2,t_1,...,t_{16}), \quad \text{or}$$

$$\underline{F} = [F_c F_T] \tag{3.4}$$

3.5 MFCC Computation

The frequency scale represented in the auditory signal analysis in the basilar membrane can be applied in the extraction of signal features which are perceptually important and have proven effective in speech classification. Such is the pitch or mel scale, which is a non-linear representation of the linear frequency scale. The mapping from Hertz to mel can be approximated using:

$$F_{mel} = \left[\frac{1000}{\log_{10}(2)}\right] * \left[\log_{10}\left(1 + \frac{F_{Hz}}{1000}\right)\right] \tag{3.5}$$

where F_{mel} is the nonlinear frequency in mel and F_{Hz} is the linear frequency in Hertz. The formula depicts the nonlinear behavior of the mel scale above 1000 Hz according to human perception (Fant, 1968). At all frequencies below 1000 Hz we can roughly assume a 1:1 correspondence between mel and Hertz scales. The mel-to-mapping is given in Figure 3.5. To simulate the subjective spectrum of the mel scale, a filter bank approach

can be used, where one filter represents each critical band. Each filter bank can typically be approximated a triangular shaped bandpass magnitude response. The spacing andbandwidth of each band is such that maintains constant width and spacing on the mel scale. However, on the Hertz frequency scale the filters are positioned closer to one another and are narrower in the low part of the spectrum. They gradually spread apart as the frequency increases, thus in the high portion of the spectrum they have wider bandwidths.

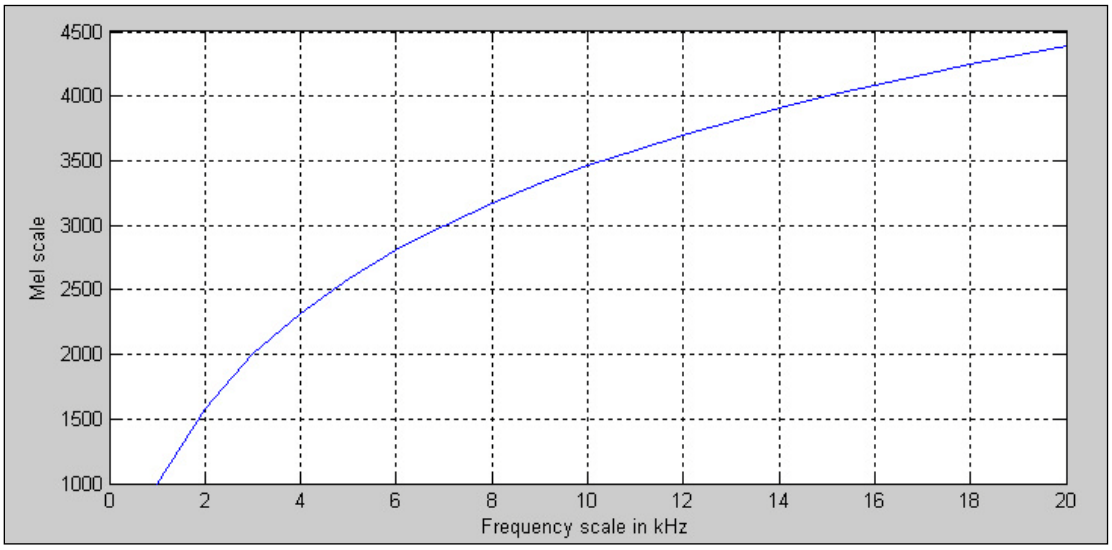


Figure 3.5: Mel-Hertz frequency plot.

The base of the triangular window for each band is the defined by its center frequency. In Davis and Mermelstein (1980), 20 triangular bandpass filters are used for band limited signals with 4 kHz Nyquist frequency. The filter banks are distributed as follows: 10 linearly spread between 0 and 1kHz, the next 5 are logarithmically distributed between 1 kHz and 2 kHz and the last 5 are spread in the 2-4 kHz band also following logarithmical distribution. Each mel filter bank, as applied in frequency domain, takes the

weighted sum of the spectrum for that band, thus resembling a histogram bin. The mel-frequency cepstral coefficients or MFCC may be estimated by the expression:

$$MFCC_n = \sum_{k=1}^{20} (\log_{10} S_k) \left[n\left(k - \frac{1}{2} \right) \frac{\pi}{20} \right] \tag{3.6}$$

where n = 1, 2, … M and M is the number of mel cepstral coefficients, S_k is the energy output of the kth filter and k = 1, 2, … , 20. This procedure is applied at each frame of the signal. Frames may vary in length, but typically they are between 20 and 40 ms using 50 – 75 % overlap and Hamming-weighted.

3.6 GMM Deployment

The Gaussian Mixture Model (GMM) has been used extensively in many engineering applications in which data can be viewed as generated from multiple mixed sources of a certain type of distribution, based on a set of corresponding prior probabilities. In the GMM case, the distribution type is assumed to be Gaussian. Besides its inherent modeling power of fitting probability densities arbitrarily (Duda and Hart, 1973), it is made particularly attractive for statistical pattern classification applications by the use of the Expectation-Maximization (EM) procedure (Dempster et al., 1977) for GMM parameter estimation. There have also been many efforts reported using GMM in emotion recognition with good results as regards to the recognition rate. However, there are still some fundamental issues which are of great interest to investigate in the employment of GMM for emotion recognition task, like for any other GMM applications such as: model initialization, determination of the number of components, the estimation

of prior probabilities of classes, normalization of feature vectors and handling of singularity issue. In this work, in addition to recognition performance, the effect of using different a number of components, the normalization of feature vectors, and how to handle matrix singularity in EM computation, were investigated and reported. These super-parameters can make a significant impact on the computation load, convergence and system performance.

In emotion recognition, the feature space has generally high dimensionality with well over ten features being involved and such issues become more critical. The initial number of features contained in each of the systems subject to our study was: 11 classical prosodic features, 16 ToBI features, 5 consecutive Glottal Symmetries from each analysis window, MFCCs of the glottis and MFCCs of the speech signal, thus bringing the total number of feature domains to five. Since the classical and ToBI domains were combined in a new domain, for corpus three the tested domains were only four.

In this work, all GMMs for each different emotion states have diagonal covariance matrices instead of full covariance. The number of components used in GMM regarding the different system configurations related to the different dimensionalities of feature vectors and the size of training samples. To handle overflow in computing the covariance matrices and their inverse matrices as well, appropriate techniques were employed, including variance flooring (Bimbot et al., 2000), relative variance flooring (Zhang and Scordilis, 2004) and appropriate normalization of the features, as explained in Section 5.1.

3.7 Other Classification Methods

Part of our goal, as in any other machine learning task, is to make an optimal classification decision by using the least amount of features by relaxing computational constraint and yet finding the best set in description space that best matches the set of examples in the training set. One very important aspect of the training process is to avoid overfitting. That is why simple rule descriptions are used and more specific complex concepts are avoided in the training process. This will be later discussed by means of searching for mutual information among different features. The main idea was to first describe the dataset and find a shorter and simpler feature description, which also fits the training set. This is achieved through a heuristic process called *backward pruning* or *post-pruning*. One of the issues that post-pruning can resolve is in a situation where two features do not contribute much when used individually, but make a very informative and powerful bond when used together.

C 4.5

One of the most popular decision tree classifier in use today is C4.5. It is a tree-building system which uses *subtree rising* as part of its pruning process. It this scheme, the entire subtree of the most popular branch is raised, thus needing to reclassify all 'child' nodes (Witten and Frank, 2005). That procedure can make the process computationally very expensive. Of course there is always the question when to replace an internal node with a leaf, in which case the error must be estimated based on the training data. Then we consider all the instances that lead to each node, and assume that the majority of nodes that belong to the same class represent that node. Knowing the total number of instances N, we can estimate the error E, which represents the rest of the nodes

belonging to different classes for that node. So given certain confidence factor *c*, the confidence limit *z* can be obtained by:

$$P\left[\frac{f-q}{\sqrt{\frac{q(1-q)}{N}}} > z\right] = c \tag{3.7}$$

where, q is the true probability of error at the node, $f = \frac{E}{N}$ - is the observed error. In general C4.5 works well on real data while providing good accuracy, which is the reason for its widespread use.

Support Vector Machines

The basic advantage of Support Vector Machines lays in its very nature: the use of a linear approach to solve nonlinear problems. For that reason SVM is using nonlinear mapping to convert from the existing to another attribute space (Witten and Frank, 2005). In short, it creates a linear solution for linearly non-separable data sets in their original subspace by creating a so-called *maximum margin hyperplane*. This is done by calculating all the possible tree-factor products. The hyperplane represents the linear model. The goal in the newly created subspace is that a maximum separation between the classes is achieved. The *support vectors* are the ones that hold a minimum distance to the hyper plane, and so exclusively defining the maximum margin hyperplane as:

$$x = a + \sum_k \alpha_k y_k \left(b(k) \cdot b\right)^n \tag{3.8}$$

where, x is the outcome, k - the support vector, $b(k)$ – training instances representing the support vectors, y_k – the class of $b(k)$, a and α_k are determined by the learning procedure and describe the hyperplane, vector b represent a test instance. The expression: $b(k)\cdot b$ represents the dot product of one of the support vectors with a training instance. When using a hyperplane, every time a classification of an instance is made the dot product $\forall k$ must be calculated. This can be a rather expensive calculation in the subspace, so this process is performed before the nonlinear mapping in the original low-dimensional space. In these terms, n represents the number of factors in transformation. The other powerful feature of SVM is that overfitting is not likely to happen, which is important when dealing with small datasets. In general, overfitting is caused by extremely relaxed constrains in the decision margin. SVM represents the whole dataset, which is the reason why overfitting is not likely to occur.

Naïve Bayes Classifier

One implementation of the general Bayes Classifier is the so-called Bayes algorithm. There, the general conditional probability formulation of the Bayes's rule:

$$P[E \mid A] = \frac{P[A \mid E]P[E]}{P[A]} \tag{3.9}$$

takes the form:

$$P[E \mid A] = \frac{P[A_1 \mid E] \times P[A_2 \mid E] \times \cdots \times P[A_n \mid E] \times P[E]}{P[A]} \tag{3.10}$$

where, n is the number of attributes, E is the emotion classes, A represents the attributes, $P[A]$ is the probability of a given attribute, and $P[E|A]$ is the probability of certain class E conditioned upon feature A. We must note that the sum of all prior probabilities of all

classes is: $\sum P[E]=1$, i.e., it holds true without knowing the occurrence of any feature. In the expression above we can multiply all the probabilities, assuming the events are independent. This assumption is very simple or 'naïve' and in reality may cause a big problem in case only one of the elements in the numerator equals zero, thus making the final probability to be zero as well. This may transpire in the training process due to non occurrence of one feature, and is avoided by using the *Laplace estimation*, which adds 1 to all elements of the numerator in equation (3.10). In general naïve Bayes works well with real data, especially when redundant features are eliminated, which is a procedure also undertaken in this study (Witten and Frank, 2005).

k-Nearest Neighbors

One other simple instance-based learning classifier is k-NN. It is also known as a "lazy" learning, because unlike other methods which create assumptions as soon as the date had been seen, it first compares the new instance to what it already has in memory. This is done through calculating a simple distance metric to all known entries. The new class assigned is the one to the lowest computed distance from the known instances, which is why it is called *nearest neighbor*. In the case when there are multiple nearest neighbors used, the majority class of the nearest *k* neighbors is assigned (Witten and Frank, 2005).

Optimum-Path Forest

In this work the theory related to the Optimum-Path Forest classifier is described, in which its training set is thought of as a complete graph, whose nodes are the samples and the arcs link all pairs of nodes. The arcs are weighted by the distance between the feature vectors of their corresponding nodes. Any sequence of distinct samples forms a

path connecting the terminal nodes and a connectivity function assigns a cost to that path (e.g., the maximum arc-weight along it). The idea is to identify prototypes in each class such that every sample is assigned to the class of its most strongly connected prototype. That is, the one which offers to it a minimum-cost path, considering all possible paths from the prototypes. By estimating prototypes as the closest samples from distinct classes, the OPF can handle all three cases with the maximum arc-weight function. In the case of overlapping between classes, these prototypes will be class defenders in the overlapped regions of the feature space.

The OPF algorithm classifies the samples of the feature space by taking into account the connectivity between them, not only the simplest distance of their corresponding feature vectors, such as the k-NN classifier. Another question concerns with the optimality criteria, because OPF's classification rule is given by an optimal search of the whole feature space, avoiding some misclassifications due to the using of local decision functions. The OPF is a fast, simple, multi-class, parameter independent, does not make any assumption about the shape of the classes (such that ANN-MLP and SVM), and can handle some degree of separation between classes.

CHAPTER 4

Feature Extraction

4.1 Glottal Waveform Extraction

4.1.1 Speech Production Model

Voiced speech can be viewed as the output of a production system consisting of three concatenated linear and time-varying subsystems as shown in Figure 4.1.

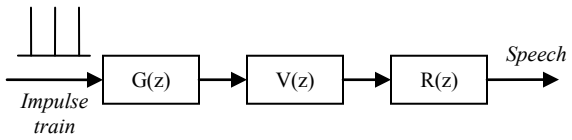


Figure 4.1: The speech production model.

During voicing the excitation is produced primarily at the glottis and it has a quasi periodic nature. The impulse train sequence models the timing of the air puffs through the glottis, which result from the oscillation of the vocal folds during the production of voicing. The timing of the glottal waveform controls the fundamental frequency of speech. The spectrum of the resulting speech measured at the lips, S(z), can be expressed in the complex frequency domain as:

$$S(z) = G(z)V(z)\,R(z), \tag{4.1}$$

where, G(z) is the glottal model, V(z) is the vocal tract transfer function, and R(z) is the effect of the radiation at the lips.

In order to obtain the glottal waveform spectrum from the available speech signal the vocal tract and lip radiation system functions need to be provided. For R(z), a simple first order filter is considered effective:

$$R(z) = 1 - \alpha z^{-1}, \tag{4.2}$$

where $0.95 \le \alpha \le 1.0$. If the radiation effects and the vocal tract transfer function can be adequately modeled then inverse filtering is an obvious method for estimating the glottal source model.

During the production of voicing the vocal tract can be modeled as an all-pole filter expressed as:

$$V(z) = \frac{1}{1 + \sum_{i=1}^{p} a_i z^{-i}}, \tag{4.3}$$

The coefficients of the V(z) filter can be readily obtained using a linear prediction (LP) analysis approach, such as the autocorrelation or covariance methods (Rabiner and Schafer, 1978), where *p* is the prediction order. Solving for G(z) from eq. (4.1) provides an inverse filtering estimation of the glottal signal as:

$$G(z) = \frac{S(z)}{V(z)R(z)}, \tag{4.4}$$

If the vocal tract model can be accurately determined over a short time basis then inverse filtering could provide the glottal signal provided the constituent parts of the speech production model are linearly separable and do not interact with one another. In reality, the vocal fold operation is affected by the vocal tract, which results in variations in the glottal volume flow and in the fine structure of the glottal wave shape. Furthermore, the glottal area variations, as observed using a laryngograph, do not always reflect in similar variations in the glottal airflow (Flanagan, 1972; O'Shaughnessy, 2000; Quatieri, 2002). Nevertheless, if only the chief characteristics of the glottal signal, such as the open and closed parts and the ratio between opening and closing phases, is what is important, rather the fine details of the waveform, then inverse filtering can be effective in providing the required glottal information. In this work, as it will be discussed in the following sections, inverse filtering can provide the level of accuracy required to use glottal information effectively for the specified emotion classification task.

In an experiment, the glottal contact area signal obtained from speakers fitted with neck impedance contacts using the EG2-PC laryngograph by Glottal Enterprises and the associated speech were digitally recorded with at sampling rate of f_s = 22,050 Hz. Temporal alignment between the two signals was achieved by considering the time difference between glottal events and speech recorded by a microphone located at distance d from the mouth of the speaker with vocal tract length l. Then the time difference in sampling periods n_0 between the source and the recorded speech signal is given as:

$$n_0 = \frac{(d+l)f_S}{c} \quad , \tag{4.5}$$

where c = 343 m/s is the speed of sound in air. Examination of the two signals confirmed the relationship between the glottal closing instant and the region of maximal disturbance in the speech signal.

A typical glottal pulse is shown on Figure 3.3. Points **A** and **B** describe the highest and lowest values in the waveform within a period. In Figure 4.2 an actual extracted glottal signal by using inverse filtering is shown above and its corresponding speech waveform of the voiced signal is depicted below.

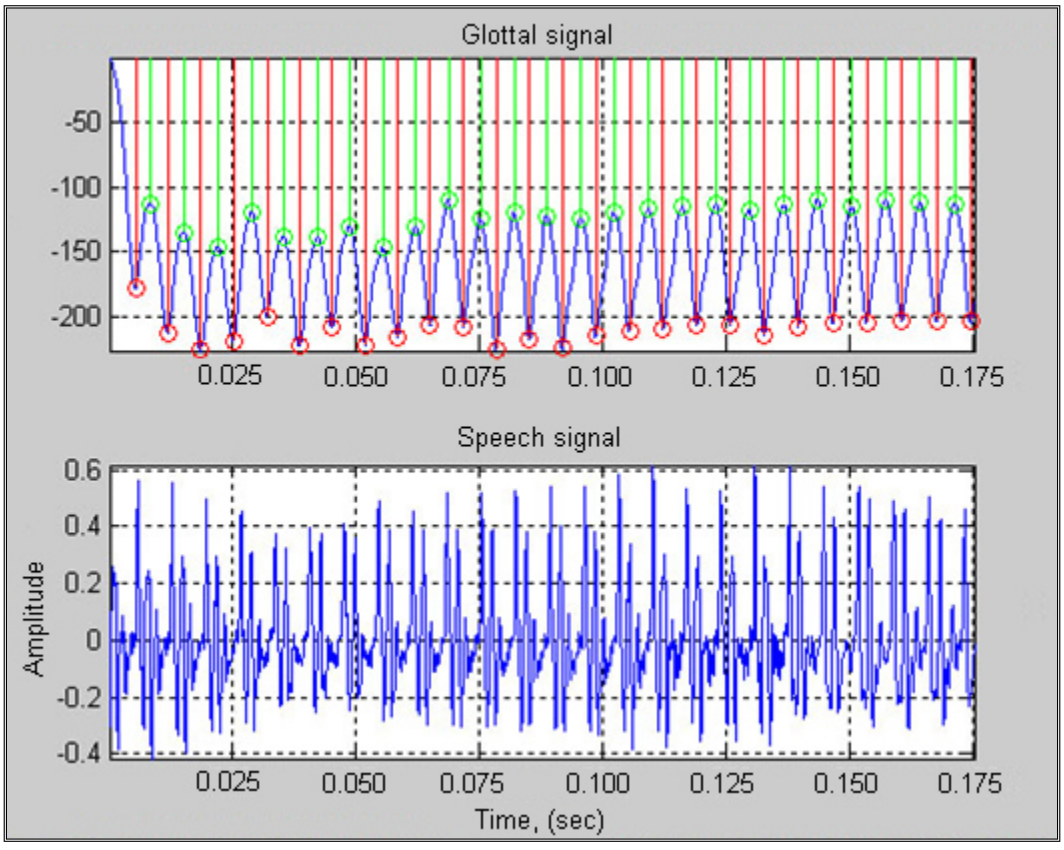


Figure 4.2: Extracted glottal signal and its corresponding speech signal.

4.1.2 Glottal Waveform Inverse Filtering

Inverse filtering can be an effective, practical method for the estimation of the glottal volume velocity waveform from speech. The quality of the obtained result is critically dependent on the accuracy of the estimation of the properties of the supra glottal part of the production system, which will subsequently be filtered out from the speech signal in order to provide a copy of the glottal waveform. An important factor that also influences the shape of the glottal airflow is the phonation type.

Many methods dealing with inverse filtering have been proposed (Rothenberg, 1973; Wong et al., 1979; Moore et al., 2003; Brookes et al., 2006), but the main studies in the area are based on two basic procedural groups according to the way the volume velocity waveform was recorded: recorded in the mouth by Rothenberg (1973), and recorded outside (away from) the mouth, thus accounting for the radiation of the lips as in Wong (1979). Quatieri (2002) described a model of the source/track interaction, which approaches the coarse and fine structures of the glottal flow derivative separately. The Liljencrants-Fant (LF) model (Fant, 1986) consisting of seven parameters is used to represent the coarse structure of the flow derivative, while the fine structure (ripple) component is estimated by subtracting the coarse model from the glottal flow derivative obtained via inverse filtering.

In a very detailed study, Rothenberg (1973) used a specifically designed mask to record the volume velocity at the mouth during voicing. Analysis was restricted to frequencies up to 1 kHz. The first two formant frequencies and bandwidths were estimated with the aid of narrowband spectrogram and used to filter out the vocal tract contribution so that the glottal signal could be obtained. Although the setup used in this

study is quite restrictive and it cannot be applied on pre-recorded speech databases, it nevertheless provided great insight into the properties of the glottal waveform. Because this technique uses a flow mask, naturally one of the most important contributions is that the resulting glottal volume velocity waveform bears useful amplitude information. Assuming that the recording device is adequately calibrated then both the *minimum flow* and the *AC-flow* (see Figure 3.3) of the glottal volume velocity waveform can be reliably obtained after inverse filtering. As stated by Rothenberg, in practice inverse-filtering is typically limited to slightly nasalized or non-nasalized vowels. The resonances of the vocal tract are represented by complex conjugate pole pairs and are referred to as *formants*. The effect created by the formants has to be cancelled (inverse-filtered) by introducing a complex zero to every complex pole in the vocal tract, plus a pole or a first-order resonance at zero frequency. The advantages of this method are: impervious to low-frequency room noise; the obtained signal achieves accuracy to zero frequency; and better calibration of the amplitude levels using constant air flow. The main disadvantage is that it is performed in laboratory conditions using specialized tools, thus making it non-practical for processing speech recorded under normal conditions.

Several techniques have also been proposed dealing with the estimation of the glottal volume velocity waveform from the acoustic pressure speech signal via inverse filtering. Because of the substantial coupling between the acoustic properties of the glottis and those of the dynamically changing supra glottal section the properties of the vocal tract need to be estimated when the glottis is closed. Therefore, the fidelity of these methods is based on the reliable estimation of the glottal opening instants (GOI) and the glottal closure instants (GCI). Generally the closure of the glottis is more abrupt than the

opening. This can be observed in Figure 4.3, where a drastic change of energy of the glottal signal is evident. This fact alone makes the identification of GCI mush easier than the GOI. It also puts higher precision constraints of the glottal signal extracting algorithm, since the changes are short in time. This leads to the necessity of correctly choosing the size of the analysis window. The analysis is usually done in small window frames of between 30 and 60 ms so that few glottal cycles can be captured (usually between 3 and 6). This allows for a precise estimation of the GCI or within 1 to 2 ms of that time frame, which plays a crucial part in estimating the vocal tract coefficients. Only when the precise location of the GCI is known, the VT coefficients can be determined with better accuracy, since at that point (right after the GCI) in the absence of glottal excitation, the sound pressure inside the vocal tract becomes freely decaying oscillation due to VT resonance.

4.1.3 Sequential Covariance Method

Wong et al. (1979) assumed an all-pole model for the vocal tract during vowel production, and used covariance analysis to estimate the vocal tract transfer function during the glottal closure phase which was determined using the resulting normalized linear prediction error energy waveform. Even though this method can provide high resolution estimates of the glottal airflow, that is essentially a representation of the AC-flow and the true offset value or the amplitude quotient, defined as the ratio between the amplitude of the AC-flow of the glottal waveform and the amplitude of the minimum of the flow derivative is essentially missing from the obtained signal (Alku and Vilkman, 1996). This glottal inverse filtering method is using recording of the speech that is

produced in an open space environment, at a certain distance from the microphone. Speech is normally recorded this way and therefore this method provides us with more practical implementation. In this case, in addition to the inverse filtering which removes the effects of the formants, the radiation of the lips also has to be cancelled by integrating the output of the inverse filter. One of the disadvantages is though the *minimum flow* portion of the signal is lost, thus the *AC-flow* is not a true representation. A very important factor that influences the shape and the uniqueness of the glottal flow is the phonation type. Wong's method was used in this work because of its substantially lower computational requirements, in comparison with other techniques, the adequacy of the obtained results for the glottal symmetry information required for the emotion classification task and because it was a good match when compared to key features of corresponding laryngograph waveforms. In a little more details, this method shows that covariance analysis presents a least-squares estimate of the all-pole model of the vocal tract *V(z)* eq. (4.3). It also includes details about determining the instants of glottal closures and openings by the use of normalized error energy. The analysis filter to be obtained has the form:

$$A(z) = \sum_{i=0}^{M} a_i z^{-1}, \tag{4.6}$$

where, $a_0 = 1$ and M represents twice the number of formants in the speech signal *s(n)*. Once *s(n)* is passed through the analysis filter described above, the error residue signal is obtained at the output, which takes the form:

$$\varepsilon(n) = s(n) + \sum_{i=0}^{M} a_i s(n-i), \tag{4.7}$$

The analysis filter coefficients for *A(z)* are obtained by minimizing the total squared error $\alpha_M(n)$, which is given by:

$$\alpha_M(n) = \sum_{j=n}^{n+N-M-1} \varepsilon^2(j), \tag{4.8}$$

A(z) is obtained through linear prediction covariance method over a window frame applied to the speech signal *s* starting at point *(n-N)* and ending at *(n+N-M-1)*. The window length is *N* samples and total squared error $\alpha_M(n)$ is computed by shifting it sample by sample. An instance of closure is defined where $\alpha_M(n_1) = 0$ or at (n_1-1). An instance of opening is found when the next sample different than zero is found at n_2 or at $(n_1+N-M-1)$. The error of the system must be normalized so it is not biased by external system gain, thus the normalized total squared error is used:

$$\eta(n) = \frac{\alpha_M(n)}{\alpha_0(n)}, \tag{4.9}$$

where, $\alpha_0(n)$ is the energy of the input signal. The advantage of this method is that it is relatively simple and does not require much computational power. In general, it provides for a more accurate detection of the moment of glottal closure than opening. Thus, preemphasis of the signal is needed before computing the LP covariance method, so the openings are better defined. One other issue is that the inverse filter *A(z)* needs to be carefully adjusted so it can only remove the poles from the speech. The estimated *V(z)* model may have poles at zero-frequency or at the folding frequency. The first problem is

addressed by applying a highpass equiripple FIR filter. In this work two cascaded highpass filters were designed. They had a stopband frequency at 0 Hz and 15 Hz, passband at 40 Hz, and stop-band attenuation at 80dB. The signal was subsequently pre-emphasized with a factor of 0.95. In the case of poles occurring at the folding frequency, they are not removed since it is practically more feasible. They have minimal effect on the filtering process although in practice it adds noise to the final outcome. One major disadvantage of this inverse filtering technique is that it is very sensitive to recording quality. In Figure 4.3 one such problem is depicted, where a phase distortion is

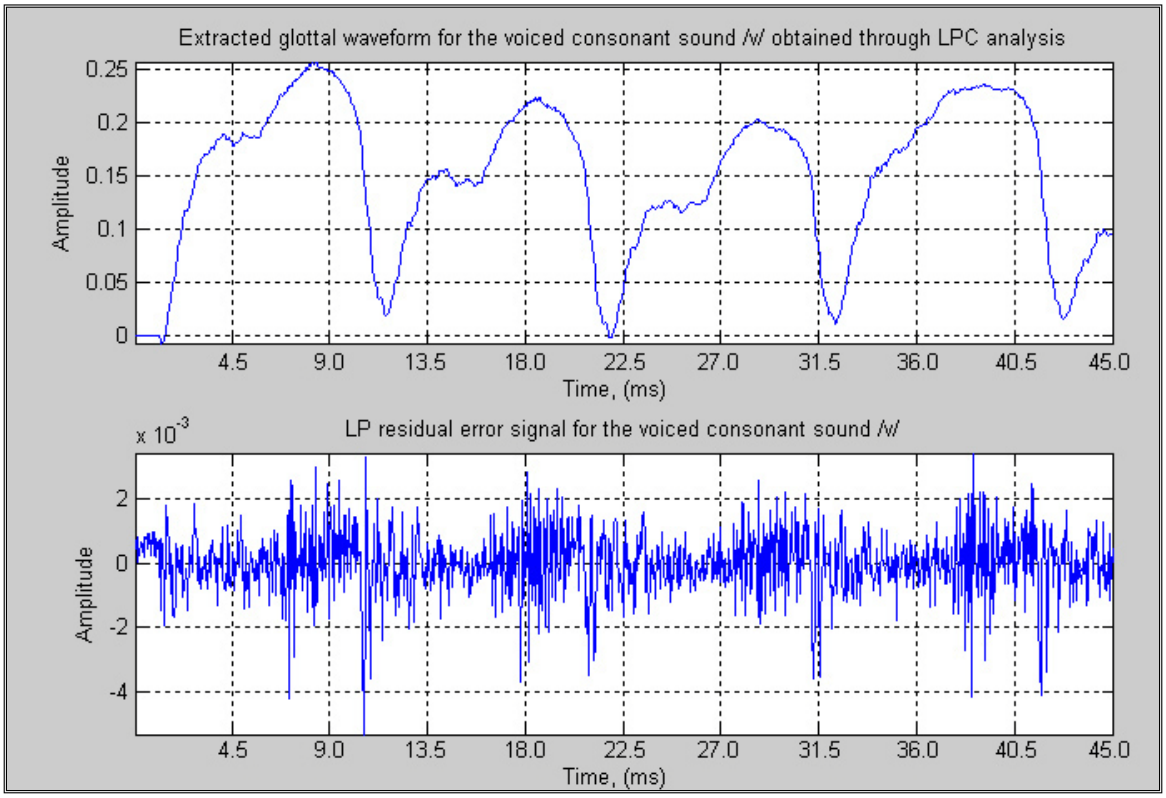


Figure 4.3: Phase distortion on the glottal signal and its subsequent LP error signal.

introduced to the glottal signal due to poor recording conditions. Although we observe the problem above, in the LP residue we are still able to tell the moments of glottal closure corresponding to the steep negative peaks. Another problem can arise when a low

frequency bias is introduced in the recording thus leading to the tilted slope of the glottal waveform. This is rectified by the use of the HPF filters described earlier. These are the reasons why this method was used only in a controlled studio environment, so the ambient noise can be avoided, as well as any low-frequency bias produced by the speaker's breathing. The later is achieved with the assist of a pop-filter. Furthermore this method is sensitive to distortion introduced by the recording devices. All of this rendered the method not suitable for real life applications, thus other methods were adopted.

Error correction procedure was developed for the falsely detected openings and closure instances - *false detection*, as well as for the ones that were not correctly identified - *missed instance*. Considering the fundaments frequency range of an average child speaker, any detection of neighboring GOI or GCI fallen 5 ms or less apart from one another was discarded. Due to window boundary conditions in the analysis, an overlap allowed for rescanning of a particular portion of the signal for both glottal instances. Phase distortion sometimes may obscure the detection of a glottal event. In such cases either the GOI or the GCI were recovered depending on whichever one was missing. If there were two neighboring glottal openings, apparently a glottal closing detection was omitted and vice versa.

4.1.4 Iterative Glottal Estimation Using the LP Residual Signal

More and Clements (2004) proposed a procedure, which is less sensitive on the accurate estimation of the glottal closure interval, and which yields the minimum airflow values as well. The GCI are estimative through an iterative procedure and the results match those provided with the aid of an electroglottograph. As in Wong, an LP

covariance method was used to obtain the residual signal. The most negative peaks in it were considered the times of glottal closure. To find the best estimate of the glottal waveform, an iterative procedure was adopted by using each of the glottal closure peaks as a midpoint. An LP estimate was obtained for each of the negative peaks by shifting around with a window starting point at each closure and length - the LP order *P*, such that the amount of total shift around each closure was *2P*. To pick the smoothest glottal waveform a first order LP autocorrelation method was used on the glottal derivatives obtained from the iterative procedure, such that:

$$\alpha_1 = \frac{-r(1)}{r(0)}, \tag{4.10}$$

where, $\alpha_1 \approx 1$ when the glottal waveform is the smoothest. The numerator essentially represents the autocorrelation at lag 1 and the denominator the autocorrelation at lag 0.

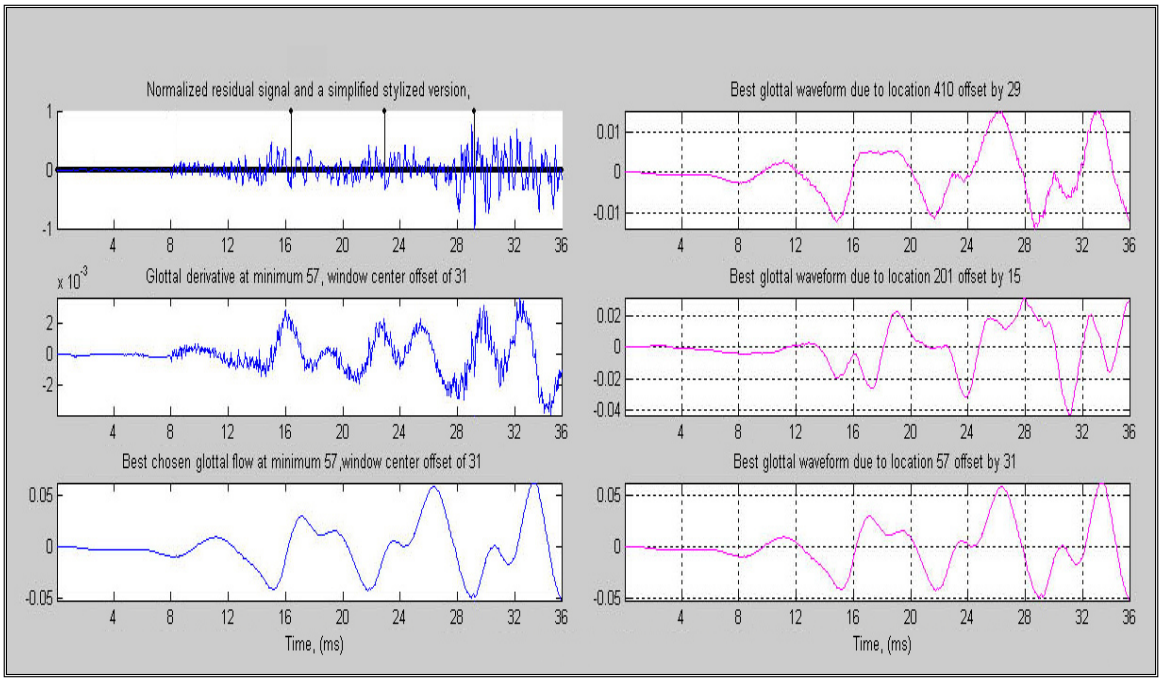


Figure 4.4: Glottal shapes obtained by applying LP autocorrelation on each closure.

This process is depicted in Figure 4.4, where a window of 36 ms length was used for the analysis. As depicted in the figure, three different iterations were performed, one for each glottal cycle.

As expected the iterative procedure added a large computational strain on the method, but providing a smoother glottal shape estimate. The algorithm does not search for the exact glottal closure instance and yet produces a very decent glottal shape. This makes it suitable for more realistic estimates of the glottal signal in a real life environment as compared to Wong's method. In practice however there were voiced regions in which both algorithms failed to represent the exact glottal form. The result of such failures was most likely due to the general assumption that the vocal tract is an

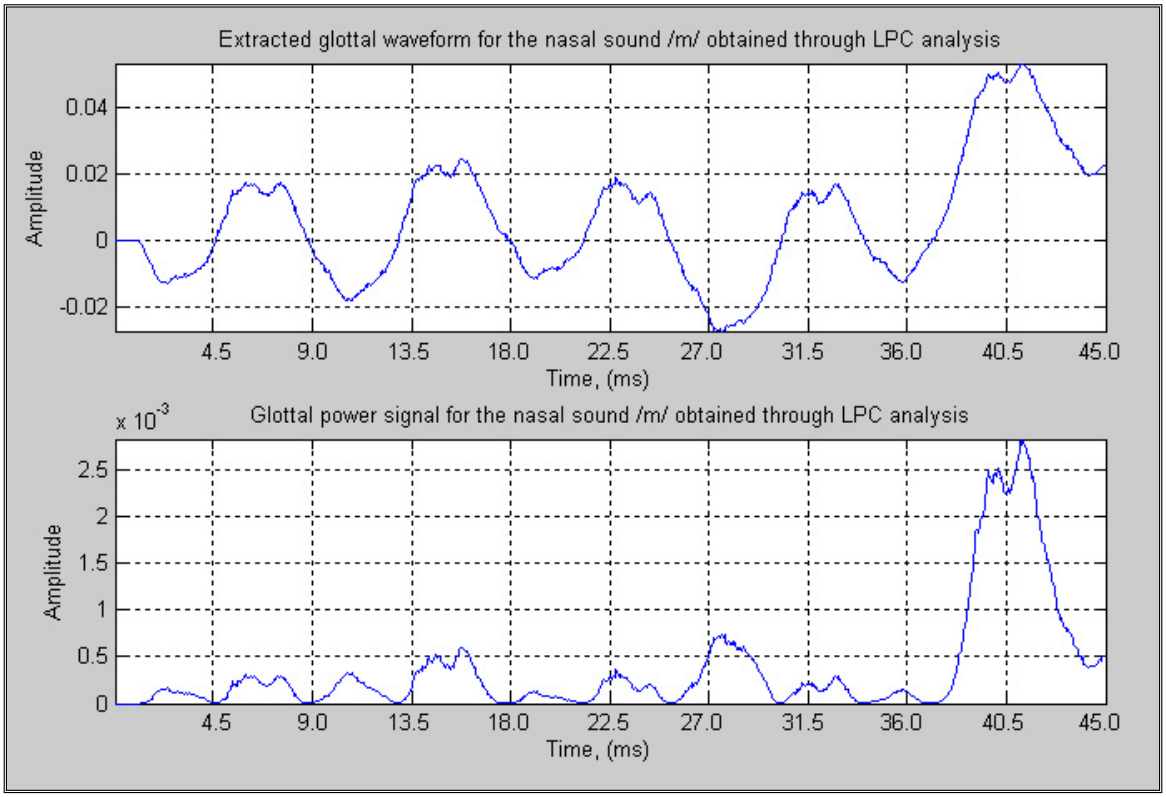


Figure 4.5: LPC analysis of the nasal sound /m/: glottal signal (above), power of the glottal signal (below).

all-pole system, which does not perform as well for voiced consonants as compared to vowels in some speakers. This is so because in nasal consonant production (*m*, *n*) the oral cavity is closed and this introduces zeros in the vocal tract filter response. As a result the obtained vocal tract model does not reflect the true physical model since some closures may not be as apparent in the residual signal. In Figure 4.5 the plots depict the extraction of the nasal consonant sound /m/. As can be seeing from the figure, the closing phase of the vocal folds or the GCI, are very gradual and not as steep as in the vowel /a/ shown in Figure 4.6. There, we can clearly see the abrupt moments of glottal closure, which corresponds to the high peak in the power of the glottal signal below. In more extreme

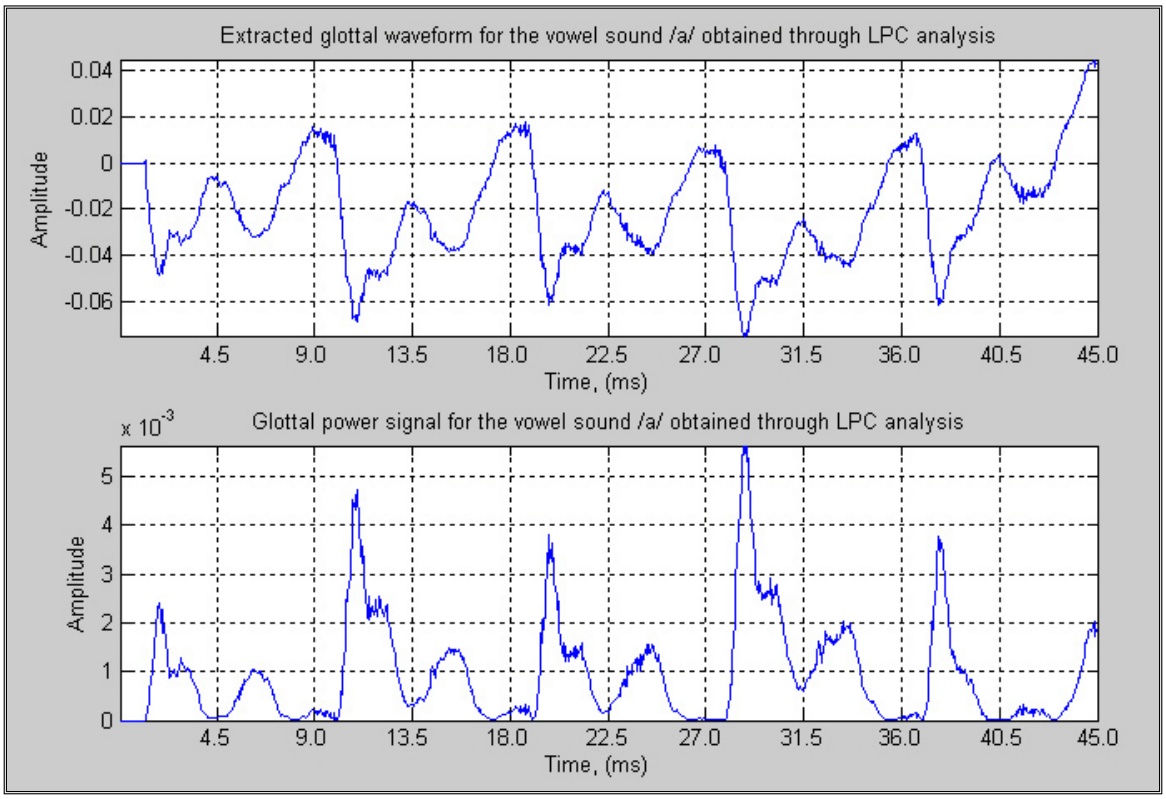


Figure 4.6: LPC analysis of the vowel sound /a/: glottal signal (above), power of the glottal signal (below).

cases, when working with nasal consonants, there may be a problem of determining the exact moment of closure, due to smearing, as extracted using the conventional LPC analysis.

Another problem may arise when dealing with voiced consonants voiced consonants (*b*, *d*, *g*, *v*, *z*). In some speakers, there may be additional excitations due to turbulence at the point of constriction at the lips, which in turn adds colored noise in the glottal signal thus obscuring the closures and leading to poor LPC filtering. This is graphically depicted in Figure 4.7, where the voiced consonant vowel /v/ was analyzed. There is a noticeable smoothing of the closure slopes, which inevitable leads to obscured LP residue (below), thus the exact moments of closure are not easily detectable.

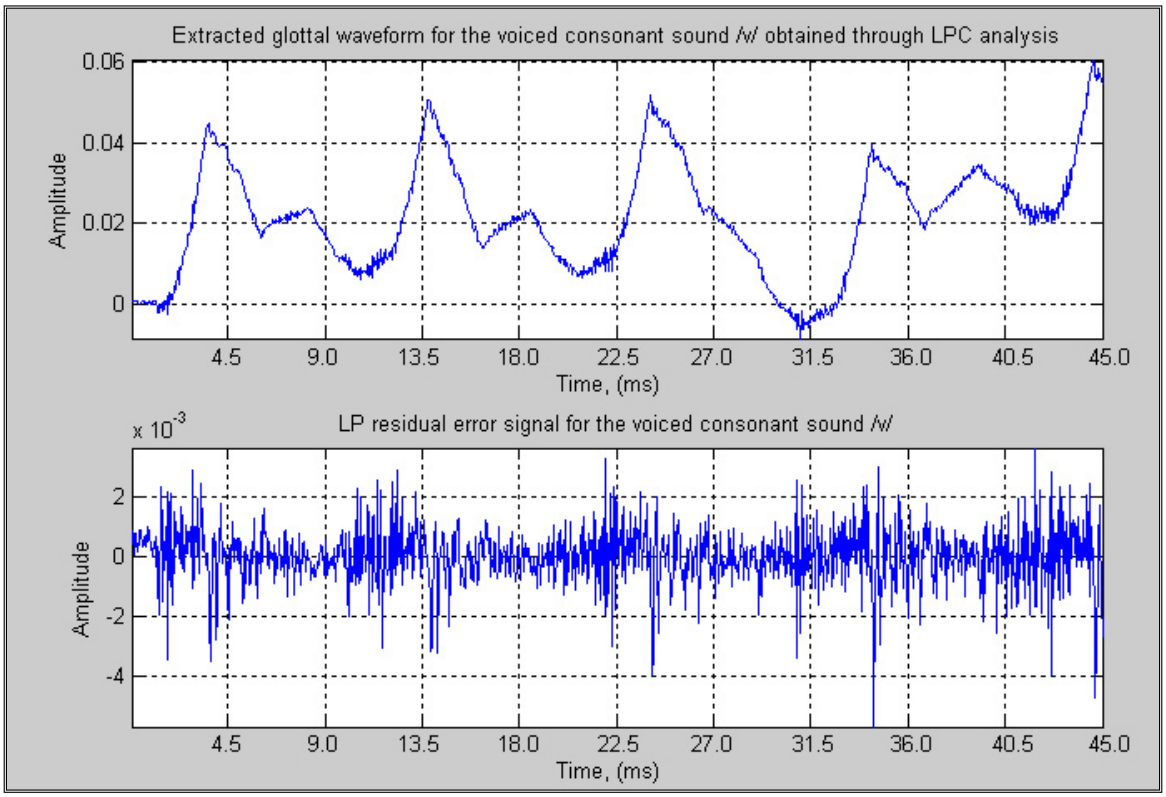


Figure 4.7: LPC analysis of the voiced consonant sound /v/: glottal signal (above), LP residue signal (below).

Although the best glottal signal can be picked using iterations of the residual signal, clearly there is a problem detecting the exact GCI moments in certain instances when using the traditional covariance LPC method. In general, if the prediction order p is high enough, the all-pole system can provide a reasonable estimation equally well for most speech sounds. Another important issue is that this method used autoregressive LPC modeling, thus provided a good spectral estimate of the glottal flow, but did not explicitly deconvolve the quasi-periodic glottal pulse train from the vocal tract transfer function. The same could be said for the method described in Section 4.1.3. Similar error correction mechanism as the one developed for the previous inverse filtering method was adopted for this procedure as well.

As it was shown there are numerous problems that can obstruct the exact representation of the shape of the glottal signal through the inverse filtering process. But even though errors are introduced to the glottal signal, the glottal symmetry is generally immune to variation in the detailed signal representation since it only takes into account the GCI and GOI. This makes the glottal symmetry feature very robust against this type of noise and thus suitable for the emotion recognition problem at hand. In addition since the ripples introduced to the inverse filtered glottal signal are due to low frequency distortion they are not expected to affect the MFCC performance of the glottal signal either.

4.1.5 Autocorrelation Linear Prediction Method

The all-pole linear prediction system should be such that can be uniquely identified when only previous output samples from the system are available. The

motivation behind using the all-pole assumption is because most of the time there is no access to the incoming sequence of samples and therefore this model gives system of equations that could be solved very efficiently. A general all-pole system can be expressed as:

$$\tilde{y}(n) = \sum_{m=1}^{p} \alpha_m y(n-m), \tag{4.11}$$

where, p is the prediction order, α_m is the set of prediction coefficients, y is the true output of the system, and $\tilde{y}$ is the output of the linear predictor. The difference between the actual and predicted signals is represented by the error:

$$e(n) = y(n) - \tilde{y}(n), \tag{4.12}$$

The goal is to minimize the prediction error by better estimating the prediction coefficients. For that, we first examine the performance of the system over *N* number of samples. Then based on the observations, we set the prediction order of the system *p*. The predictor coefficients are calculated next, such that they minimize the energy of the error signal over *N*. This process leads to a system of *p* equations in *p* unknowns and is known as *least-squares* minimization.

Although there are several known linear prediction coding methods, the most commonly used algorithm for extraction of the glottal signal is the covariance method discussed in the previous chapter. As it was documented by Brooks et al., the autocorrelation linear prediction carries certain advantages over the covariance method, as applied to the extraction of glottal signal in noisy conditions. In the autocorrelation

predictor, the waveform signal is considered to be bound within the interval $[\,0, N-1\,]$ and thus can be expressed like:

$$y(n) = y(n+k)\,w(n), \tag{4.13}$$

where, *w(n)* is a finite length Hamming window and $n \in [\,0, N-1\,]$. The signal is windowed with *N* point window length, the result of the liner predictor is equivalent to the short-time autocorrelation function, or:

$$\phi = \sum_{n=0}^{N-1-j+k} y(n)\,y(n-k+j), \tag{4.14}$$

where $j \in [\,1, p\,]$ and $k \in [\,0, p\,]$. The boundaries of the window postulate that the signal is zero outside the *N* sample region. The autocorrelation formulation of the least squares fit of the predictor coefficients from equation (4.14) constructs a system of linear equations that can be represented in matrix form. It can then be solved via standard Gaussian elimination for example. In practice, the autocorrelation solution of equations can be achieved a lot more efficiently. This is because the autocorrelation coefficients in the matrix form of the equations have a very simple symmetric structure, which allows for a recursive solution. One such popular solution is offered by Levinson and Durbin, where each predictor coefficient may be derived from the previous coefficient. In the autocorrelation LP method, the [*p* x *p*] matrix resolution of correlations is a Toeplitz matrix, with a symmetry over the diagonal where all elements across are equal. This gives grounds for implementing faster computational solution of this method.

In order to make the current emotion recognition model more indifferent to the alignment between analysis frames and larynx cycles, as well as to gain better noise robustness, autocorrelation LPC was applied for the analysis of corpus 3.

When comparing the covariance and the autocorrelation linear prediction techniques, there are three main issues to be considered: the number of multiplications for the computation of the correlation matrix and to find the solution of the matrix equation, the amount of storage used, and stability of the system. All of these values are well summarized by Rabiner and Schafer, 1978. For the covariance method the number of multiplications for the correlation matrix is $* p$, and to find the solution for its equation it requires $\frac{(p^3+9p^2+2p)}{6}$ multiplications, p divisions, and p square roots. In comparison, the autocorrelation method needs the same amount of multiplication for the correlation matrix or $N * p$, but the solution to the matrix equation needs much less computational power or precisely p^2 multiplications. The amount of storage needed for the data in the covariance method corresponds to the number of analysis point *N* and for the correlation matrix that number is $\frac{p^2}{2}$. For the autocorrelation method these numbers are: *N* for the data point and p for the autocorrelation matrix, which again is less than the one needed in the covariance method. Finally, the stability of the autocorrelation method is almost always guaranteed when it is computed with sufficient accuracy, which in turn means using high enough prediction order. In addition the stability of the predictor polynomials will normally remain stable when using a pre-emphasis filter. However, the stability of the prediction polynomials in the covariance method cannot be guaranteed. In general, if the number of samples in the analysis window is large enough, both methods

will lead to a similar solution. Considering the characteristics of both the covariance and the autocorrelation linear predictors the later remains in focus for the analysis of corpus 3. Figure 4.8 shows the LP residue of a given speech signal. The moments of glottal closure are easy to see, as they are represented by the peaks of the residual signal.

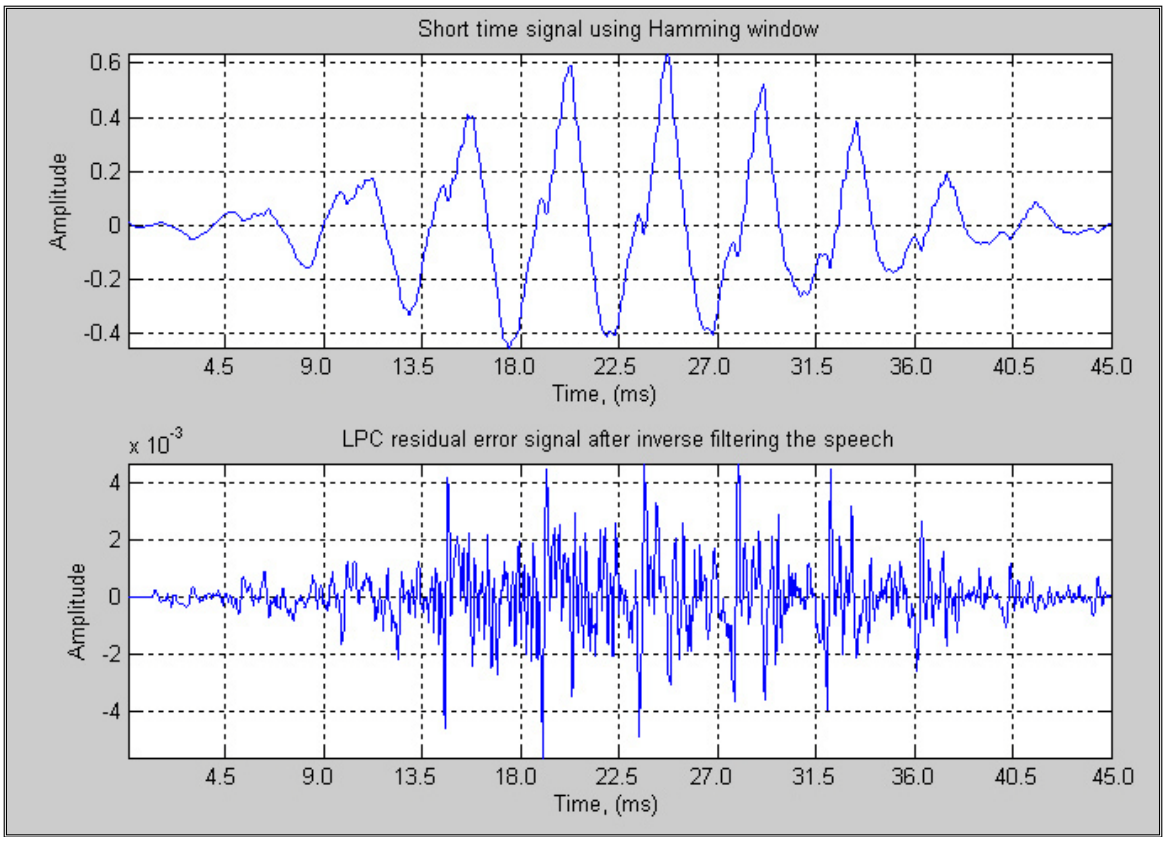


Figure 4.8: Short time speech using Hamming window (above) and its LPC residue signal after inverse filtering (below).

4.1.6 Group Delay

A side effect when using a highpass filter, as described in Section 4.1.3, is its characteristic of introducing a frequency dependent delay into the output signal. In theory, when sending a square wave at the input of a highpass filter, in order to reproduce

it back at the output, all frequencies must arrive at the same time. In practice however the frequency dependant delay introduced by the filter will smear the shape of the square wave at the output as it passes through it. This is because all frequency components of the signal enter the filter synchronized in time but they appear with low frequencies slightly bit more delayed than the high frequencies, which results in time smear or phase shift. One of the best ways to address this problem is to lower the cutoff frequency of the highpass filter as much as possible. In general, the lower the cutoff and stop band frequency of the filter, the lower the group delay in the passband. This is the reason why the highpass filters had stop band frequencies at 0 Hz and 15 Hz to allow for a bit wider transition band. Finding the moments of glottal opening and closing in the inverse filtered glottal signal is dependent on a precise filtering methodology, thus to correctly reproduce the spikes in the LP residue, the problem of group frequency delay must be addressed.

In a separate study on identifying glottal closures, Brooks et al. (2006) have addressed many of the issues discussed herein. As already established in Section 4.1.4, there may be various problems as to correctly localizing the GCI in the residual LP signal. Thus the group delay employment in this work is essential for addressing the issue. In essence, this techniques computes the frequency-averaged group delay applied to the LP residue, while using a sliding window with length of 5 ms. As in all preceding techniques, linear prediction analysis was performed on the closed glottal portion of the glottal waveform (portion B_0-B_1 in Figure 3.3) to obtain the VT coefficients. The autocorrelation LP technique was adopted, to address the issues discussed in the previous sections. The noise of any FM system is primarily generated by the higher frequencies rather than the lower frequencies. For that purpose, FM systems implement a system of

pre-emphasis in which the higher frequencies are increased in amplitude. Thus the original speech signal needs to be pre-emphasized by either using a RC differentiator during the recording process or passing the recorded signal through a first order pre-emphasis filter of the form:

$$H(z) = 1 - 0.95z^{-1}, \tag{4.15}$$

By applying the LP inverse filter on the pre-emphasized speech signal, the residual signal is obtained. As pointed out by Brookes et al., the use of the LPC residual signal requires three assumptions: *a)* the VT is assumed to be an all-pole system, which was discussed in more detail earlier; *b)* the filter should be estimated solely from the speech waveform; and *c)* the LP residual signal will carry the timing instances or identifying the GCI for voiced speech. One of the main points in this method is the addition of a new energy-weighted group delay measure to assist for more precise localization of the impulse, identifying the instance of CGI in the residue. The group delay measurement is applied to the residue signal obtained by the LP, applied to the pre-emphasized signal. The group delay function is then averaged across the frequency spectrum, thus detecting impulses corresponding to the glottal events. This step shows that identifying the GCI is more robust as compared to all previously described inverse filtering techniques.

In general, group delay can be defined as the derivative of the phase, in radians, with respect to frequency. Group delay is caused by filters and may also mean an average of this delay over some frequency band. It can also be described as the time delay through a given filter for a pulse of given sine-wave. In the case where the group delay is

non-uniform, meaning it varies with the sine-wave frequency, the time domain response of a sharp input signal change may overshoot and show ringing. At any sine-wave frequency, the group delay of the filter equals the derivative of the filter phase shift in regard to frequency, or:

$$\tau_r = \frac{-d\arg(X_r)}{d\omega}, \tag{4.16}$$

where, X_r is the Fourier Transform (FT) of a given signal $x(r)$, and r is the beginning sample for the FT and $\omega = 2k\pi/N$ is the frequency. As a general rule it can be said that a completely uniform group delay is corresponding to a perfectly linear phase response. One of the great features about addressing the issue of group delay in our case is that it will provide better localization of an impulse within an analysis frame. This will give more accurate calculation of the GCI and thus will help estimate the Glottal Symmetry needed for the emotion recognition task at hand. Expanding equation (4.16), the group delay of a sampled signal $x_r(n)$ with window size *n=0, ..., N-1*, becomes:

$$\begin{aligned}\tau_r(k) &= -\Im\left(\frac{d\ln(X_r)}{d\omega}\right)\\ &= -\Im\left(\frac{1}{X_r}\frac{dX_r}{d\omega}\right)\\ &= \Re\left(\frac{\sum_{n=0}^{N-1} n x_r(n) e^{-\frac{2j\pi nk}{N}}}{X_r(k)}\right),\end{aligned} \tag{4.17}$$

where, the numerator in the last expression is the discrete Fourier transform of the sampled signal $nx_r(n)$, and $\Re$ is the real part. As expected, if noise is introduced to the signal the group delay will vary, which in turn may jeopardize the identification precision of the GCI. In this case the group delay should be averaged for all *k*. In these terms Brooks offered four different terms in which *k* is restricted to only take integer values, thus offering alternative solutions of how to best estimate the delay. One such value is the Average Group Delay, which is given as:

$$d_{AV}(r) = \frac{1}{N}\sum_{k-0}^{N-1}\frac{\widetilde{X}_r(k)}{X_r(k)}, \tag{4.18}$$

where, $\widetilde{X}(k)$ is the Fourier transform of $nx_r(n)$. The major problem in this method is that the denominator can approach zero for some *k*, in which case the residual quotient will dominate the expression. This will have a negative effect on the term as the group delay d_{AV} will approach infinity. Brooks proposed a new group delay measure that addresses this problem. This method is called *Energy-Weighted Group Delay* or d_{EW}. It limits the bounds of the summation by weighting both the numerator and denominator by $|X_r(k)^2|$. This term represents the energy at k^{th} frequency, and is defined as:

$$d_{EW}(r) = \frac{\sum_{k=0}^{N-1}|X_r(k)|^2\ \tau_r(k)}{\sum_{k=0}^{N-1}|X_r(k)|^2}, \tag{4.19}$$

$$= \frac{\sum_{k=0}^{N-1} \widetilde{X}_r(k) * X_r(k)}{N \sum_{k=0}^{N-1} x^2{}_r(k)}, \tag{4.20}$$

After further simplification the term in eq. (4.20) reduces to:

$$d_{EW}(r) = \frac{\sum_{n=0}^{N-1} n x^2{}_r(n)}{\sum_{n=0}^{N-1} x^2{}_r(n)}, \tag{4.21}$$

This new measure is now bounded within the interval $[0, N-1]$, assuming $x_r(n) \neq 0$. One of the strong points of this algorithm is its resistance to noise. It detects almost all zero crossings and the GCI in turn. Experiments showed that the measure d_{EW} is resistant to SNR up to 14dB. This is due to the nature of the weighted measure as shown in eq. (4.21). When a noise is introduced to it, the center of the energy for the calculated window is shifting, thus compromising the detection accuracy of GCI.

Another important parameter is the analysis window size. If a window with too short of a length is used (shorter than the length of one glottal period), then the captured signal is noise and many zero crossings will be detected. If a window is too large, each impulse at GCI contains a smaller portion of the energy in the frame. This in turn will degrade the time resolution for each GCI. This is a known tradeoff between time accuracy and detection accuracy. The group delay analysis window length used was 20ms.

In this work, all three inverse filtering techniques described in this section were adopted. Moor's method was applied on the clean signal, recorded in an anechoic

chamber, in a controlled environment where the glottal signal was simultaneously recorded with a laryngograph as well. The third method was implemented using the Dynamic Programming Projected Phase-Slope Algorithm platform as described by Kounoudes et al. (2002), Brookes et al. (2006), and Naylor et al. (2007).

A highpass filter was used to remove any low-frequency bias. All signals were pre-emphasized before processing. In addition the recorded database used a pop filter to further reduce breading into the microphone. The parameters of the vocal tract were estimated in a two step procedure:

1. Voiced components were located using the envelope of the signal and obtained from the magnitude of the Hilbert transform and the help of zero crossing rate detection. The Hilbert transform corresponds to allpass filtering the signal with a 90° phase shift. The parameters of the vocal tract system function, V(z), were estimated in the regions where the glottis is closed, so no glottal pulse train was present, and only the ringing of the vocal tract's acoustical environment was analyzed.
2. Speech is inverse filtered with the obtained linear prediction model of V(z) to provide the glottal waveform. The linear prediction order was eight. A laryngograph device was used to simultaneously record the glottal waveform and the obtained signal was subsequently used to verify the results.

Time domain glottal parameters were finally estimated, such as open quotient (*OQ*), closing quotient (*CQ*) and speed quotient (*SQ*). In Figure 3.3, the period *T* of one glottal cycle is between B_0 and B_2 or: $T = |B_0B_1| + |B_1A_2| + |A_2B_2|$. Thus the following can be determined: $OQ = (|B_1A_2| + |A_2B_2|) / T$; $CQ = |A_2B_2| / T$; $SQ = |B_1A_2| / |A_2B_2|$.

4.2 ToBI and Classical Prosodic Features

The classical prosodic features were: pitch and energy in time. Automated extraction for ToBI feature domain was implemented and its performance was compared to classical domain features. Tone tier elements alone are used to depict the four distinct emotional states. Their detection is based on pitch recognition using the Simple Inverse Filter Tracking (SIFT) algorithm (Markel, 1972). The classical prosodic method used both pitch and energy in time for the extraction of useful data for emotion recognition. Fine tuning of the ToBI and classical prosodic feature selection was proposed by calculating the Mutual Information (MI) of all attributes from each domain. The feature set was further reduced after reevaluating it with Sequential Forward Selection (SFS). A new and improved feature domain was developed as a combination of the two, which demonstrated more stable performance in clear speech conditions. All tests in this corpus were performed in a text and speaker independent environment.

Phrasing of the language specific utterances is based on two intonation classes: intonation phrases (IP) and intermediary phrases (ip). Any given IP naturally consist of several ip's. The pitch for each IP is extracted at the beginning using the SIFT algorithm proposed by Markel (1972), and then feature collection follows. In Section *3.3* was established what 11 classical prosodic features were chosen. The first six elements were extracted from the pitch in the current IP. The rising and falling of the pitch in each ip was stored throughout the IP and then the rising-falling pitch ratio was calculated. The next five elements were extracted using the energy of the signal in time domain. The average pause length was detected by using cross correlation or the zero crossing rate.

Then the average of all pause regions was taken across the IP. The speaking rate represented the ratio of number and length of all voiced segments Wang (2004).

The first two ToBI tone tier parameters detected were the highest **H*** (*Topline*) and lowest **L*** (*Baseline*) pitch of a given IP. Then the mid-point average of the pitch was obtained. After all **H*** and **L*** were marked for all IPs, a check for sequential occurrences of **H*** & **L*** was next. When such bursts were detected, only one mark was kept, placing it at 60% from the beginning (Jilka et al., 1999). Then the initial mid-pitch boundary tone **%H** was found, which is located at the beginning of the IP, and was detected if only appeared to be 20% higher than the midpoint. It is important to note that some ToBI elements are mutually exclusive. Good examples are the final low boundary tone **L%** and the final high boundary tone **H%**. They are always displayed in combination with preceding pitch information. That is one of the reasons ToBI was chosen as feature extraction method, since it provides further in-depth relationships among different elements across the tone tier. This in a way described the shape of the glottal waveform, which is found to be in accordance with the emotional state. The next step was to determine the location of all low **L-** and the high phrase accents **H-** in each voiced region within the IP. In the current implementation, any pitch fallen below the mid-point average value of the pitch and not equal to the *Baseline* was marked with **L-**, and any pitch higher than the mid-point average of the pitch and 25% lower than **H*** was marked as **H-**. Once all the low and high phrase accents were detected, it was possible to determine the 'phrase accents and boundary tones combinations' as well as the 'bi-tonal pitch accents'. If there was a very low point in the speakers range, and the following and final pitch element in the IP was lower than the preceding one, then **L-L%** was detected.

If there was a low phrase accent followed by high boundary tone, then **L-H%** was detected. If there was a high phrase accent (between 50% and 75%) slightly lowered at the boundary, then **H-L%** was detected. If there was high phrase accent followed by higher boundary tone (upstepping), then **H-H%** was detected. In the case of **L-H%** and **H-H%** if the very last pitch element was **H***, it was then overwritten by either one of them.

Some ToBI elements, after being detected and in combination with preceding or following ones, may convert and represent different elements. An example can be **H*** and **L+H***. In the case where within the ip we had a presence of **H*** and **L**, and the former preceded, a rising peak bi-tonal phrasal accent **L+H*** was detected and the *Topline* at this point was overwritten. If within a given ip there was a presence of **L*** and **H**, and the former followed, a scooped bi-tonal phrasal accent **L*+H** was detected and the *Baseline* at this point was overwritten.

Finally the IP had to be examined for downstepping '**!**', which can only occur in front of an **H***. In the case when **H-** was detected and it appeared to be the highest pitch to the end of the IP sequence, then it became the new highest pitch, marked **!H***. But sometimes depending on the preceding pitch, **!H*** could change too. For example if **L** was detected before the **!H*** within the same ip, then rising peak accent with lower *Topline* (bi-tonal PA) was detected, or **L+!H***. In the case where **L*** was detected before the **!H*** within the same ip, then scooped accent with lower **H** was detected, or shortly **L*+!H**. When **H** between the *Topline* and the **!H*** was detected such that is was higher in pitch than the former, a downstepped bi-tonal phrasal accent was marked, or **H+!H***. In the case when **H-** was found before **!H***, and it was equal or lower than **!H***, the former

downstepped marker remained unchanged. The grouping **!H+H*** is not possible. When a downstep was detected, the mid-point average stayed relevant to the original mid-point and did not change, and therefore the following pitch marks did not change. If at the end of the last ip we had **H-L%**, and was preceded by an **H**, which had a higher pitch than any of the pitch in the **H-L%**, then the former was changed to **!H-L%**. The downstepping procedure continued for each ip until the end of the utterance was reached. In the case when either one of the four combinations **L-L%**, **L-H%**, **H-L%**, or **H-H%** were detected in the last ip, they were not overwritten by **!H***.

4.3 MFCC Extraction

There were four steps used to compute the MFCCs:

1. Apply the Fast Fourier Transform for each window of a given signal;
2. By using the triangularly shaped windows, a mapping from Hertz to mel scale of the spectral powers is created;
3. Compute the log for each one of the newly obtained mel frequencies;
4. Take the DCT of the mel log powers, thus obtaining the MFCCs, which are represented by the amplitudes of the newly obtained spectrum.

To obtain the power "spectrum of the spectrum" of a signal or the power cepstrum, the following sequence flow is frequently used:

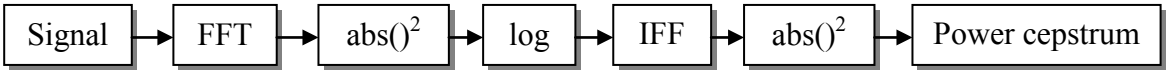


Figure 4.9: Block diagram for obtaining the power cepstrum.

For the MFCC analysis performed for the second round of experiments using the studio recording at the anechoic chamber, the number of features varied from four to ten for the glottal and six to twelve for the speech signal as summarized in Table 4.1. The MFCCs used on the third corpus, were of order 4^{th} and 6^{th}.

Table 4.1: MFCC order used for the glottal and speech signals.

Type of signal	Order of MFCC				
	4	6	8	10	12
Glottal	√	√	√	√	-
Speech	-	√	√	√	√

4.4 System Architecture

Detailed system description of extracting the glottal symmetry from the speech is visually depicted in Figure 4.10. After the speech is loaded in block 0, the signal is highpass filtered in order to get rid of the unnecessary low frequencies below 40 Hz. To maintain the same SNR across the frequency spectrum the signal is pre-emphasized next. The envelope in block 3 is computed by using the real part of the Hilbert transform. Detailed explanation of the Vocal Tract (VT) model is given in Section 2.2. The key component in Figure 4.10 is block 6 where the glottal signal is obtained, because not only the GS can be collected, but the MFCCs of the glottis can be computed as well. After the glottal signal is obtained through inverse filtering, finding the opening and closing points of the glottis and thus the GS were determined. The inverse filtering techniques considered in this work are discussed in details in Section 4.1, and a typical glottal signal and its corresponding time waveform are depicted in Figure 4.2.

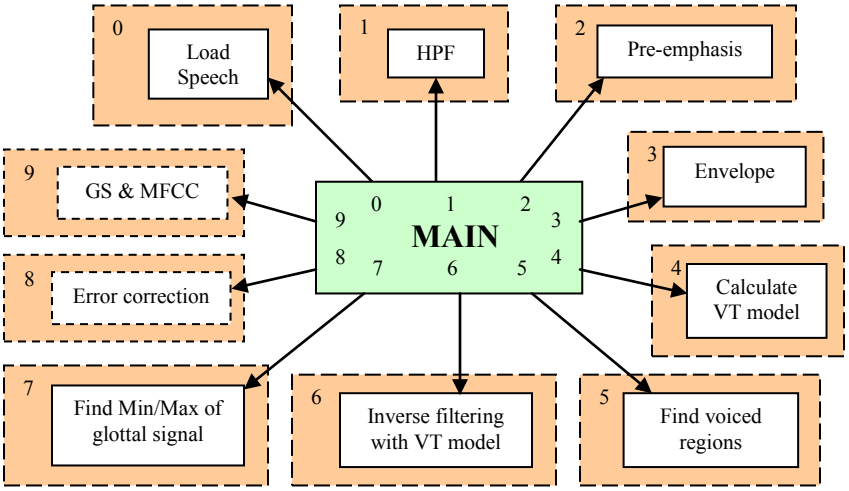


Figure 4.10: Extraction of glottal signal and MFCCs.

Detailed diagram of the ToBI system is shown in Figure 4.11. In block 0, the corpus is loaded in memory for reading. As can be seen, there are two main processes associated with block one. Firstly, the corpus has to be created by manually finding the markers for the beginning of each emotion. Then, all time markers were processed so that the beginning and end of each emotion is recorded sequentially as they appeared in the corpus. The time markers from the speech data collection are then read in block 2 and later stored in memory. Separation of each emotion available in the corpus follows after the reading, so that each emotional utterance is grouped in its own class. This is done in block 3. Feature selection for ToBI is performed in block 4 and training in block 5. When applied to corpus 1, *happy* had the least amount of utterances as shown in Table 3.1. For each of the four emotions in this corpus, the balanced training set was formed with 242 utterances for each emotional class. Testing is performed in block 6 for 20% of the samples, or 61 utterances for each of the four emotions. All results from testing are reported and discussed in Section 7.

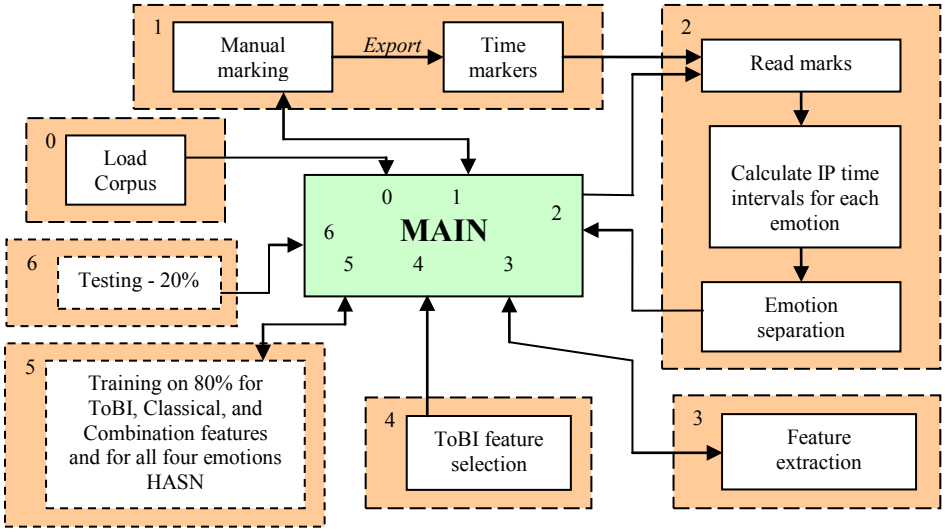


Figure 4.11: ToBI and Classical system design.

After emotion separation is completed, the signal is passed to the SIFT algorithm to extract the pitch and to also error correct it. The extraction of features for corpus 1 involved classical prosodic features and ToBI features only, for corpus 2 they were MFCC of the speech and glottal signals, and the GS, and for corpus 3 all of the above features were included. In the case of corpus 1, the data varied greatly among the different feature domains, therefore normalization was adopted. Finally after all data was detected, error corrected, and normalized, it was divided in three testing categories: ToBI (T), classical (F), and combined (C).

CHAPTER 5

Feature Selection

5.1 Feature Selection and Normalization

As expected, the extracted features had a very dynamic ranges due to their diverse nature. The features extracted by the classical method were within the same typical range, while in ToBI the occurrence of all such features in a given IP was not possible as it has already been shown that some elements are mutually exclusive. This meant that each time automatic extraction of ToBI elements was made, the length of the feature vector varied or was sparse. That was so because ToBI features were extracted sequentially as they appeared in time and in the classical approach they represented generalization of parameters throughout the whole sequence. To clarify, in the first case vector lengths always consisted of the same eleven elements, and in the second case they were unpredictable since it was not known what elements will be detected. All of this made the classification process more difficult. To overcome these limitations, three steps were taken: first a fixed length vectors were created for each IP, then, normalization was performed and finally feature's weight was reevaluated. To create a fixed vector for ToBI all sixteen elements were mapped and then their occurrences counted. It should be noted

that after fixing the problem with the fixed length vector we lost each feature time reference. The numbers collected under each feature showed if an element appeared in the particular IP analyzed. As a result of their mutually exclusive nature, many of them had a zero occurrence. That created the need of reevaluating the newly created feature space in order to determine what features are truly going to contribute for the classification task at hand. Also, as stated in the literature (Wang, 2004) when too many features are used for classification, the overall success rate may diminish, which was confirmed in this work as well. Moreover some of the features may be highly correlated, hence leading to strenuous computational payload. A method for automatic feature selection refinement was necessary, thus Mutual Information was employed. The results were further examined by computing the Sequential Forward Selection.

Normalization of all utterances followed before the classification modeling process. It was done for all 27 features in two steps and was verified experimentally. To avoid divisions by zero, an offset was first added to each sample set. To scale the data for better compatibility, a standard deviation difference threshold was used.

5.2 Information Content and Mutual Information

Constructing the feature vectors is part of the solution, since the weight of each feature needs to be tested for optimization. In real-time applications there is a need for fast processing and thus reduction of computationally expensive feature extraction procedures is needed. Therefore there is a need to determine the mutual content between different features. This is achieved by calculating their mutual information (MI). MI is a procedure evaluating the amount of information one arbitrary variable contains about

another (Cover and Thomas, 1991). In the feature selection case information content (IC) for each feature shows its contribution in the classification process, while reveals their redundancy as well. This process can be depicted as:

$$I(C,x) = \sum_{c} p(c,x) \log \frac{p(c,x)}{p(c)p(x)} \tag{5.1}$$

where: $X:(x_1,x_2,\cdots,x_n)$ is the feature vector, and $C:(c_1,c_2,c_3,c_4)$ is the random variable representing the four emotion class labels (*Happy, Angry, Sad, Neutral*). In short $I(C,x)$ is the reduced amount of the class uncertainty after observing feature *x*, Cover (1991). When $I(C,x)$ is higher the given feature is more effective and vice versa. The IC content of investigated ToBI features in this study is displayed in Table 5.1, and the MI of all classical prosodic features is depicted in Table 5.2. IC was calculated by training the classifier for one feature at a time then testing its accuracy based on the outcome from the testing samples. Here, the goal was to reduce the amount of features based on their IC. Although there is no certain limit below which some features should be discarded, it was aimed for the top eight features from each feature space. Next we needed to estimate the performance of all features when paired up. These calculations may additionally be facilitated by using the greedy method described by (Battiti, 1994). In this scenario instead of estimating $I(C,X)$, which represents the IC of a certain feature vector *X* as it relates to class *C*, the algorithm is aiming to find the MI between individual feature x_i and class *C* depicted as $I(C,x_i)$, as well as $I(x_i,x_j)$, which is the MI across individual features. If the value of $I(x_i,x_j)$ is large, then the investigated pair of features carry very similar information hence one of them could be discarded. As a result

there may be insignificant variation in classification performance, but achieving better computational cost.

Table 5.1: Information content of investigated ToBI feature space.

Feature:	MI:
Lowest pitch (L*)	0.062090
Scooped accent (L*+H)	0.036618
Rising peak accent (L+H*)	0.067108
Downstepped accent (H+!H*)	0.050160
Initial mid-pitch boundary tone (%H)	0.035180
Low phrase accent (L-)	0.056227
High phrase accent (H-)	0.089602
Very low point in the speakers range (L-L%)	0.027052
Low phrase accent followed by high boundary tone (L-H%)	0.067331
High phrase accent slightly lowered at the boundary (H-L%)	0.036640
Extremely high in the speakers range (upstepping) (H-H%)	0.047299
Scooped accent with lower topline (L*+!H)	0.044631
Rising peak accent with lower topline (L+!H*)	0.045288
Downstep - compression of pitch range or lowered topline (!H*)	0.132360
Downstepped high phrase accent slightly lowered at the end (!H-L%)	0.078516
Highest pitch (H*)	0.025279

Table 5.2: Information of investigated classical prosodic feature space.

Feature:		MI:
Pitch	mean	0.011873
	medium	0.010036
	STD	0.013141
	maximum	0.031162
	rising - falling ratio	0.211880
	maximum of falling range	0.034868
Energy	mean	0.038756
	STD	0.064665
	maximum	0.085876
	average pause length	0.213740
	speaking rate	0.149000

Table 5.3: Mutual information content levels between selected ToBI features.

	L*	L*+H	L+H*	H+!H*	%H	L-	H-	L-L%	L-H%	H-L%	H-H%	L*+!H	L+!H*	!H*	!H-L%	H*
L*	0.792	0.356	0.047	0.025	0.003	0.084	0.059	0.009	0.005	0.002	0.004	0.031	0.023	0.066	0.004	0.035
L*+H	0	0.570	0.013	0.005	0.001	0.034	0.032	3E-5	0.004	2E-6	2E-5	0.045	0.010	0.050	0.008	0.020
L+H*	0	0	1.018	0.010	0.008	0.125	0.104	0.003	0.012	0.005	0.013	0.016	0.034	0.082	0.009	0.217
H+!H*	0	0	0	0.598	0.008	0.100	0.126	0.015	0.006	0.007	0.003	0.002	0.093	0.148	0.005	0.040
%H	0	0	0	0	0.511	0.020	0.015	8E-6	2E-5	0.001	0.001	0.003	0.011	0.022	0.001	0.010
L-	0	0	0	0	0	2.136	0.741	0.022	0.026	0.022	0.026	0.010	0.149	0.271	0.029	0.204
H-	0	0	0	0	0	0	2.010	0.021	0.019	0.025	0.017	0.010	0.140	0.278	0.022	0.204
L-L%	0	0	0	0	0	0	0	0.458	0.054	0.063	0.023	0.016	0.013	0.013	0.010	0.005
L-H%	0	0	0	0	0	0	0	0	0.556	0.095	0.034	0.012	0.013	0.030	0.015	0.019
H-L%	0	0	0	0	0	0	0	0	0	0.593	0.040	8.0E-5	0.006	0.049	0.018	0.015
H-H%	0	0	0	0	0	0	0	0	0	0	0.351	0.005	0.008	0.013	0.006	0.021
L*+!H	0	0	0	0	0	0	0	0	0	0	0	0.177	0.027	0.024	0.002	0.006
L+!H*	0	0	0	0	0	0	0	0	0	0	0	0	0.944	0.204	0.005	0.046
!H*	0	0	0	0	0	0	0	0	0	0	0	0	0	1.413	0.087	0.154
!H-L%	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.207	0.008
H*	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.607

In Tables 5.3 and 5.4 respectively, the MI across ToBI and Classical prosodic features are portrayed. Observing the ToBI data sets, few pairs that carry redundant information can be named, for example: L* and L*+H. Referring back to Table 5.1 it can be seeing that L* has a higher value than L*+H, therefore the later were omitted. The decision was taken based on both tables. The most effective features were evaluated and thus the final feature vector was reduced from 16 to 8 ToBI features, namely: L*, L+H*, H+!H*, L-, H-, L-H%, !H* and !H-L%. The pruning process could go down even further, so the newly selected feature set had to be validated by extensive testing with the Sequential Forward Selection algorithm.

A refined selection of classical prosodic features was also performed using the same mechanism of pruning. Based on the observations in Table 5.4, the new classical vector was constructed, containing: *mean* and *maximum of pitch*, *rising-falling pitch ratio*, *maximum of falling pitch range*, *STD* and *maximum of energy*, *average pause length*,

speaking rate. It is obvious from Table 5.2 that the chosen elements carry meaningful information, but based on Table 5.4 fewer features with unique information might have been chosen. For example based on the MI between *STD of pitch* to *maximum of falling pitch range* one should eliminate the first since it carries less meaningful information. The same conclusion can be drawn when the *STD of pitch* is compared to the *speaking rate*. In this case from three features the system decreased down to two. But since three of them carry higher MI, it might be better to use *STD* of pitch instead of the other two, in which case it will prune down from three to one feature. It was therefore needed to verify the validity of this claim by using the SFS method as well.

Table 5.4: Mutual information between selected classical prosodic features.

	mean pitch	medium pitch	STD pitch	max pitch	rise/fall pitch	max fall range	mean energy	STD energy	max energy	avg ps length	speaking rate
mean pitch	2.341	1.162	0.304	0.210	0.098	0.285	0.090	0.128	0.150	0.289	0.272
medium pitch	0	2.614	0.428	0.349	0.172	0.271	0.096	0.161	0.154	0.344	0.313
STD pitch	0	0	2.602	0.271	0.129	0.319	0.152	0.133	0.154	0.364	0.377
max pitch	0	0	0	1.742	0.098	0.182	0.116	0.136	0.129	0.219	0.216
rise - fall pitch	0	0	0	0	1.182	0.194	0.054	0.059	0.065	0.112	0.135
max fall range	0	0	0	0	0	1.971	0.113	0.107	0.122	0.380	0.272
mean energy	0	0	0	0	0	0	1.264	0.255	0.105	0.128	0.159
STD energy	0	0	0	0	0	0	0	1.323	0.364	0.122	0.132
max energy	0	0	0	0	0	0	0	0	1.570	0.171	0.142
mean pause length	0	0	0	0	0	0	0	0	0	2.359	0.264
speaking rate	0	0	0	0	0	0	0	0	0	0	1.221

5.3 Sequential Forward Selection

When it comes to determine the optimal number of features, there are several widely accepted methods. Exhaustive search also known as ‘Brute Force” method is a technique which considers all possible cases sequentially, when there is no better known

technique to obtain efficient solution. Another well known heuristic search method is the Beam Search, where breadth-first search is used to construct the tree-structure. This method only keeps the best partial solutions closest to the goal for the next step. The Genetic Algorithm (GA) is also well known technique of global search heuristics, trying to find an optimum solution. It uses chromosomes to form a population of samples that evolves with generations based on a predetermined fitness function using crossover, mutation, and selection. A chromosome with the highest fitness in a given population is chosen to be the solution. More detailed description to GA is provided in (Srinvas and Patnaik, 1994).

In addition to the methods already mentioned, there are more and various search techniques such as Greedy Search, PTA, and Sequential Backward Search (SBS), but their detailed description is out of the scope of this research. Rather our goal was to find tangible evidence leading to the choice of the right search mechanism. In (Hongwei et al., 2003), the performance of the GA was tested against PTA, GPTA, Sequential Floating Forward Search (SFFS), and Sequential Floating Backward Search (SBFS). It was concluded that they all perform reasonably well, but none of them consistently outperformed the other. SFFS is in general the more refined version of SFS, but comes at a higher computational cost for a marginal improvement. It can be therefore chosen to use the SFS algorithm for selecting the best set of ToBI features. SFS had previously been used in emotion recognition problem where the goal was to select more informative features to improve class-separability (Altun et al., 2007). As stated in there and cited in (Reunanen, 2003) more computationally intensive search methods like the SFFS does not

necessarily surpass their simpler counterparts such as the SFS, despite of the widespread believe of the contrary.

In SFS at each iteration exactly one feature from the full set *A* is given to the set *B*, where $A = \{a_i\}$ and *i* = 1,2, … *k*. In the case of this study *k* = 16. Initially the set *B* is empty and the goal is to determine the best combination from set *A* to construct set *B* which will maximize the final outcome. In these terms, $A \geq B$, because $A \supseteq B$. Every newly introduced feature a^+ must satisfy: $a^+ = \arg\max\left[J\left(B + a^+\right)\right]$, where $a \notin B$ and *J* is the objective function in Altun (2007). This process continues until no further improvement of the final outcome is observed. After careful consideration of the results obtained from MI and after using the SFS, the final eight ToBI features were selected: L*, L+H*, H+!H*, L-, H-, L+!H*, !H*, and H*. The same pruning process was repeated to the classical feature set, thus the final selection contained the seven features: *mean*, *medium*, *maximum*, and *STD of pitch*, *rising-falling pitch ratio*, *maximum of energy*, and *average pause length*. Since this process can be computationally expensive, it is not recommended to be used for large feature sets.

CHAPTER 6

Emotion Modeling

6.1 Classification Based On GMM

In GMM, assuming a D-dimensional sample space and a feature vector X extracted from a speech segment under a certain emotional state E, then its mixture density is:

$$p^E(X \mid \lambda) = \sum_{i=1}^{M} \omega_i^E \, p_i^E(X) \quad , \tag{6.1}$$

where *i* is the number of mixture components, i.e. the number of Gaussians, and ω_i^E is the weight of component i. In other words, it is a weighted linear combination of M Gaussian densities, $p_i^E(X)$, each characterized by a mean μ_i^E, and a covariance matrix Σ_i^E. The mixture weights ω_i^E sum to one: $\sum_i \omega_i^E = 1$, and the model description is $\lambda : \{\mu_i^E, \Sigma_i^E, \omega_i^E, \quad i = 1,2,\cdots,M\}$.

$p_i^E(X)$ has the general form:

$$p_i^E(X) = \frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_i^E\right|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}\left(X-\mu_i^E\right)\Sigma_i^{E^{-1}}\left(X-\mu_i^E\right)\right\} \quad (6.2)$$

,

As previously indicated, the form of the covariance matrix Σ_i^E in this study is diagonal and for each emotion state there is a corresponding GMM. The classification decision of underlying emotion state ES for observed vector *X*, will be made by selecting the class that maximizes the posterior probability (maximum a posteriori or MAP) as:

$$ES = \arg\max_E \{\frac{P^E(X)}{\sum_{i=1}^{4} P^{E_i}(X)P(E_i)}\} \quad (6.3)$$

,

There have been many efforts to use GMM in emotion recognition. In most cases GMM was used as the main classifier (Jiang and Cai, 2004; Chuang and Wu, 2005) and in other cases it was evaluated as an alternative emotion classification method to be compared with other classifiers (Wang and Guan, 2004; Hung et al., 2004). However, there are some fundamental issues of great interest in investigating the deployment of GMM for classification applications, including emotion recognition tasks. Determination of the number of mixture components and appropriate normalization of features to handle the overflow in computing the covariance matrix, are two of them. In this work, in addition to the recognition rate, the effect of using different number of components and the normalization of the feature were investigated.

It is always of great interest, both practically and theoretically to establish how many GMM components are sufficient (Duda and Hart, 1973). It is a key factor to the computation load of the system and one of the weakest assumptions in GMM. There have

been many proposed theoretically elegant methods to determine the optimal number of the components in GMM, when most of them need exhausted searching. This study investigated the issue, and reported the observations in a practical way straightforwardly. The method was evaluated on the same system configurations except for using various numbers of the components. Following is the summary of recognition rates for four and six emotions by using different number of mixture components.

From the experiments the following observations were drawn: 1) Simply increasing the number of components, *M*, used in GMM doesn't guarantee the improvement of the performance; 2) for different emotions, they have different optimum *M*, regarding the three numbers chosen for interest; 3) using 16 Gaussians dynamically degraded the system performance, so roughly the range from 15 to 18 should be avoided when choosing *M* for the task.

Another significant observation comes with the ToBI parameters when used for GMM training. Because of the large range the parameters fall in, the singularity of the covariance matrices and their inverse matrices have been observed frequently. To make GMM training more reliable, the parameters were normalized so that the singularity warning in computing of GMM covariance matrices and inverse matrices has been reduced significantly. Furthermore, if singularity has still been observed after feature normalization, the singular elements in covariance matrix was overflowed to a global fixed small value, or another localized value which could be assigned as 0.5% percent of biggest value observed in that covariance matrix (Zang and Scordilis, 2004). This method has proven to be very efficient.

6.2 Optimum-Path Forest (OPF)

In this section the theory related to the OPF classifier is described. The training set is thought of as a complete graph, whose nodes are the samples and the arcs link all pairs of nodes. The arcs are weighted by the distance between the feature vectors of their corresponding nodes. Any sequence of distinct samples forms a path connecting the terminal nodes and a connectivity function assigns a cost to that path (e.g., the maximum arc-weight along it). The idea is to identify prototypes in each class such that every sample is assigned to the class of its most strongly connected prototype. That is, the one which offers a minimum-cost path to it, considering all possible paths from the prototypes. By estimating prototypes as the closest samples from distinct classes, the OPF can handle all three cases with the maximum arc-weight function. In the case of overlapping between classes, these prototypes will be class defenders in the overlapped regions of the feature space.

The OPF algorithm classifies the samples of the feature space by taking into account the connectivity between them, not only the simplest distance of their corresponding feature vectors, such as the k-NN classifier. Another question concerns with the optimality criteria, because OPF's classification rule is given by an optimal search of the whole feature space, avoiding some misclassifications due to the using of local decision functions. The OPF is a fast, simple, multi-class, parameter independent, does not make any assumption about the shape of the classes (such that ANN-MLP and SVM), and can handle some degree of separability between classes.

Let Z_1 and Z_2 be training and test sets with $|Z_1|$ and $|Z_2|$ samples of a given dataset. The samples can be points, images, voxels (3D pixels), and contours. Here, a

feature vector extracted from speech signals was used as sample. Let $\lambda(s)$ be the function that assigns the correct label i, $i = 1,2,\ldots,c$, of class i to any sample $s \in Z_1 \cup Z_2$, $S \subset Z_1$ be a set of prototypes from all classes, and v be an algorithm which extracts n features from any sample $s \in Z_1 \cup Z_2$ and returns a vector $\vec{v}(s)$. The distance $d(s,t)$ between two samples, s and t, is the one between their feature vectors $\vec{v}(s)$ and $\vec{v}(t)$. One can use any distance function suitable for the extracted features, but the most common is the Euclidean norm $\|\vec{v}(t) - \vec{v}(s)\|$. A pair (v, d) then describes how the samples of a dataset are distributed in the feature space. The current problem consists of projecting a classifier which can predict the correct label $\lambda(s)$ of any sample $s \in Z_2$. A classifier which creates a discrete optimal partition of the feature space is described, such that any sample $s \in Z_2$ can be classified according to this partition. This partition is an optimum-path forest (OPF) computed on Z_1 by the image foresting transform (IFT) algorithm (Falcão et al., 2004).

Let (Z_1, A) be a complete graph whose nodes are the training samples and any pair of samples defines an arc in $A = Z_1 \times Z_1$ (Figure 6.1a). The arcs do not need to be stored and so the graph does not need to be explicitly represented. A path is a sequence of distinct samples $\pi_t = \langle s_1, s_2, \ldots, t \rangle$ with terminus at a sample t. A path is said trivial if $\pi_t = \langle t \rangle$. To each path π_t, a cost $f(\pi_t)$ is assigned, given by a connectivity function f. A path π_t is said optimum if $f(\pi_t) \leq f(\tau_t)$ for any other path τ_t. It is also denoted by $\pi_s \cdot \langle s,t \rangle$ the concatenation of a path π_s and an arc (s,t).

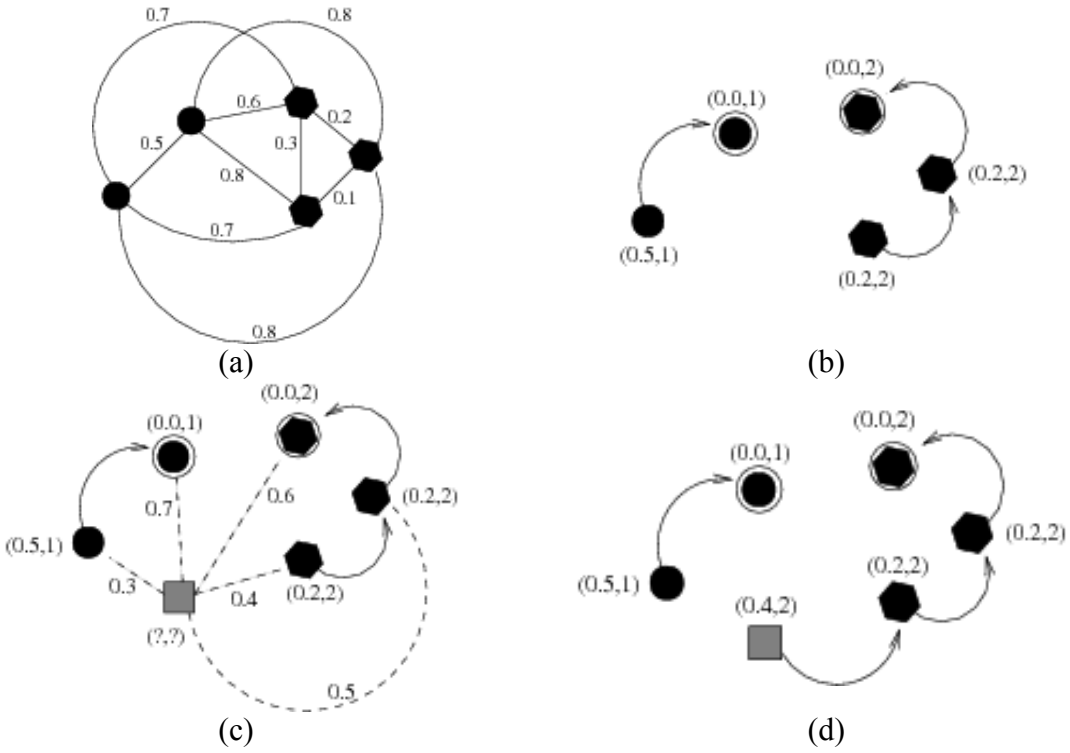


Figure 6.1: (a) Complete weighted graph; (b) Resulting OPF; (c) Test sample and its connections; (d) The optimum path from the most strongly connected prototype.

In Figure 6.1 (b) the resulting optimum path forest for $f_{\max}$ and two given prototypes (circled nodes) are depicted. The entries (x, y) over the nodes are, respectively, the cost and the label of the samples. The directed arcs indicate the predecessor nodes in the optimum path. In Figure 6.1 (c) a test sample (gray square) and its connections (dashed lines) with the training nodes are shown. In Figure 6.1 (d) the optimum path from the most strongly connected prototype is represented. Its label 2 and classification cost 0.4 are assigned to the test sample. The test sample is classified in the class hexagon, although its nearest training sample is from the class circle.

The OPF algorithm may be used with any smooth connectivity function which can group samples with similar properties (Falcão et al., 2004). A function f is smooth

in (Z_1, A) when for any sample $t \in Z_1$, there exists an optimum path π_t which either is trivial or has the form $\pi_s \cdot \langle s,t \rangle$, where

1. $f(\pi_s) \leq f(\pi_t)$,
2. π_s is optimum,
3. for any optimum path τ_s, $f(\tau_s \cdot \langle s,t \rangle) = f(\pi_t)$.

The connectivity function $f_{\max}$ is described as:

$$f_{\max}(\langle s \rangle) = \begin{cases} 0 & if\ s \in S, \\ +\infty & otherwise \end{cases}$$

$$f_{\max}(\pi_s \cdot \langle s,t \rangle) = \max\{f_{\max}(\pi_s), d(s,t)\}, \tag{6.4}$$

such that $f_{\max}(\pi_s \cdot \langle s,t \rangle)$ computes the maximum distance between adjacent samples along the path $\pi_s \cdot \langle s,t \rangle$. The OPF algorithm minimizes $f_{\max}$, by storing the minimum costs in a map C,

$$C(t) = \min_{\forall \pi_t \in (Z_1, A)} \{f_{\max}(\pi_t)\}, \tag{6.5}$$

and assigning one optimum path $P^*(t)$ from S to every sample $t \in Z_1$. Its result is an optimum-path forest P (a function with no cycles which assigns to each $t \in Z_1 \setminus S$ its predecessor $P(t)$ in $P^*(t)$ or a marker *nil* when $t \in S$, as shown in Figure 6.1b). The root $R(t) \in S$ of $P^*(t)$ can be obtained from $P(t)$ by following the predecessors backwards along the path, but its label is propagated during the algorithm by setting $L(t) \leftarrow \lambda(R(t))$.

6.2.1 Training

It can be said that S^* is an optimum set of prototypes when the OPF Algorithm minimizes the classification errors in Z_1. S^* can be found by exploiting the theoretical relation between minimum-spanning tree (MST) (Cormen et al., 1990) and optimum-path tree for $f_{\max}$ (Allène et al., 2007; Miranda et al., 2008). The training essentially consists of finding S^* and an OPF classifier rooted at S^*.

By computing an MST in the complete graph (Z_1, A), a connected acyclic graph is obtained whose nodes are all samples of Z_1 and the arcs are undirected and weighted by the distances d between adjacent samples. The spanning tree is optimum in the sense that the sum of its arc weights is minimum as compared to any other spanning tree in the complete graph. In the MST, every pair of samples is connected by a single path which is optimum according to $f_{\max}$. That is, the minimum-spanning tree contains one optimum-path tree for any selected root node.

The optimum prototypes are the closest elements of the MST with different labels in Z_1. By removing the arcs between different classes, their adjacent samples become prototypes in S^* and the OPF Algorithm can compute an optimum-path forest in Z_1 (Figure 11b). Note that, a given class may be represented by multiple prototypes (i.e., optimum-path trees) and there must exist at least one prototype per class.

It is not difficult to see that the optimum paths between classes tend to pass through the same removed arcs of the minimum-spanning tree. The choice of prototypes as described above aims to block these passages, reducing the chances of samples in any given class be reached by optimum paths from prototypes of other classes.

6.2.2 Classification

For any sample $t \in Z_2$, we consider all arcs connecting t with samples $s \in Z_1$, as though t were part of the training graph (Figure 6.1c). Considering all possible paths from S^* to t, we find the optimum path $P^*(t)$ from S^* and label t with the class $\lambda(R(t))$ of its most strongly connected prototype $R(t) \in S^*$ (Figure 6.1b). This path can be identified incrementally, by evaluating the optimum cost $C(t)$ as

$$C(t) = \min\{\max\{C(s), d(s,t)\}\}, \forall s \in Z_1 \tag{6.6}$$

Let the node $s^* \in Z_1$ be the one that satisfies Equation 6.6 (i.e., the predecessor $P(t)$ in the optimum path $P^*(t)$). Given that $L(s^*) = \lambda(R(t))$, the classification simply assigns $L(s^*)$ as the class of t (Figure 6.1d). An error occurs when $L(s^*) \neq \lambda(t)$.

For the GMM, BC, OPF and SVM classifiers a 10-fold cross validation was performed based on a 80/20 training/testing, randomly generated set partition with a balance representation with regards to speakers, speaker gender and emotion. For k-NN and C4.5, a 5-fold cross validation method was used as shown in Table 6.1. All ten speakers were represented in all training and test sets.

The LibSVM package was used to implement the SVM (Chang, 2001) with radial basis function (RBF) kernel, parameter optimization and the one-versus-one strategy for the multi-class problem. The LibOPF package was used for the OPF (Papa et al., 2008) and for k-NN and C4.5 classifiers we used *Weka* (Witten and Frank, 2005). All classifiers were trained and tested for each of the four emotions, using speech from all 10 subjects.

Table 6.1: Data training/testing arrangement for classification.

Classifier	Data arrangement
BC	80/20 split, 10-fold average
C4.5	5-fold cross validation
GMM	80/20 split, 10-fold average
k-NN	5-fold cross validation
OPF	80/20 split, 10-fold average
SVM	80/20 split, 10-fold average

CHAPTER 7

Emotion Classification Results

7.1 Corpus 1: Four Emotions in "Waiting for Godot"

During the training and testing phases experiments were completed with 14 feature combinations using the four emotional states. The results are shown in Figures 7.1 through 7.3. For simplicity, the following abbreviation scheme was adopted: **A** – for *Angry*, **H** – for *Happy*, **S** – for *Sad*, **N** – for *Neutral*. Depending on the type of features used for training and testing, we adopted: **(T)** for ToBI, **(F)** for classical prosodic features, **(C)** for combined features (**T+F**). All results are based on 61 test utterances and are obtained based on a 10-fold average performance on a balanced corpus.

Figures 7.1 through 7.3 depict the GMM results, where the x-axis has four bar clusters, one for each emotional class, where the classification rate for each is displayed. In Figure 7.1 can be seeing that using the classical features recognizing S is very effective, followed by A, N and H. It is also evident that N was often confused as H, while H was often confused as A or S. From the figures can be established that a longer vector size provides no obvious advantage over a shorter feature vector.

The same cannot be said for GMM T-16 in Figure 7.2 where **S** had equal probability of being misclassified as **A** and H had almost equal chances of being misclassified as A. This shows that the full ToBI vector had relatively inferior performance, which improved for the reduced ToBI (T-8) vector after pruning. Specifically, recognition performance on emotions A and N was above 65%, while for A and S was between 40% and 50%.

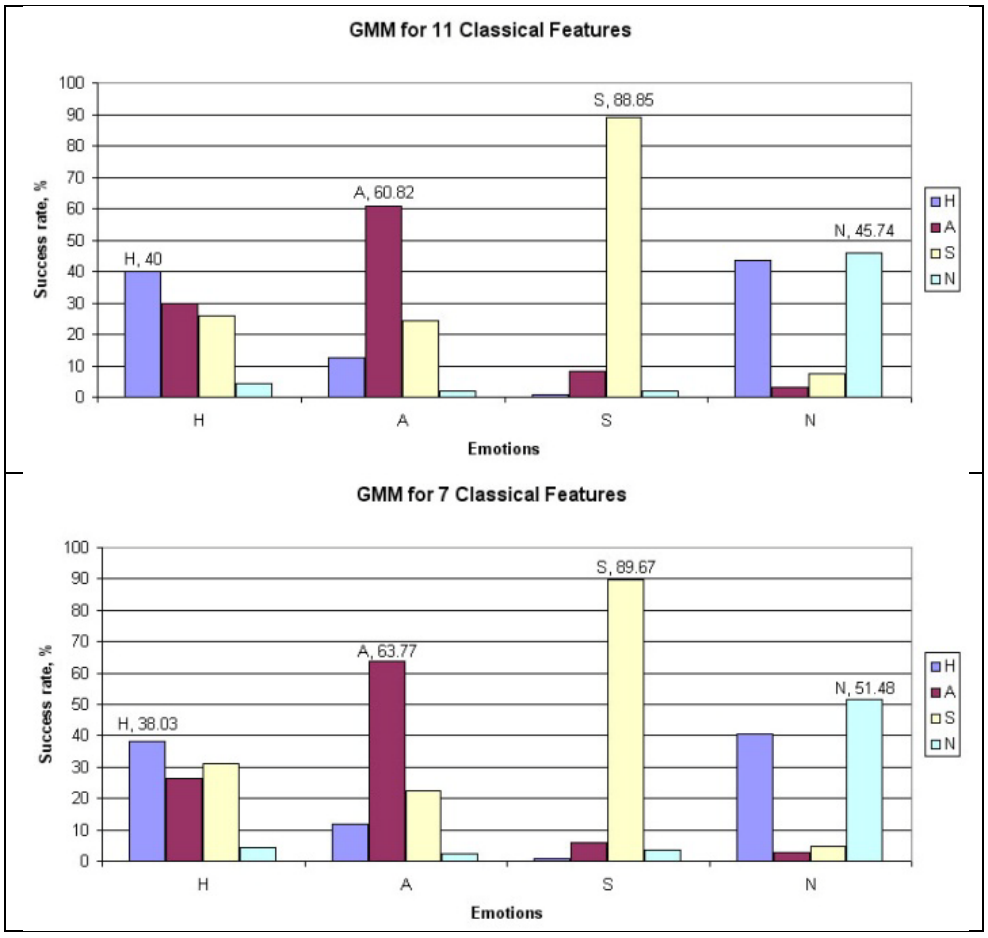


Figure 7.1: Recognition rates for five feature vectors across four emotions using GMM.

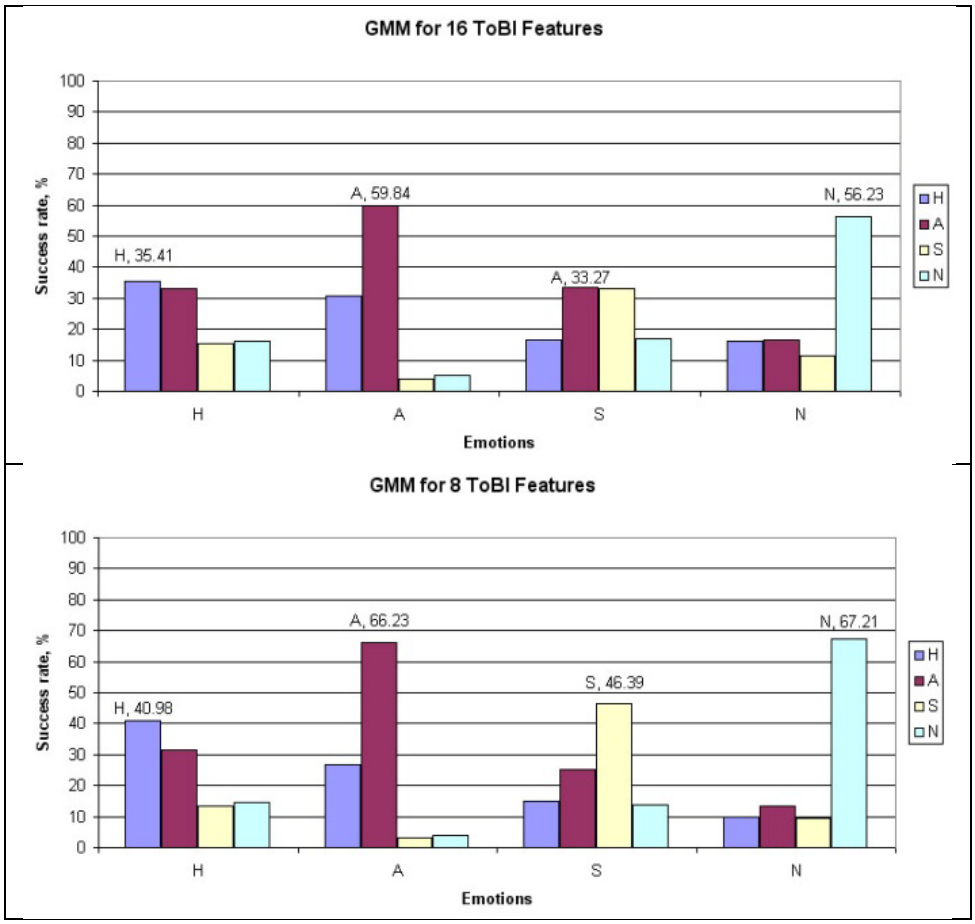


Figure 7.2: Recognition rates for five feature vectors across four emotions using GMM.

So far it is apparent that pruning the classical and ToBI vectors according to mutual information content of features improves classification performance. Even though for all cases recognizing H remained problematic. The situation changed when the two feature domains, classical and ToBI, were combined as shown Figure 7.3. There we see classifier performance for a full feature vector comprising 27 combined features (C-27), and pruned alternatives with vector lengths reduced to 19 and 15 features (C-19 and C-15) by discarding mutual information content. For all cases, the performance was markedly better than classical or ToBI alone, and pruning from C-27 to C-19 improved performance substantially, while from C-19 to C-15 improvement was noticeable but

slight. In particular, recognition of H improved dramatically, from below 50% for classical or ToBI alone to over 60% for the combined features. Recognition of N was also noticeably improved.

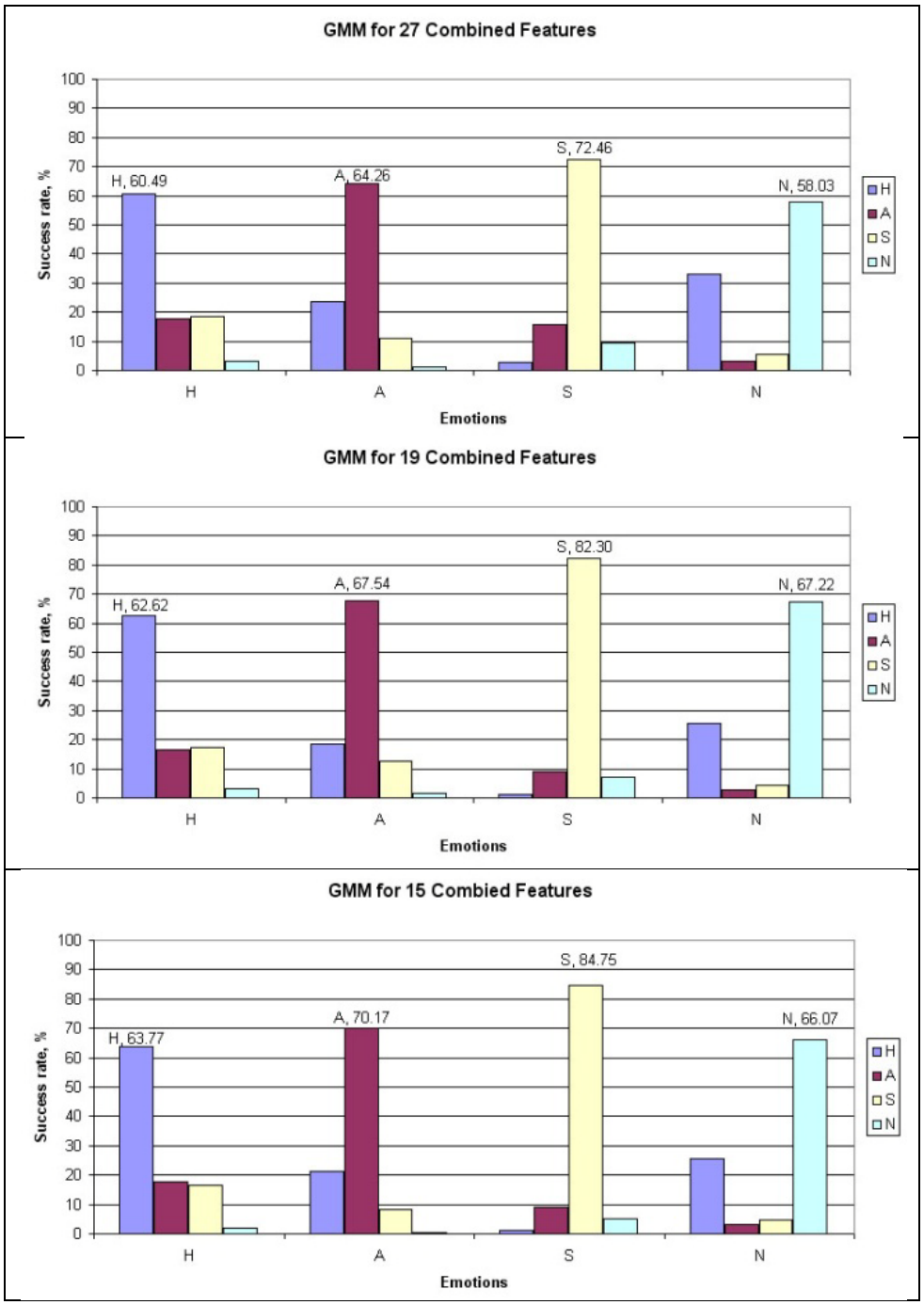


Figure 7.3: Recognition rates for five feature vectors across four emotions using GMM.

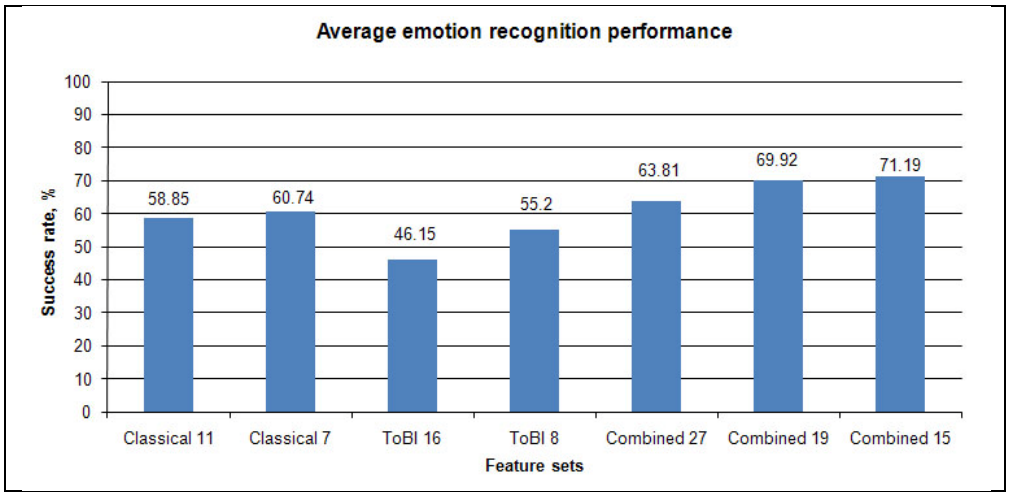


Figure 7.4: Average emotion recognition performance for Corpus 1 (4 emotions).

The average emotion recognition rate is depicted in Figure 7.4. As can be seeing in all three cases: Classical, ToBI, and Combined, the recognition rate improved after removing the mutual information. In general, the classical feature domain outperformed ToBI when used alone. The significance of both domains when performing together is obvious by the last three bars, which show substantial improvement. This is better summarized in Figure 7.5. There, the average emotion recognition performance improvement for corpus 1 is shown over the standard 11 classical features (F-11). Comparing GMM (F-11) and GMM (C-19) there was an overall system improvement of 18.81%. The best enhancement was achieved for GMM (C-15) with 20.97%. In the case of (C-15) we not only gained the best system accuracy but also the lowest computational cost as seen in Tables 7.2 and 7.3. The top average processing time improvements from (C-27) based on 10-fold average were roughly: 2 times for GMM (C-15).

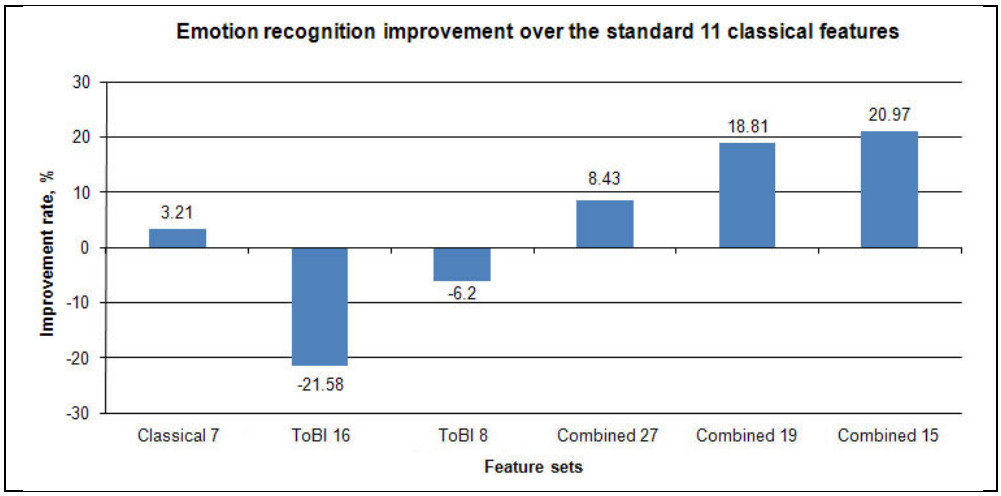


Figure 7.5: Average emotion recognition performance improvement for Corpus 1 (4 emotions) over the standard 11 classical features vector (F-11).

Table 7.1: System accuracy improvement from (F-11) to (C).

Classifier	*(C-27)*	*(C-19)*	*(C-15)*
GMM	8.43%	18.81%	20.97%

Table 7.2: Average processing time in sec.

Classifier	*(C-27)*	*(C-19)*	*(C-15)*
GMM	7.97	4.08	2.58

Table 7.3: Average processing time improvement from (C-27).

Classifier	*(C-19)*	*(C-15)*
GMM	95.34%	208.91%

7.2 Corpus 2: Four Emotions in Short and Long Emotional Phrases with Speech and Laryngograph Signals

The speech signal, its sequential covariance 12^{th} order LP error, along with the laryngograph-recorded and inverse filtered glottal signals is displayed in Figure 7.6. It

can be seen that the recorded and inverse filtered glottal signals have similar characteristics. The reason for that is that recording was completed in a noise-controlled studio environment, which facilitated glottal signal extraction. Examining the two glottal signals, it can be seen that while the plateau of the opening region of the laryngograph waveform is longer and flatter in comparison to the glottal flow waveform calculated via inverse filtering the exact moment of opening is of essence, which is where they both match. The moments of closure match as well, which shows that glottal symmetry can be obtained effectively through inverse filtering in the given studio environment where the samples were collected.

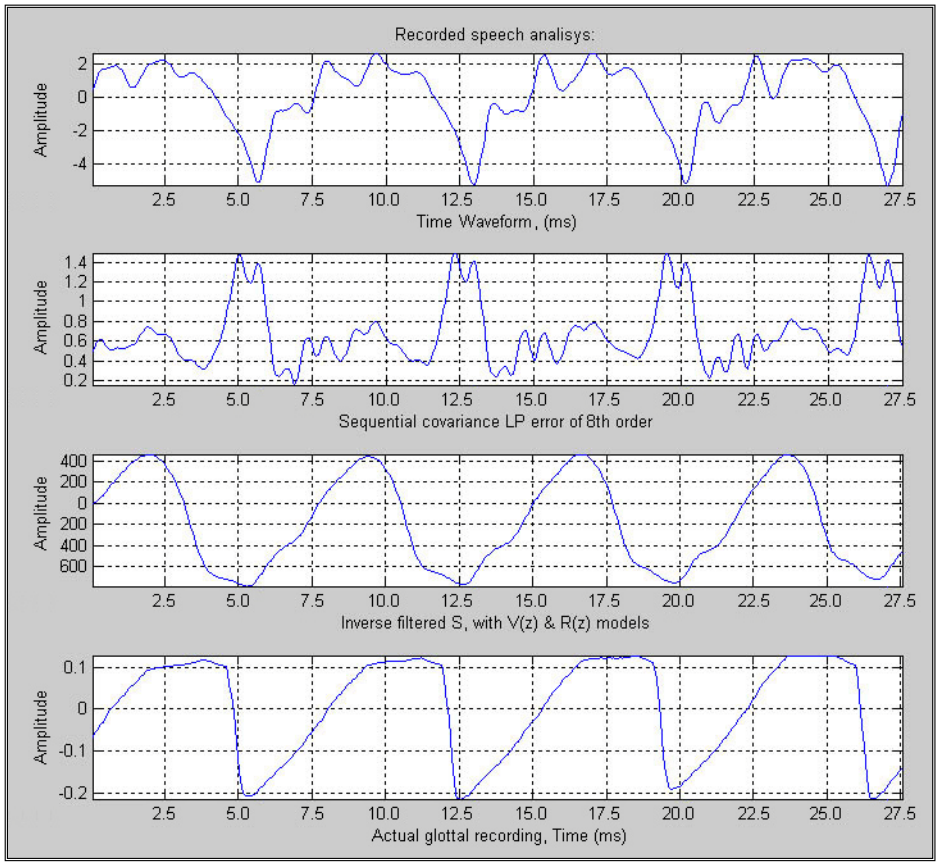


Figure 7.6: Comparison between recorded and inverse filtered glottal signal.

In Figure 7.7, a comparison between a synthesized and recorded speech analysis had been included. As can be seeing, there are important similarities between the two glottal signals extracted via inverse filtering. Importantly, the Rosenberg-type excitation pulse used in for the synthesis was perfectly reconstructed.

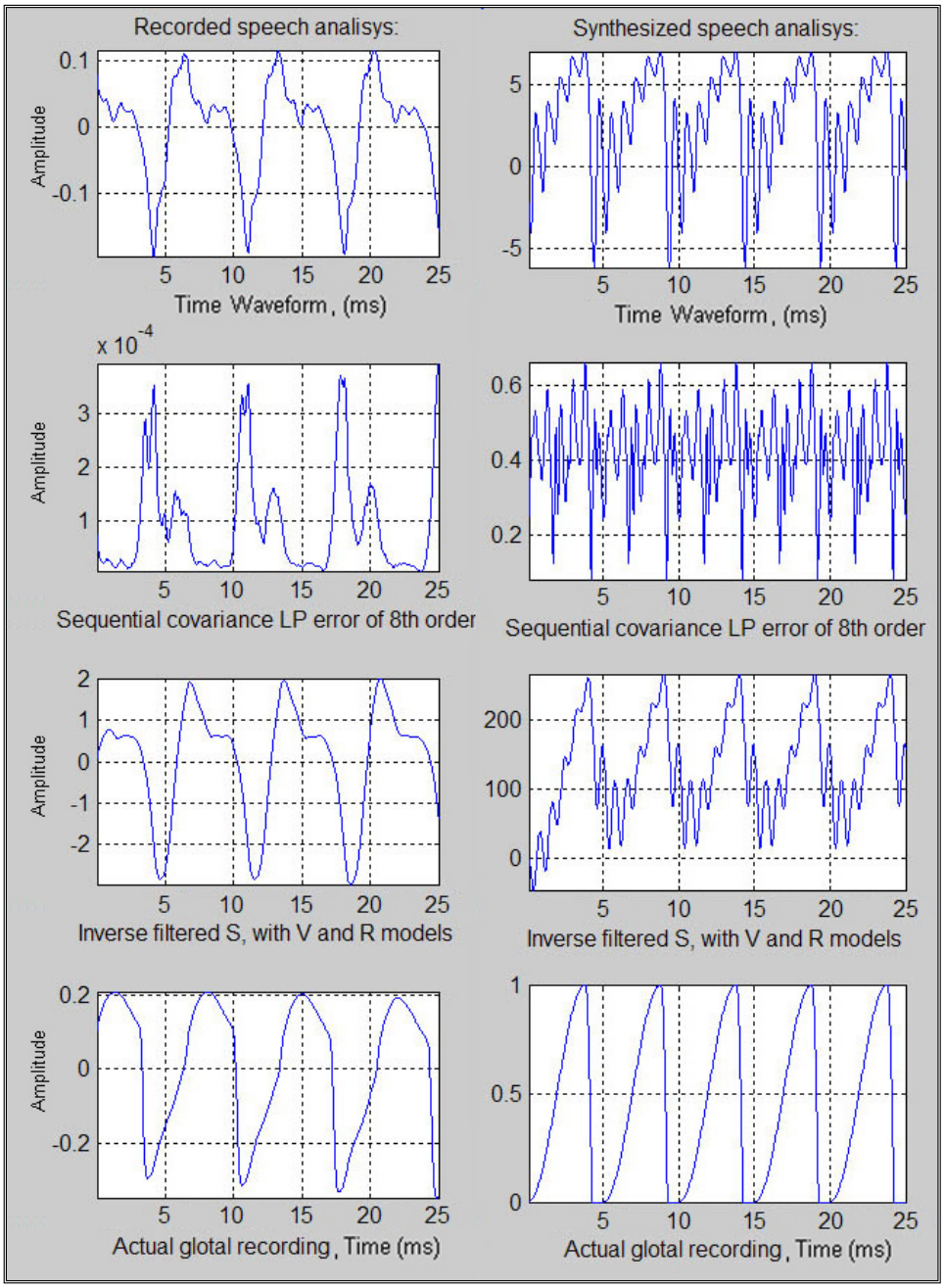


Figure 7.7: Comparison between synthesized and recorded speech signals.

In this corpus, the performance of the Optimum-Path Forest (OPF) classifier was compared against that of the other six methods and the results are summarized in Figures 7.8 and 7.9. After establishing the features providing best performance, three double-feature sets of data containing combinations of MFCCs for both glottal and speech signals and those results are shown in Figure 7.9.

Figure 7.8 summarizes the performance results for each classifier for glottal symmetry (GS) alone, and MFCC features for glottal and speech signals. We have to note that in the case of GS, many examples from known classes were presented for training, thus resembling a *clustering* learning procedure. As it can be seen, OPF, SVM, k-NN, and BC performed better than, GMM, and C4.5 in all data cases. The performance of the GS dataset was very good for that single feature reaching a 97.8% recognition performance for the k-NN classifier, thus reinforcing the initial rationale for its selection. For that feature alone, performance of the other classifiers exceeded the 90% mark (90% for C4.5), except of GMM which scored a 55.8% recognition rate.

Glottal-only MFCC features of orders 4, 6, 8 and 10 were also used to train and test the classifiers and their performance was generally slightly better that for GS alone. Recognition results ranged from 57% for GMM with 10 MFCCs, to 100% using 4 and 6 glottal MFCCs for OPF, BC, KNN and SVM to 97.6% for C4.5 and 82.8% for GMM for the same features. Among the classifiers for the glottal-only case SVM provided the best overall results, closely followed by OPF.

For the case of speech-only MFCC features of orders 6, 8, 10 and 12 performance was substantially lower overall, so that the glottal case with recognition performance

ranging from 29.8% for 10 MFCC for GMM to 95% for 6 MFCC for SVM. Overall, the 6 MFCC case provided best performance, with SVM providing the best results, closely followed by OPF, as in the glottal case.

Classifier performance for the combinations of the best glottal and speech cases was also tested. Features comprising MFCCs of orders 6 to 8 for each signal type were used for training each system. The obtained average recognition performance ranged from 67% for GMM and MFCCs of order 8 for both glottal and speech [8(gl)+8(sp)], to 99% for several classifiers particularly in the 6(gl)+6(sp) case. In terms of computation times, k-NN was the fastest system followed by C4.5 and OPF. The most demanding systems was the SVM. Comparing the two overall best classifiers, the OPF was about 3000 times faster than SVM when measuring the complete training and testing cycle time.

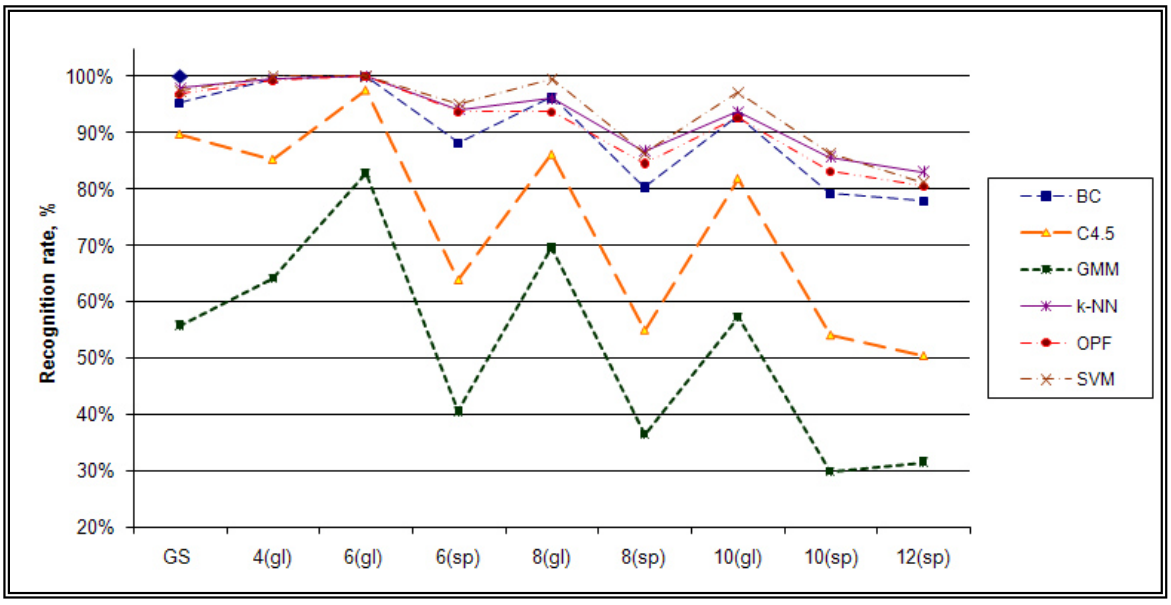


Figure 7.8: Mean recognition rates for 9 individual features using ANGRY versus HAPPY versus SAD versus NEUTRAL after 10 rounds.

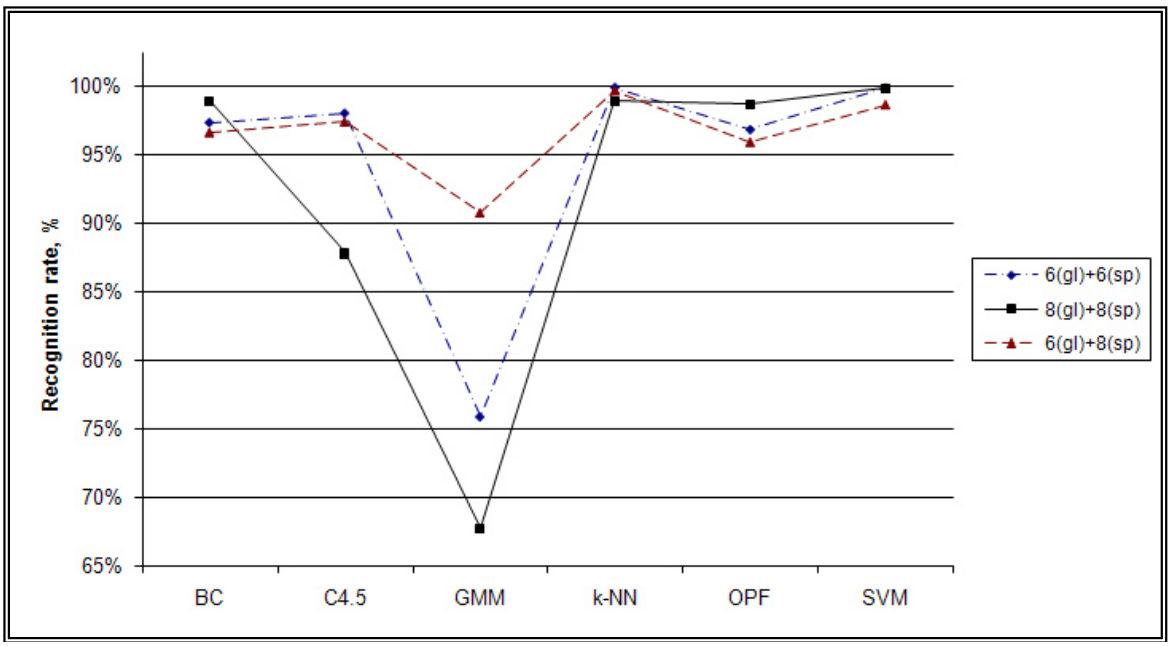


Figure 7.9: Mean recognition rates for 3 combinations of features using ANGRY versus HAPPY versus SAD versus NEUTRAL after 10 rounds.

7.3 Corpus 3: Six Emotions in "Waiting for Godot"

I. Glottal Symmetry tests for individual subjects in speaker dependent, text independent conditions using clean speech

Individual test on glottal symmetry of clean speech were performed first to compare them to the combined tests performed later. This was done separately for each of the three male subjects. The glottal symmetry data in this corpus was extracted in a sequence of 5 consecutive glottal pulses at a time. They were later fed to the GMM for classification. The results are shown in Tables 7.4 through 7.15 as percentage of correct recognition. As can be observed from the balanced test results, for each speaker the correct emotion was recognized over 50% of the time with the exception of Surprise for speaker 3 which was correctly recognized 47.5% of the time. As expected, in the unbalanced test case the results were not as accurate although they all produced high

recognition rates as well. All results were obtained through a 10-fold average performance.

Subject 1:

Table 7.4: GS confusion matrix for Subject 1 for 4 emotions on a balanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**94.41**	4.55	0.47	0.57
HAPPY	24.87	**74.77**	0.22	0.14
SAD	37.67	1.40	**60.61**	0.32
NEUTRAL	1.25	13.80	19.39	**65.56**

Table 7.5: GS confusion matrix for Subject 1 for 4 emotions on an unbalanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**44.85**	38.00	7.02	10.13
HAPPY	5.11	**78.38**	8.13	8.38
SAD	11.28	28.38	**51.57**	8.77
NEUTRAL	22.13	10.21	4.94	**62.72**

Table 7.6: GS confusion matrix for Subject 1 for 6 emotions on a balanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**61.47**	14.91	3.08	1.68	18.64	0.22
HAPPY	19.57	**66.56**	1.25	0.54	12.04	0.04
SAD	25.02	5.45	**55.52**	1.04	12.90	0.07
NEUTRAL	13.94	4.09	16.42	**52.36**	13.05	0.14
FEAR	16.31	2.01	12.76	0.65	**68.16**	0.11
SURPRISE	12.94	2.33	11.15	0.65	6.56	**66.37**

Table 7.7: GS confusion matrix for Subject 1 for 6 emotions on an unbalanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**48.46**	8.34	4.98	12.43	24.47	1.32
HAPPY	18.21	**38.72**	6.09	9.36	25.28	2.34
SAD	20.98	5.40	**36.65**	10.51	25.06	1.40
NEUTRAL	3.06	2.85	3.74	**66.78**	21.91	1.66
FEAR	3.11	2.47	6.72	7.49	**78.12**	2.09
SURPRISE	11.91	5.28	18.34	6.98	27.06	**30.43**

Subject 2:

Table 7.8: GS confusion matrix for Subject 2 for 4 emotions on a balanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**80.49**	8.13	10.78	0.60
HAPPY	32.47	**56.57**	10.48	0.48
SAD	9.64	2.89	**87.17**	0.30
NEUTRAL	3.31	19.16	9.04	**68.49**

Table 7.9: GS confusion matrix for Subject 2 for 4 emotions on an unbalanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**94.20**	2.52	1.38	1.90
HAPPY	34.90	**62.43**	1.24	1.43
SAD	29.52	2.71	**65.20**	2.57
NEUTRAL	3.00	23.05	24.14	**49.81**

Table 7.10: GS confusion matrix for Subject 2 for 6 emotions on a balanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**60.96**	10.30	4.40	3.37	7.17	13.80
HAPPY	22.71	**47.53**	5.66	3.92	5.30	14.88
SAD	12.89	3.92	**70.84**	1.57	3.13	7.65
NEUTRAL	5.18	2.35	5.36	**73.25**	4.28	9.58
FEAR	3.55	2.35	13.31	0.54	**70.13**	10.12
SURPRISE	0.00	1.02	2.77	0.54	5.66	**90.01**

Table 7.11: GS confusion matrix for Subject 2 for 6 emotions on an unbalanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**69.38**	10.71	11.90	1.48	4.48	2.05
HAPPY	26.86	**53.91**	13.14	1.52	3.43	1.14
SAD	26.86	2.24	**63.32**	2.62	3.10	1.86
NEUTRAL	19.81	3.00	17.00	**54.57**	3.33	2.29
FEAR	34.19	2.86	17.00	1.95	**40.05**	3.95
SURPRISE	25.10	2.86	13.71	2.05	4.43	**51.85**

Subject 3:

Table 7.12: GS confusion matrix for Subject 3 for 4 emotions on a balanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**93.76**	3.00	2.95	0.29
HAPPY	38.62	**55.00**	6.14	0.24
SAD	16.14	0.57	**83.24**	0.05
NEUTRAL	0.57	20.00	13.19	**66.24**

Table 7.13: GS confusion matrix for Subject 3 for 4 emotions on an unbalanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**43.98**	21.02	23.55	11.45
HAPPY	23.86	**41.81**	23.19	11.14
SAD	23.25	13.07	**51.39**	12.29
NEUTRAL	5.18	11.69	10.00	**73.13**

Table 7.14: GS confusion matrix for Subject 3 for 6 emotions on a balanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**72.57**	5.62	6.62	4.62	8.05	2.52
HAPPY	21.67	**50.14**	9.38	4.33	11.62	2.86
SAD	3.90	1.14	**83.19**	3.05	6.86	1.86
NEUTRAL	2.86	0.43	20.10	**66.28**	7.71	2.62
FEAR	3.24	1.05	27.90	1.95	**61.24**	4.62
SURPRISE	5.43	0.33	30.62	0.19	15.90	**47.53**

Table 7.15: GS confusion matrix for Subject 3 for 6 emotions on an unbalanced text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**35.48**	17.53	7.05	9.34	8.43	22.17
HAPPY	17.35	**35.17**	7.29	9.10	9.40	21.69
SAD	18.43	10.66	**31.51**	10.36	8.92	20.12
NEUTRAL	5.78	4.58	7.23	**58.01**	7.35	17.05
FEAR	5.54	1.99	3.67	4.94	**66.45**	17.41
SURPRISE	10.60	3.80	8.61	1.75	7.77	**67.47**

The average recognition performance for each subject separately as well as for all subjects together are shown in Table 16. The results obtained in corpus 3 are consistent

with the findings in corpus 2, where the glottal symmetry was also investigated. The overall better performance for the 4-emotions case in corpus 3, as compared to that of corpus 2, is because of the way the tests were conducted. In the latter the average GMM performance for the GS, as shown in Figure 7.8, is taken from a combined test for all subjects. In corpus 3 however the average is taken for all subjects, but over their individually tested performance, as shown in Table 16.

Table 7.16: Average recognition performance for subjects 1, 2, and 3 separately and combined for 4-emotions and 6-emotions.

Number of emotions:	*Subject 1*	*Subject 2*	*Subject 3*	*All subjects*
4	66.61	70.55	63.57	**66.91**
6	55.80	62.15	56.26	**58.07**

II. Glottal Symmetry tests for speaker and text independent system using: clean speech, LPF, SNR-30dB, and SNR-10dB

Emotion recognition performance for all subjects with added white Gaussian noise and distortion due to lowpass filtering was examined next. These conditions were applied prior to glottal flow extraction via the inverse filtering thus demonstrating emotion recognition performance in a real world conditions.

A. Clean Speech: Glottal Symmetry results using balanced and unbalanced training sets

Table 7.17: Four emotions on an unbalanced speaker and text independent test.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**96.92**	2.55	0.44	0.09
HAPPY	36.46	**63.24**	0.16	0.14
SAD	25.11	1.58	**73.21**	0.10
NEUTRAL	1.49	27.78	22.41	**48.32**

Table 7.18: Four emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**98.74**	0.51	0.42	0.33
HAPPY	24.11	**75.47**	0.18	0.24
SAD	35.24	0.13	**64.59**	0.04
NEUTRAL	0.38	20.56	15.18	**63.88**

Table 7.19: Four emotions for a balanced speaker and text independent test - GS based: 4'167 GS's per emotion, per speaker, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**91.54**	1.68	3.10	3.68
HAPPY	16.86	**79.96**	1.42	1.76
SAD	37.52	0.86	**59.68**	1.94
NEUTRAL	0.26	21.16	12.70	**65.88**

Table 7.20: Six emotions on an unbalanced speaker and text independent test - all emotions, all utterances, all speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**68.50**	24.62	1.89	0.76	4.19	0.04
HAPPY	22.61	**70.32**	1.75	1.02	4.30	0.00
SAD	15.28	11.89	**69.34**	0.57	2.88	0.04
NEUTRAL	11.95	8.88	16.15	**60.44**	2.48	0.10
FEAR	20.07	14.13	16.43	1.05	**48.29**	0.03
SURPRISE	15.32	9.90	14.19	0.85	4.49	**55.25**

Table 7.21: Six emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**85.37**	12.18	0.44	1.56	0.18	0.27
HAPPY	30.44	**67.32**	0.09	1.44	0.44	0.27
SAD	26.96	5.91	**66.02**	0.49	0.42	0.20
NEUTRAL	15.67	4.04	23.44	**56.51**	0.18	0.16
FEAR	22.31	8.64	10.31	1.29	**57.07**	0.38
SURPRISE	20.27	4.58	7.49	0.56	0.33	**66.77**

Table 7.22: Six emotions for a balanced speaker and text independent test - GS based: 4'167 GS's per emotion, per speaker, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**78.40**	5.90	7.24	6.80	0.52	1.14
HAPPY	34.20	**54.70**	5.90	3.84	0.38	0.98
SAD	33.24	4.00	**56.90**	4.22	0.66	0.98
NEUTRAL	19.02	0.56	20.20	**59.00**	0.46	0.76
FEAR	27.58	1.50	15.30	5.58	**48.86**	1.18
SURPRISE	24.54	3.98	17.08	4.82	0.74	**48.84**

Table 7.23: Average recognition performance for all subjects combined for 4-emotions and 6-emotions in clean speech.

Number of emotions:	*ANGRY*	*HAPPY*	*SAD*	*NEUTRAL*	*FEAR*	*SURPRISE*	*All emotions*
4	95.73	72.89	65.83	59.36	-	-	**73.45**
6	77.42	64.11	64.09	58.65	51.41	56.95	**62.11**

Table 7.23 shows the average recognition performance of all subjects combined in both 4 and 6 emotion cases, where the average performance is also depicted.

B. Lowpass Filtered Speech:

In this section all clean speech was passed through a lowpass filter with passband frequency of 600 Hz and stopband frequency of 800 Hz. As a result, the newly filtered speech was severely altered and unintelligible. The classification results from these tests are displayed in Tables 7.24 through 7.29. The average recognition performance in the case of lowpass filtered signal is depicted in Table 7.30.

Table 7.24: Four emotions unbalanced speaker and text independent test after LPF.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**62.01**	27.44	1.65	8.90
HAPPY	0.50	**86.83**	1.91	10.76
SAD	18.91	19.13	**55.45**	6.51
NEUTRAL	16.24	9.39	5.23	**69.14**

Table 7.25: Four emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers after LPF.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**51.62**	43.70	2.35	2.33
HAPPY	0.56	**93.54**	3.14	2.76
SAD	26.35	25.34	**46.60**	1.71
NEUTRAL	10.19	23.95	16.77	**49.09**

Table 7.26: Four emotions for a balanced speaker and text independent test - GS based: 2'406 GS's per emotion, per speaker, random sequence of speakers after LPF.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**42.77**	17.17	28.54	11.52
HAPPY	14.10	**48.02**	26.69	11.19
SAD	6.96	9.27	**73.85**	9.92
NEUTRAL	2.45	16.15	0.37	**81.03**

Table 7.27: Six emotions on an unbalanced speaker and text independent test - all emotions, all utterances, all speakers after LPF.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**47.56**	17.14	9.38	14.74	8.98	2.20
HAPPY	13.04	**47.69**	9.08	16.71	10.89	2.59
SAD	13.52	10.00	**55.71**	11.72	7.30	1.75
NEUTRAL	1.44	8.31	5.82	**74.36**	8.07	2.00
FEAR	7.22	12.44	17.34	14.53	**45.88**	2.59
SURPRISE	5.93	11.89	16.26	12.33	10.98	**42.61**

Table 7.28: Six emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers after LPF.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**44.64**	28.76	5.21	15.28	6.11	0.00
HAPPY	7.35	**64.38**	4.83	15.85	7.59	0.00
SAD	16.99	24.66	**39.42**	13.18	5.75	0.00
NEUTRAL	10.06	19.23	8.03	**57.38**	5.30	0.00
FEAR	9.27	26.07	5.64	7.61	**51.41**	0.00
SURPRISE	8.76	16.56	6.22	6.60	9.23	**52.63**

Table 7.29: Six emotions for a balanced speaker and text independent test - GS based: 2'406 GS's per emotion, per speaker, random sequence of speakers after LPF.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**49.37**	13.68	6.15	14.89	4.18	11.73
HAPPY	19.69	**45.83**	6.03	13.06	4.62	10.77
SAD	21.77	8.48	**43.07**	11.75	4.10	10.83
NEUTRAL	0.94	2.62	3.16	**79.44**	4.03	9.81
FEAR	7.28	2.56	9.50	10.54	**54.03**	16.09
SURPRISE	6.74	4.16	5.32	7.34	8.36	**68.08**

Table 7.30: Average recognition performance for all subjects combined for 4-emotions and 6-emotions after LPF.

Number of emotions:	*ANGRY*	*HAPPY*	*SAD*	*NEUTRAL*	*FEAR*	*SURPRISE*	*All emotions*
4	52.13	76.13	58.63	66.42	-	-	**63.33**
6	47.19	52.63	46.07	70.39	50.44	54.44	**53.53**

As expected, the average recognition performance results from all the tests and all emotions shown in Table 7.30 are lower than the ones displayed in Table 7.23, due to the use of a lowpass filtered signal in the former. For visual comparison, the waveform of a sample utterance for one emotion of clean speech is given at the top of Figure 7.10. Below are the noisy counterparts with SNR=30dB and SNR=10dB, then clean signals energy and its corresponding lowpass filtered version. It can be observed that the change introduced to the speech signal after LPF is applied is quite dramatic.

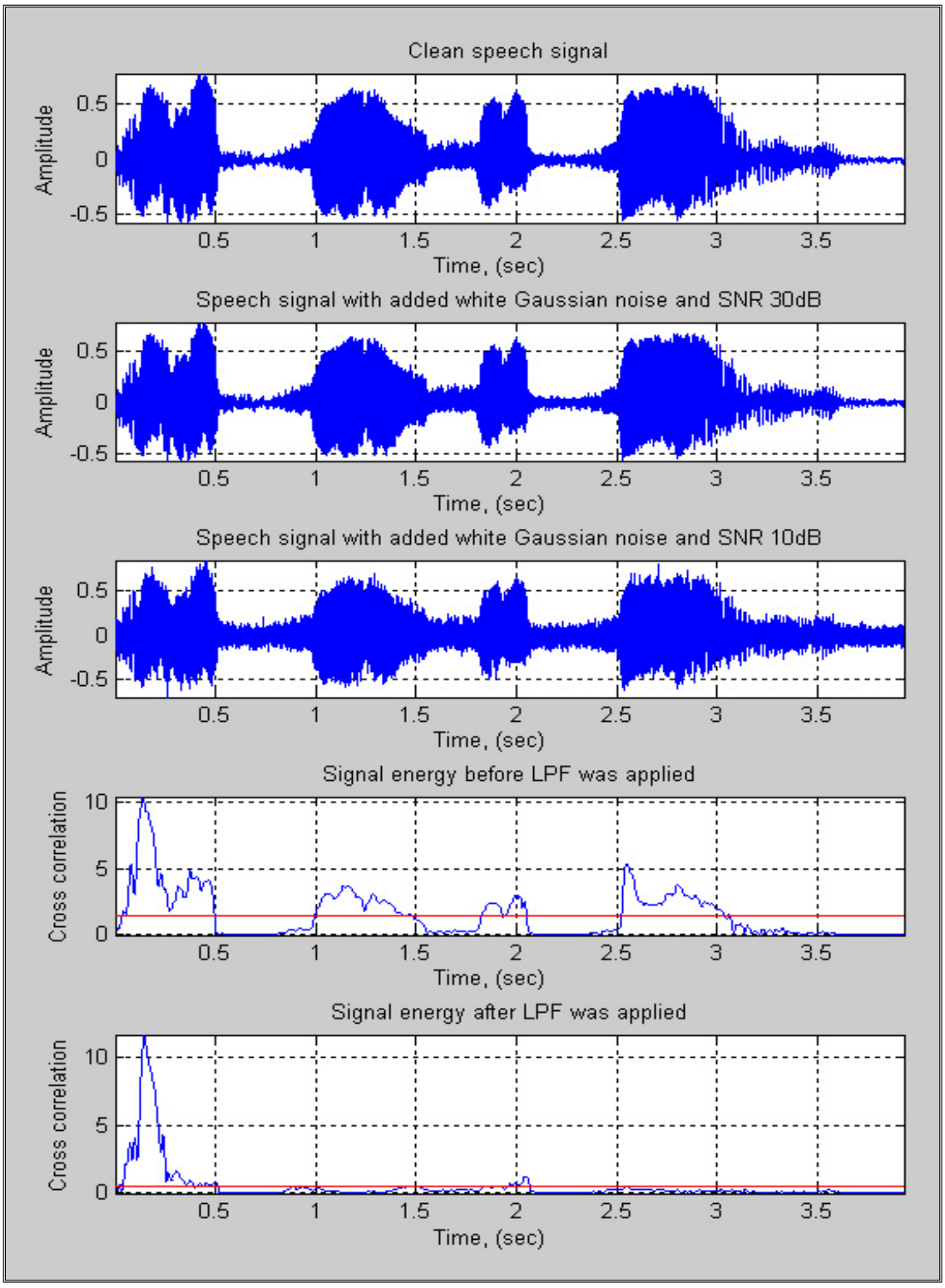


Figure 7.10: Clean and distorted speech.

C. SNR = 30 dB:

In addition to the tests performed in clean and LPF speech signals, a white Gaussian noise was generated and added to the clean speech at a signal to noise ratios of 30 and 10 dB before the glottal symmetry was obtained. These tests are displayed in Tables 7.31 and 7.36.

Table 7.31: Four emotions balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 30 dB.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**93.91**	0.41	3.51	2.17
HAPPY	39.63	**54.70**	3.05	2.62
SAD	26.52	0.71	**70.51**	2.26
NEUTRAL	0.75	23.74	23.70	**51.81**

Table 7.32: Six emotions balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 30 dB.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**80.77**	0.90	12.84	2.28	1.40	1.81
HAPPY	37.20	**45.38**	11.18	2.86	1.66	1.72
SAD	24.09	0.58	**70.07**	2.45	1.46	1.35
NEUTRAL	22.92	0.54	27.48	**47.10**	1.08	0.88
FEAR	30.26	0.45	30.92	2.30	**33.96**	2.11
SURPRISE	28.37	0.47	27.27	2.39	1.33	**40.17**

Table 7.33: Average recognition performance for all subjects and all emotions combined for 4-emotions and 6-emotions for SNR = 30 dB.

Number of emotions:	*All emotions*
4	**67.73**
6	**52.91**

D. SNR= 10 dB:

Table 7.34: Four emotions balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 10 dB.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**50.96**	42.09	4.28	2.67
HAPPY	34.01	**58.54**	4.52	2.93
SAD	22.45	29.35	**45.34**	2.86
NEUTRAL	14.01	28.27	7.40	**50.32**

Table 7.35: Six emotions balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers, SNR = 10 dB.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**37.14**	18.87	7.79	14.28	20.84	1.08
HAPPY	16.20	**40.79**	8.65	12.96	20.58	0.82
SAD	12.79	12.04	**43.66**	13.03	17.45	1.03
NEUTRAL	2.12	8.82	18.70	**50.84**	18.29	1.23
FEAR	8.05	12.04	7.98	11.61	**59.29**	1.03
SURPRISE	3.75	9.47	10.02	8.44	19.21	**49.11**

Table 7.36: Average recognition performance for all subjects and all emotions combined for 4-emotions and 6-emotions for SNR = 10 dB.

Number of emotions:	*All emotions*
4	**51.29**
6	**46.81**

Logically, the average performance of the system when WGN with SNR=10dB was added is lower than the one with SNR=30dB. The results in the case of the clean speech signal also show consistency as they are notably higher. Despite of the lower performance in noisy environments the average performance rate was high enough to

make positive emotion identification. It should be noted here that the LPF system performed better in the case of 6-emotions showing 53.53%, as compared to the one with SNR=30dB showing 52.91%. In the case of 4-emotions was the opposite where the latter outperformed the former with 67.73% to 63.33% respectively.

III. 6th order MFCC tests for speaker and text independent system using: clean speech and SNR-10dB on speech signal

In this group of tests, the detection accuracy for four and six emotions was tested using 6th order MFCC coefficients on both clean and noisy environments. As before, the signal-to-noise ratio of 10dB was applied to the original speech before the feature extraction. The results are depicted in Tables 7.37 through 7.42. The average results of the performance of the MFCC from clean speech are shown in Table 7.39 and the ones from the SNR=10dB are depicted in 7. 42. From there can be concluded that GS performs better than MFCC of 6th order in both clean and noisy conditions. This result is also consistent with the one obtained from corpus 2 as shown in Figure 7.8.

A. Clean Speech: MFCC results

Table 7.37: Four emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**46.12**	21.70	15.46	16.72
HAPPY	33.58	**38.30**	13.62	14.50
SAD	27.73	10.74	**45.98**	15.55
NEUTRAL	14.32	28.73	15.50	**41.45**

Table 7.38: Six emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**31.57**	26.51	8.60	8.56	10.39	14.37
HAPPY	19.00	**41.47**	9.83	7.12	9.39	13.19
SAD	21.97	20.70	**29.29**	7.21	9.26	11.57
NEUTRAL	16.20	17.38	20.39	**23.41**	9.52	13.10
FEAR	18.78	17.38	11.97	5.55	**29.68**	16.64
SURPRISE	19.26	20.35	12.49	11.22	9.65	**27.03**

Table 7.39: Average recognition performance for all subjects and all emotions combined for 4-emotions and 6-emotions for MFCC on clean speech.

Number of emotions:	*All emotions*
4	**42.96**
6	**30.41**

B. SNR=10 dB:

Table 7.40: Four emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL
ANGRY	**39.09**	25.93	10.35	24.63
HAPPY	23.16	**41.43**	11.60	23.81
SAD	23.81	17.06	**35.58**	23.55
NEUTRAL	21.73	21.99	15.50	**40.78**

Table 7.41: Six emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**27.62**	16.49	12.21	11.90	11.17	20.61
HAPPY	20.69	**25.06**	14.85	10.61	9.61	19.18
SAD	17.01	12.42	**30.40**	11.43	9.00	19.74
NEUTRAL	12.77	10.26	19.26	**30.39**	8.92	18.40
FEAR	15.19	9.70	15.32	9.00	**29.49**	21.30
SURPRISE	14.59	10.69	18.74	18.83	8.01	**29.14**

Table 7.42: Average recognition performance for all subjects and all emotions combined for 4-emotions and 6-emotions for MFCC on speech with SNR=10dB.

Number of emotions:	*All emotions*
4	**39.22**
6	**28.68**

IV. Classical prosodic feature tests for speaker and text independent system using: clean speech

Since in corpus 1 only four emotions were tested, here the performance of the classical prosodic features were compared to the previous features using six emotions.

Table 7.43: Six emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**40.00**	23.85	10.38	3.08	13.46	9.23
HAPPY	23.46	**32.70**	11.15	2.69	20.00	10.00
SAD	8.85	5.77	**58.84**	9.62	11.92	5.00
NEUTRAL	2.69	2.69	31.15	**52.70**	5.77	5.00
FEAR	6.54	5.00	1.15	1.92	**79.62**	5.77
SURPRISE	10.77	14.23	15.00	4.23	5.00	**50.77**

Table 7.44: Average recognition performance for all subjects and all emotions combined for 6-emotions using classical prosodic features.

Number of emotions:	*All emotions*
6	**52.44**

Comparing the results from Table 7.44 to the ones in Table 7.23 it can be concluded that the GS in clean speech performs with higher confidence than the 11 classical prosodic features alone. Moreover the GS performs slightly better than the classical features in noisy conditions in the case of SNR=30dB as shown in Table 7.33.

V. Combined features (Classical with ToBI) for speaker and text independent system using: clean speech

Combined features from the classical and ToBI domains were used for comparison in the six emotions case. These results are displayed in Table 7.45. As shown in corpus 1, the average performance of the combined system has higher confidence than the one using classical prosodic features alone. This is also confirmed in the case of 6-emotions, which becomes evident by comparing Tables 7.44 and 7.46.

Table 7.45: Six emotions for a balanced speaker and text independent test - utterance based: 100 utterances per emotion, random sequence of speakers.

Actual \ Detected	ANGRY	HAPPY	SAD	NEUTRAL	FEAR	SURPRISE
ANGRY	**43.84**	26.15	13.46	3.85	8.85	3.85
HAPPY	20.77	**43.45**	14.62	4.62	11.54	5.00
SAD	6.92	10.38	**71.93**	2.31	6.15	2.31
NEUTRAL	5.00	3.46	12.69	**69.62**	5.77	3.46
FEAR	4.23	1.92	12.31	1.15	**67.70**	12.69
SURPRISE	1.15	0.38	3.08	0.00	5.00	**90.39**

Table 7.46: Average recognition performance for all subjects and all emotions combined for 6-emotions using combined features (Classical with ToBI).

Number of emotions:	*All emotions*
6	**64.49**

One other observation is that the combined system shows better average performance then the GS in clean speech, but in noisy environment GS shows the most confident results by far.

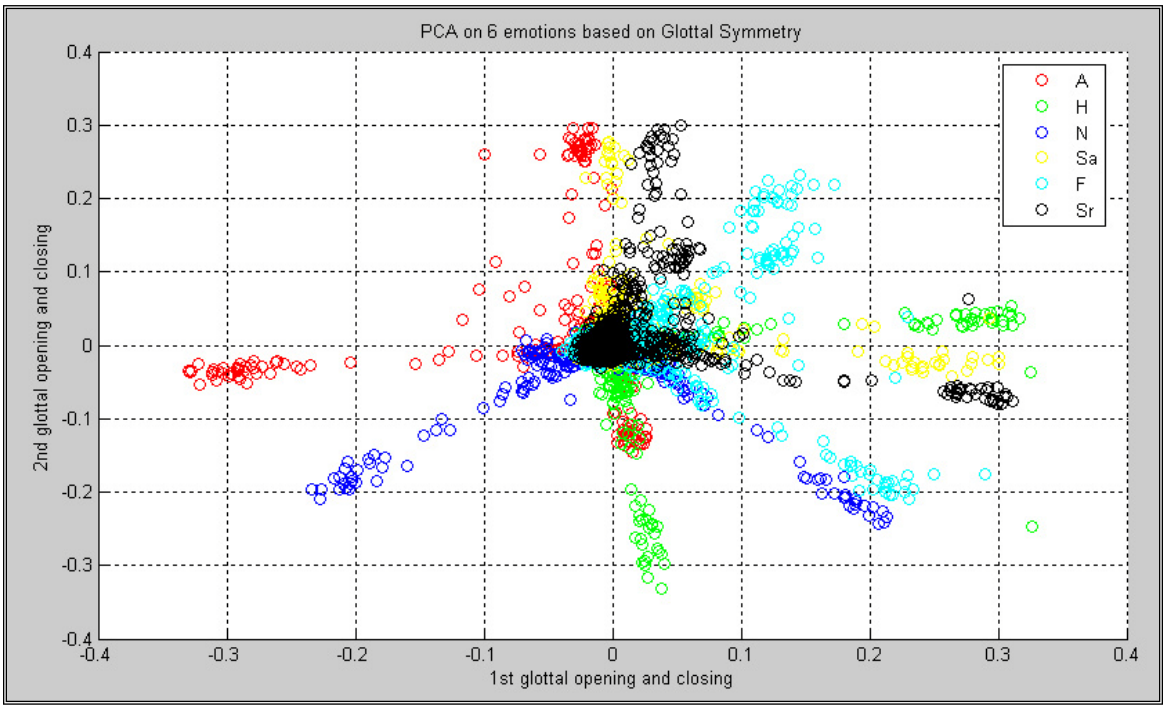


Figure 7.11: PCA analysis of the 1st vs. 2nd Glottal Symmetry for 6 emotions.

To compare the first to the second glottal opening and to observe their relation to one another, Principal Component Analysis (PCA) method was adopted to graphically depict their relation from 3D to 2D plane as shown in Figure 7.11. It is apparent that

although they have a substantial overlap in the middle, a separation for each emotional class by just looking at two neighboring glottal pulses in a sequence is possible due to the clusters formed in the sides.

The average performance for each of the three corpora is shown in Table 7.47. To be fare, they were taken over a clean speech signal only. As expected corpus 2 had higher performance confidence since it was recorded in an anechoic chamber and the original speech quality was much higher. It also had a prerecorded glottal signal used for reference, which was collected through a laryngograph. Corpus 3 showed a slightly lower overall performance confidence than corpus 1 for two reasons. In the result for corpus 3 both 4-emotions and 6-emotions cases are considered and the latter consistently shows lower performance rate as observed from the results. The second reason is that it used a larger variety of feature domains, which were not present in corpus 1.

Table 7.47: Average recognition performance on clean speech for all three corpora using all subjects, features, emotions and classifiers combined.

Corpus 1 (Godot: 4 emotions in a real world dialog)	*Corpus 2 (4 emotions in short and long emotional, read sentences)*	*Corpus 3 (Godot: 6 emotions in a real world dialog)*
60.84	**81.05**	**54.31**

CHAPTER 8

Summary

8.1 Evaluation of Emotion Classification Performance

Corpus 1

One of the new features subject to this study was ToBI, for which the automatic detection followed general rules of ToBI annotation in the tone tier alone. Thus the system provided solely the pitch relationships. More specifically, it described the shape of the glottal waveform within any given intonational phrase.

It was found adequate that in ToBI the number of occurrences of different accents in the tonal tier is passed to the classifiers, rather than the exact time sequence of these accents. One reason for this is that the classical approach naturally provides fixed length vectors without the truncation of useful data, which is suitable for classification.

The overall system evaluation when ToBI, classical prosodic feature domain, and all combined features were used shows that the combined method performed better than when the first two systems used separately. An improvement of computational speed and

system accuracy was obtained after pruning achieved by analyzing the Mutual Information among features. The results were refined using optimization with Sequential Forward Selection, and consequently the number of features was reduced. On average, a 45% reduction in the number of features resulted in a 3 times lower computational cost while achieving over 20% recognition improvement.

Performing more tests along with the use of a larger database with a variety of speakers of different genders and age would be very beneficial for system improvement, which is why corpus 2 was developed. Furthermore other classification algorithms in addition to GMM were considered in corpus 2, such as k-Nearest Neighbor (k-NN), Support Vector Machine (SVM), and others.

Corpus 2

The glottal source signal, which is quasi-periodic during voicing conveys important characteristics about the identity of the speaker as well and the speaking style. A number of studies have pointed out the key role of voicing in the production of loud, soft and stressed speech, and in singing. In this study the role of the glottal signal is investigated for the classification of spoken emotion. The glottal waveform of the first voiced portion of a speech utterance was estimated via inverse filtering. The glottal symmetry as well as MFCC vectors of various lengths were computed and used for the training and testing of seven different classification systems. The corresponding speech signal was also processed and the effectiveness of similar feature vectors of various lengths was examined. A combination of the most effective glottal and speech features was also tested.

The Optimum Path Forest (OPF), a new classification method with application to emotion recognition exhibits good performance for non-easily separable classes. This was also the case with the problem at hand. This method owes its efficiency to the image foresting transform (IFT) algorithm it includes. Its effectiveness is due to the choice of prototypes based on the minimum-spanning tree (MST), which minimize the errors of the optimum-path forest in the training set. The method's performance was compared with previously established classification systems as well.

While error bounds of the proposed method have not yet been theoretically established, a sizeable body of experimental results indicates that the classification error in the training set is a minimum, due to a theoretical relationship between the MST and Shortest Path Trees (SPT) for the f_{max} path-cost function used by the OPF. Therefore, assuming that the training set is representative of the problem, these errors in the test set will be minimized as well. However, when the prototypes are estimated by the MST, an error may still occur in the training set whenever a sample t can be reached by two or more optimum paths from prototypes of different labels. If the optimum paths that reach t belong to the MST, then the prototypes in the path also avoid the error. Therefore, assuming that the results in the training set can be extended to the test set, one may conjecture that the error upper bound corresponds to the number of those tied samples, which is usually low. Note that this is certainly the error upper bound for the training set.

Results confirm that glottal information is rich in emotion clues and it presents a very effective source for achieving recognition of spoken emotion. Parameters as simple as the sequence of glottal symmetry values at the beginning of a spoken utterance proved quite effective for classification. Low order glottal MFCC resulted in best recognition

performance indicating that real time deployment of such classifiers is quite feasible. Glottal information was more effective than speech alone. Furthermore, combinations of glottal and speech features slightly degraded performance.

Best classification performance was provided by SVM and OPF; lowest performance was that of GMM. In terms of computation times, k-NN was the fastest. For the top two classifiers, OPF was substantially faster than SVM making the former quite appealing for this application.

The obtained average recognition performance on randomly selected test was perfect for several the classifiers, particularly in the case of 6 glottal MFCCs. This indicates that for this medium-sized database of 5 male and 5 female speakers, speaking 10 sentences each in an anechoic environment, every sentence with different emotional expression, for a four emotions classification task, several classifiers reached the correct decision for at least one type of feature vector.

As mentioned earlier this corpus provided speaker and text depended environment, hence the results obtained were consistently better than the ones obtained from the other corpora. That facilitated the goal set when testing corpus 2, namely to establish if the MFCCs of the speech and glottal signal can play role in the field of emotion recognition. The work in corpus 3 extends the work to texts not seen by the classifiers and it includes more emotions as well. In it both better quality recording environment and noisy acoustic environment are used via various communication channels.

Corpus 3

One of the goals of corpus 3 was to study the performance of the glottal symmetry as class separator while increasing the number of emotions to six. The attribute's performance was analyzed in clean and in a variety of noisy conditions. As shown in Tables 7.17 through 7.22 recognition performance was mostly above 50% for every emotional class. A comparison in this "GS Clean Speech" category between the two balanced test, both utterance and GS based, shows that the first performs better, in both 4 and 6 emotional classes. However the difference between the utterance based balanced test and the unbalanced test is marginal. The situation is similar for the LPF speech case, with the difference that the confidence success rate is generally lower for all classes. This, as expected is due to the severe frequency band limitation of the speech quality from the filtering process.

The robustness test continued by adding white Gaussian noise (WGN) to the clean speech with SNR = 30dB where remarkably high recognition rates up to 94% for class "Angry" in 4 emotional case and 81% for the 6 emotions case were observed. For SNR = 10dB conditions, the achieved recognition rate was 55% for 4 emotions and 59% for the 6 class set. A sample of the speech signal before and after the noise and filtering were applied is depicted in Figure 7.8.

In the next round of tests the MFCCs were involved and their results are shown in Tables 7.37 through 7.42. Although their performance rates were substantially lower than the corresponding tests in clean speech, they performed consistently well in noisy environment, and particularly in the case of SNR = 10dB for both 4 and 6 emotion classes. Comparing results for SNR = 10dB in both GS and glottal MFCCs the GS is

much more robust under severe quality degradation. Furthermore, Table 7.43 depict the performance of classical prosodic features on clean speech, which are comparable to the tests, ran on corpus 1 for the 4-emotions set.

Although classical features performed well on clean speech they failed under any of the noisy environments used herein. Finally, in Table 7.45 the results for the newly introduced features in corpus 1 combined system with both classical prosodic and ToBI features are displayed. Comparing these results to their counterparts when classical features alone were used, it can be seen that ToBI improves the recognition rate when added to classical prosodic feature sets, thus confirming the importance of this domain. It did not however show satisfactory results when tested to noisy conditions.

8.2 Statistical Significance Analysis

Statistical significance indicates the probability depending on a null hypothesis. In practice, as the probability gets higher, the doubt about the null hypothesis occurring gets lower. If the null hypothesis exists, this means that a particular decision rule is not informative and the decision is random (Duda et al., 2001). Typical probability values range between 0.0001 and 1. When it rises above certain level, according to the degree of freedom d_f and level of significance α, one gains more confidence in the original hypothesis. It is widely expected to use confidence level of $\alpha = 0.01$ as acceptance probability value. The hypothesis test has to be such that it poses a reasonable question of how significant the decision is, based on random sampled data from the set $\underline{v}$. The hypothesis test first calculates the chi-square distribution (as shown in Equation 8.1). The null hypothesis is accepted if the probability of the observation holds values above

the ones given in Table 8.1 for the given α value or rejects it otherwise. In these terms, statistical significance is the observation of more extreme statistic compared to a previously computed probability, as the one shown in Tables 7.10 through 7.15. The chi-square distribution is dependent on the degree of freedom. An easy way to find the degree of freedom is by knowing the number of the classes involved. In the case of multi class comparison it means that d_f will be greater than 1. The chi-square distribution is the sum of the squared difference between the observed data *k* and the expected data *e*, divided by the expected data, or:

$$X^2 = \sum_{n=1}^{x} \frac{\left(k_n - e_n\right)^2}{e_n}, \tag{8.1}$$

The value of X^2 signifies the chi-square distribution using $n = (x - 1)$ degrees of freedom, where *x* represents the number of classes used for the test. The statistical significance results of the balanced speaker and text independent test with 100 utterances per emotion, random sequence of speakers on the glottal symmetry from clean speech for four and six emotional classes are shown in Table 8.1.

A sample table with critical values of chi-square with different degrees of freedom and confidence level of the null hypothesis $\alpha = 0.01$ presented by Duda et al. (2001) allows us to accept or reject the null hypothesis. There, the critical value of X^2 for the four emotion case has to be higher than 11.34 or higher when the degrees of freedom is $d_f = 3$, and 15.09 or higher when the degrees of freedom is $d_f = 5$.

Table 8.1: Statistical significance for four and six emotions using GS on 100 utterances of clean speech.

Number of emotions	Degrees of freedom	Chi-square	Statistical Significance
4	3	31.62	> 99.99
6	5	72.75	> 99.99

The results from Table 8.1 show that the statistical significance is very high thus the null hypothesis is rejected. This proves that the results shown in Section 7 for the balanced glottal symmetry tests have higher than 99.99% significance and did not result from a random decision. Therefore, we have high confidence that the obtained results are a true representation of the tasks investigated and the performance of the developed systems.

CHAPTER 9

Conclusions

9.1 Contributions of This Dissertation

Analyzing on the data collected from the experiments described in Chapter 7 and considering the goals set for this work, it can be concluded with great confidence that glottal symmetry contains rich emotional content and thus can be effectively used for the task of spoken emotion recognition. It was further shown that glottal information performs substantially better than classical prosodic features and is more robust to noisy conditions than all other features analyzed. The low frequency nature of the glottal signal supports its ability to survive severe lowpass filtering conditions as well. A constraint on the system could be the extraction of the instants of the glottal closures and openings, which was resolved by employing an appropriate group delay–based procedure. This was especially important when the system was tested for robustness in noisy conditions. Furthermore it was shown that even though the mel frequency cepstral coefficients (MFCC) did not perform as well in corpus 3 as compared to corpus 2 in clean speech, the MFCCs show better robustness than classical features in noisy conditions down to SNR =

10 dB. Finally, it was shown that the Tonal and Break Indices (ToBI) feature domain can help improve performance. The two separate applications, ToBI and classical prosodic feature domain were further pruned to their most effective features to obtain better classification rates. This was shown in two occasions in corpus 1 and corpus 3 for both four and six emotion recognition tasks, respectively. The computed statistical significance confirmed the confidence in the obtained results.

9.2 Future Directions

Better automatic detection of accented syllables may lead to further improvement of any later study. There may also an improvement in the ToBI feature extraction and possibly including the break tier as part of feature efforts. This may lead to better results when ToBI alone is used for recognition, and the performance of the combined system may increase as a result. Refinement of the overall selected ToBI features may also provide performance improvement.

The separation of utterances in different emotional classes when building the corpora should be verified through listening tests performed by multiple subjects. This will give more clarity of how the listeners relate to a particular dataset. This in turn will help mitigate any bias connected to either speech quality or perceptual cognition that may negatively impact the correct separation of emotions when forming the corpora.

For more practical applications, since every emotion may carry different levels of intensity, from *weak to strong,* the degree of expression each emotion may be tested. When an utterance is evaluated only a small fragment is taken for analysis. It therefore may not hold the key information of the emotion sought. This in turn requires the

implementation of a different approach in search of keywords and phrases. For on-line systems the temporal variations of features may be studied using Hidden Markov Models (HMM).

Finally, fusion of acoustic and linguistic features may also be considered for achieving more robust emotion recognition.

Bibliography

Books:

Beckman M. and Elam G., 1997. Guidelines for ToBI labeling. *The Ohio State University Research Foundation.*

Cormen T., Leiserson C., and Rivest R., 1990. Introduction to Algorithms. *MIT*.

Cover T. and Thomas J., 1991. Elements of information theory. *John Wiley & Sons, Inc.*

Duda R. and Hart P., 1973. Pattern Classification and Scene Analysis. *John Wiley, New York.*

Duda R., Hart P., and Stork D., 2001. Pattern Classification, 2nd ed., *John Wiley & Sons.*

Flanagan J., 1972. Speech Analysis, Synthesis and Perception. *Springer-Verlag*.

Hardcastle W. and Laver J., 1999. The Handbook of Phonetic Sciences, *Blackwell Publishers Ltd.*

Haykin S., 1994. Neural Networks: A Comprehensive Foundation, *Prentice Hall.*

O'Shaughnessy D., 2000. Speech Communications – Human and Machine. *IEEE Press*.

Quatieri T., 2002. Discrete-Time Speech Signal Processing Principles and Practice, *Prentice Hall*.

Rabiner L. and Schafer R., 1978. Digital Processing of Speech Signals, *Prentice Hall.*

Titze I., 1994. Principles of Voice Production, *Prentice Hall*.

Witten I. and Frank E., 2005. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed., *Morgan Kaufmann.*

Papers and Book Chapters:

Airas M. and Alku P., 2004. Emotions in Short Vowel Segments: Effects of the Glottal Flow as Reflected by the Normalized Amplitude Quotient. Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science). *Affective Dialogue Systems*, Vol. 3068, pp. 13-24.

Alku P. and Vilkman E., 1996. Amplitude Domain Quotient for Characterization of the Glottal Volume Velocity Waveform Estimated by Inverse Filtering. *Speech Communication*, Vol. 18, pp. 131-138.

Allène C., Audibert Y., Couprie M., Cousty J., and Keriven R., 2007. Some Links between Min-cuts, Optimal Spanning Forests and Watersheds. *Mathematical Morphology and its Applications to Image and Signal Processing*, pp. 253-264.

Altun H., Shawe-Taylo J., and Polat G., 2007. New Feature Selection Frameworks In Emotion Recognition To Evaluate The Informative Power Of Speech Related Features. *IEEE Signal Processing and Its Applications, ISSPA 2007 9th International Symposium*, pp. 1-4.

Ananthakrishnan S. and Narayanan S., 2005. Automatic Prosodic Event Detection using Acoustic, Lexical, and Syntactic Evidence. *IEEE Transactions On Speech and Audio Processing – ICASSP*, pp. 216-228.

Battiti R., 1994. Using Mutual Information for Selecting Features in Supervised Neural Net Learning. *IEEE Transactions on Neural networks*, Vol. 5, pp. 537-550.

Berlin N., 1999. Traffic of our stage: Why Waiting for Godot?. *The Massachusetts Review*.

Bhatti M., Wang Y., and Guan L., 2004. A Neural Network Approach for Human Emotion Recognition in Speech. *IEEE International Symposium on Circuits and Systems, Vancouver BC*, pp. 181-184.

Bimbot F., Blomberg M., Boves L., Genoud D., Hutter H., Jaboulet C., Koolwaaij J., Lindberg J. and Pierrot J., 2000. An Overview of the CAVE Project Research Activities in Speaker Verification, *Speech Communication*, Vol.31, pp.155-180.

Black A. and Hunt A., 1996. Generating F_0 Contours from ToBI Labels Using Linear Regression. *Fourth International Conference Proceedings on Spoken Language*, *ICSLP 96*, Vol. 3, pp.1385-1388.

Bosch L., 2003. Emotions, Speech and the ASR Framework. *Speech Communication*, Vol. 40, pp. 213-225.

Boser B., Guyon I., and Vapnik V., 1992. A Training Algorithm for Optimal Margin Classifiers. *5th Workshop on Computational Learning Theory*, pp. 144-152.

Brooks M., Naylor P., and Gudnason J., 2006. A Quantitative Assessment of Group Delay Methods for Identifying Glottal Closures in Voiced Speech. *IEEE Transactions On Audio, Speech and Language Processing*, Vol. 14, No. 2, pp. 456-466.

Chang C. and Li C., 2001. LIBSVM: A Library for Support Vector Machines. *Software available at:* http://www.csie.ntu.edu.tw/~cjlin/libsvm/, Accessed on 09/18/09.

Chuang Z. and Wu C., 2004. Emotion Recognition Using Acoustic Features and Textual Content. *IEEE International Conference on Multimedia and Expo (ICME)*, pp. 53-56.

Chuang Z. and Wu C., 2005. IG-based Feature Extraction and Compensation for Emotion Recognition from Speech. *Proceeding of the First International Conference on Affective Computing and Intelligent Interaction (ACII) 2005*, Beijing, China, October 22-24 2005, Vol. 3784, pp. 358-365.

Collobert R. and Bengio S., 2004. Links Between Perceptrons, MLPs and SVMs. *Proceedings of the 21th International Conference on Machine learning*, Banff, Alberta, Canada, ACM International Conference Proceeding Series, Vol. 69.

Cornelius R., 1996. The Science of Emotion. Research and Tradition in the Psychology of Emotion. *Upper Saddle River, NJ: Prentice-Hall*, pp. 260.

Cowie R. and Cornelius R., 2003. Describing the Emotional States that are Expressed in Speech. *Speech Communication*, Vol. 40, pp. 5-32.

Cummings K. and Clements M., 1995. Analysis of the Glottal Excitation of Emotionally Styled and Stressed Speech. *Journal of the Acoustical Society of America*, July 1995, Vol. 98 (1), pp. 88-98.

Davis S. and Mermelstein P., 1980. Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. Audio and Speech Signal Processing*, Vol. 28, pp. 357-366.

Dempster A., Laird N., and Rubin, D., 1977. Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, Vol. 39, pp. 1-38.

Duan K. and Keerthi S., 2005. Which Is the Best Multiclass SVM Method: An Empirical Study, *Multiple Classifier Systems*, pp. 278-285.

Eckman P., 1992. An Argument for Basic Emotions. *Cognition and Emotion*, Vol. 6 (3/4), pp. 169-200.

Falcão A., Stolfi J., and Lotufo R., 2004. The Image Foresting Transform: Theory, Algorithms, and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 1, pp. 19-29.

Fant G., 1986. Glottal Flow: Models and Interaction. *Journal of Phonetics*, Vol. 14, pp. 393-399.

Fant G., 1968. Analysis and Synthesis of Speech Processes. *In B. Malmberg (Ed.), Manual of phonetics*, *Amsterdam: North-Holland*, pp. 173-177.

Galarniotis A., Tsakoumis A., Fessas P., Vladov S., and Mladenov V., 2003. Using Elman and FIR Neural Networks for Short Term Electric Load Forecasting. *Signals, Circuits and Systems, SCS-IEEE*, Vol. 2, pp. 433-436.

Go H., Kwak K., Lee D., and Chun M., 2003. Emotion Recognition from the Facial Image and Speech Signal. *SICE Annual Conference in Fukui*, August 4-6 2003, pp. 2890-2895.

Gobl C. and Chasaide A., 2003. The Role of Voice Quality in Communicating Emotion, Mood and Attitude. *Speech Communication*, Vol. 40, pp. 189-212.

Hongwei H., Cheng-Lin L., and Hiroshi S., 2003. Comparison of Genetic Algorithm and Sequential Search Methods for Classifier Subset Selection. *Proceedings of the Seventh International Conference on Document Analysis and Recognition ICDAR-IEEE, Computer Society*, August 3-6 2003, pp. 765-769.

Howard D., Lindsay, G., and Allen B., 1990. Toward a Quantification of Vocal Efficiency. *Journal of Voice*, Vol. 4(3), pp. 205-212.

Hung X., Quénot G., and Castelli E., 2004. Recognizing Emotions for the Audio-Visual Document Indexing. *9th IEEE Symposium on Computers and Communications (ISCC'04), Alexandria, Egypt,* June 28 - July 1 2004, Vol. 2, pp. 580-584.

Iida A., Campbell N., Higuchi F., and Yasumara M., 2003. A Corpus-Based Speech Synthesis System with Emotion. *Speech Communication*, Vol. 40, pp. 161-18.

Iida, A., Campbell, N., and Yasumura, M., 1998. Emotional Speech as an Effective Interface for people with Special Needs. *Proceedings of the Third Asian Pacific Computer and Human Interaction, IEEE Computer Society*, July 15-17 1998, pp. 266-271.

Iliev A., Zhang Y., and Scordilis M., 2007. Spoken Emotion Classification Using ToBI Features and GMM, *IEEE 6th EURASIP Conference Focused on Speech and Image Processing*, pp. 495-498.

Jain A., Duin R., and Mao J., 2000. Statistical Pattern Recognition: A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 1, pp. 4-37.

Jiang D. and Cai L., 2004. Speech Emotion Classification with the Combination of Statistic Features and Temporal Features. *IEEE International Conference on Multimedia and Expo (ICME)*, June 27-30 2004, Vol. 3, pp. 1967-1970.

Jilka M., Moler G., and Dogil G., 1999. Rules for the generation of ToBI-based American English intonation. *Speech Communication*, Vol. 28, pp. 83-108.

Kang B., Han C., Lee, S., Youn, D., and Lee, C., 2000. Speaker Dependent Emotion Recognition Using Speech Signals. *Proceedings of the Sixth International Conference on Spoken Language Processing (ICSLP)*. Beijing, China, October 16-20 2000, Vol. 2, pp.383-386.

Klasmeyer G. and Sendlneier W., 1995. Objective voice parameters to characterize the emotional content in speech. *Proceedings ICPhS 95*, pp. 1,182.

Kounoudes A., Naylor P., and Brookes M., 2002. The DYPSA Algorithm for Estimation of Glottal Closure Instants in Voiced Speech. *Proceedings ICASSP'02 Orlando*, Vol. 1, pp. 349–352.

Kwon O., Chan K., Hao J., and Lee T., 2003. Emotion Recognition by Speech Signals. *Eurospeech – Geneva*, pp. 125-128.

Laukkanen A., Vilkman E., Alku P., and Oksanen H., 1996. Physical Variation Related to Stress and Emotional State: A Preliminary Study. *Journal of Phonetics*, Vol. 24, pp. 313-335.

Lee C. and Narayanan S., 2005. Toward Detecting Emotions in Spoken Dialogs. *IEEE Transactions on Speech and Audio Processing*, March 2005, Vol. 13, No. 2, pp. 293-303.

Lin Y. and Wei G., 2005. Speech Emotion Recognition Based on HMM and SVM. *Proceedings of the Fourth International Conference on Machine Learning and Cybernetics, Guangzhou,* August 18-21 2005, Vol. 8, pp. 4898-4901.

Luengo I., Navas, E., Hernáez I., and Sánchez, J., 2005. Automatic Emotion Recognition Using Prosodic Parameters. *INTERSPEECH-2005*, pp. 493-496.

Markel J., 1972. The SIFT Algorithm for Fundamental Frequency Estimation. *IEEE Transactions on Audio and Electroacoustics*, December 1972, Vol. 20, Issue 5, pp. 367-377

Miranda P., Falcão A., Rocha A., and Bergo F., 2008. Object Delineation by *k* - Connected Components. *EURASIP Journal on Advances in Signal Processing*, pp. 1-14.

Mokhtari P. and Campbell N., 2003. Automatic Measurement of Pressed/Breathy Phonation at Acoustic Centers of Reliability in Continuous Speech. *IEICE Transactions on Information and Systems*, March 2003, Vol. E86-D, No. 3, pp. 574-582.

Moore E., Clements M., Peifer J., and Weisser L., 2003. Investigating the Role of Glottal Features in Classifying Clinical Depression. *25th Annual International Conference of the IEEE EMBAS*, pp. 2849-2852.

Moore E., Clements M., Peifer J., and Weisser, L., 2008. *IEEE Transactions on Biomedical Engineering*, January 2008, Vol. 55, No. 1, pp. 96-107.

Naylor P., Kounoudes A., Gudnason J., and Brookes M., 2007. Estimation of Glottal Closure Instants in Voiced Speech Using the DYPSA Algorithm. *IEEE Transactions on Speech and Audio Processing*, Vol. 15, pp. 34–43.

Neiberg D., Elenius K., Karlsson I., and Laskowski K., 2006. *Emotion Recognition in Spontaneous Speech, FONETIK*, http://www.ling.lu.se/conference/fonetik2006/program.html, Accessed on 05/24/08.

Nicholson J., Takahashi K., and Nakatsu, R., 2000. Emotion Recognition in Speech Using Neural Networks. *Neural Computing and Applications*, Vol. 9, pp. 290-296.

Nissen S., 2003. Implementation of a Fast Artificial Neural Network Library (FANN), Department of Computer Science University of Copenhagen (DIKU). *Software available at:* http://leenissen.dk/fann/, Accessed on 08/12/06.

Noda T., Yano Y., Doki S., and Okuma S., 2006. Adaptive Emotion Recognition in Speech by Feature Selection Based on KL-divergence. *IEEE International Conference on Systems, Man, and Cybernetics in Taipei, Taiwan*, October 8-11 2006, pp. 1921-1926.

Papa J., Falcão A., Suzuki C., and Mascarenhas N., 2008. A Discrete Approach for Supervised Pattern Recognition. *12th International Workshop on Combinatorial Image Analysis*, Vol. 4958, pp. 136-147.

Papa J., Suzuki C., and Falcão A., 2008. LibOPF: A Library for the Design of Optimum-Path Forest Classifiers. *Software version 1.0 available at:* http://www.ic.unicamp.br/~afalcao/libopf/index.html, Accessed on 10/15/09.

Peterson G. and Barney H., 1952. Control Methods Used in a Study of Vowels. *Journal of the Acoustical Society of America*, Vol. 24, pp. 175-184.

Petrushin V., 2000. Emotion Recognition in Speech Signal: Experimental Study, Development, and Application. *Proceedings of the Sixth International Conference on Spoken Language ICSLP'00, Beijing, China*.

Razak A., Komaiya, R., and Abidin, M., 2005. Comparison Between Fuzzy and NN Method for Speech Emotion Recognition. *IEEE Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05)*, July 4-7 2005, Vol. 1, pp. 297-302.

Reunanen J., 2003. Overfitting in Making Comparisons Between Variable Selection Methods, *Journal of Machine Learning Research*, Vol. 3, pp. 1371-1382.

Roach P., 2000. Techniques for the Phonetic Description of Emotional Speech. *Proceedings of the ISCA Workshop on Speech and Emotion, Northern Ireland*, pp. 53-59.

Rosenberg A., 1971. Effect of Glottal Pulse Shape on the Quality of Natural Vowels. *Journal of the Acoustical Society of America*, Vol. 49, pp. 583-590.

Rothenberg M., 1973. A New Inverse-Filtering Technique for Deriving the Glottal Air Flow Waveform During Voicing. *Journal of the Acoustical Society of America*, Vol. 53, pp. l632-1645.

Scherer K., 2003. Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, Vol. 40, pp. 227-256.

Schuller B., Rigoll G., and Lang M., 2003. Hidden Markov Model-based speech emotion recognition. *ICME'2003 Multimedia and Expo Proceedings*, July 6-9 2003, Vol. 1, pp. 401-404.

Srinvas M. and Patnaik L., 1994. Genetic Algorithms: A Survey. *Computer*, June 1994, pp.17-26.

Stibbard, R., 2000. Automated Extraction of ToBI Annotation Data from the Reading/Leeds Emotional Speech Corpus. *ITRW on Speech and Emotion, ISCA, in Speech Emotion*, pp. 60-65.

Ververidis D. and Kotropoulos C., 2006. Emotional Speech Recognition: Resources, Features, and Methods. *Speech Communication*, Vol. 48, pp. 1162-1181.

Wang Y. and Guan L., 2004. An Investigation of Speech-Based Human Emotion Recognition. *IEEE 6th Workshop on Multimedia Signal Processing*, pp. 15-18.

Wong D., Markel J., and Gray A, 1979. Least Squares Glottal Inverse Filtering from the Acoustical Speech Waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-27, No. 4, pp. 350-355.

Zhang Y. and Scordilis M., 2004, Optimization of GMM Training for Speaker Verification, *Proceedings of ODYSSEY 2004, the IEEE Speaker and Language Recognition Workshop, Toledo, Spain*, May 31 – June 4, 2004.

Zhen-Hua L., Yu H., and Ren-Hua W., 2005. A Novel Source Analysis Method by Matching Spectral Characters of LF Model with STRAIGHT Spectrum. *Springer-Verlag Berlin Heidelberg*, pp. 441-448.

Zhou G., Hansen H., and Kaiser J., 1999. Methods for Stress Classification: Nonlinear TEO and Linear Speech Based Features. *Acoustics, Speech, and Signal Processing ICASSP'99 Proceedings*, pp. 2087-2090.

Zhou G., Hansen H., and Kaiser J., 2001. Nonlinear Feature Based Classification of Speech under Stress. *IEEE Transactions On Speech And Audio Processing*, Vol. 9, No. 3, pp. 201-216.

Appendix A: Work published by Alexander I. Iliev

Articles:

Iliev, A.I., Scordilis, M.S., Papa J.P., Falcão A.X., "Spoken emotion recognition through optimum-path forest classification using glottal features", Computer Speech and Language, Special Issue, ELSEVIER, 2009.

Iliev, A.I., Scordilis, M.S., "Emotion Recognition in Speech using Inter-Sentence Glottal Statistics", Proceedings of the 15th International Conference on systems, Signals and Image Processing, IEEE-IWSSIP 2008, Bratislava, Slovakia, June 25-28, 2008, pp. 465-468.

Iliev, A.I., Zhang, Y., Scordilis, M.S., "Spoken Emotion Classification Using ToBI Features and GMM", Proceedings of the 14th International Workshop on Signals and Image Processing 2007 and the 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services. IEEE-IWSSIP 2007, Maribor, Slovenia, June 27-30, 2007, pp. 495-498.

Iliev, A.I., He, X., Scordilis, M.S., "A High Capacity Watermarking Technique for Stereo Audio", Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP2004, Montreal, Canada, May 2004, vol. 5, pp. 393-396.

Iliev, A.I., Scordilis, M.S., "Multilevel High Capacity Data Hiding Technique for Stereo Audio", Proceedings of the 2004 IEEE-Asilomar Conference, Pacific Grove, California, November 7-10, 2004, pp. 1793-1797.

Jin, W., Scordilis, M.S., Iliev., A.I., "Comparison and implementation of a 16-bit fixed point audio resampler", Proceedings of the 2004 IEEE-Asilomar Conference, Pacific Grove, California, November 7-10, 2004, pp. 1798-1800.

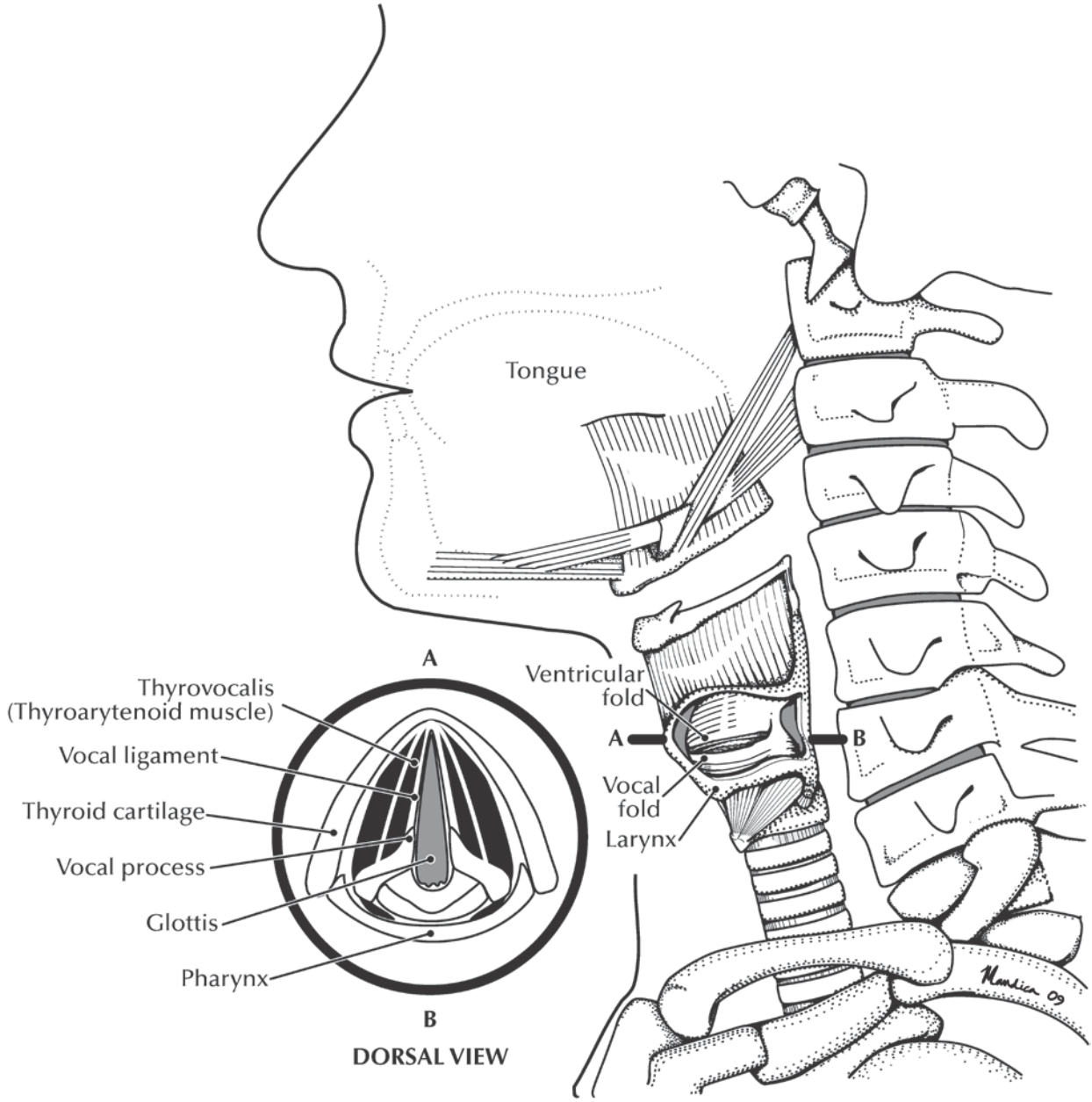
Iliev, A.I., Scordilis, M.S., "Binaural Phase Masking Experiments in Stereo Audio", Proceedings of the First Pan-American/Iberian Meeting on Acoustics, ASA, Cancun, Mexico, December 2-6, 2002 (abstract).

Patents:

"Coding a Masked Data Channel in a Radio Signal" Inventors: A.I. Iliev, M. S. Scordilis and H. Leventhal, U.S. Patent No. 7,079,633 B2, awarded on July 18, 2006.

"Auxiliary channel masking in an audio signal", U.S. Patent No. 6,996,521 B2, Inventors: A.I. Iliev and M. S. Scordilis, Assignee: University of Miami, awarded on February 07, 2006.

Appendix B: Anatomy of the Larynx



Appendix C: Sentences used for creating *Corpus 2*:

All utterances were interpreted for the four emotional states: *Angry*, *Happy*, *Sad*, and *Neutral*.

A. Short sentences:

   a. Me too.

   b. Over there.

   c. In a ditch.

   d. There you are.

   e. Among the first.

B. Long sentences:

   a. Never neglect the little things of life.

   b. We were respectable in those days.

   c. That's where we'll go for our honeymoon.

   d. Hand in hand from the top of the Eiffel Tower.

   e. He wants to know if it hurts.

VITA

Alexander Iliev Iliev was born in Sofia, Bulgaria, on May 22, 1972. His parents are Iliya Mihov Iliev and Raina Alexandrova Ilieva. He received his elementary education at “Valentin Andreev” School and his secondary education at the High School of Professional Audio, Video and Telecommunications "Alexandar Stepanovich Popov" both in Sofia, Bulgaria. In September 1991 he entered the “St. Ivan Rilski” University in the same city, from which he graduated with the MS degree in Electrical Engineering in June 1996. During the fall of 1993, he attended the “Technical University” in Sofia, second majoring in Multimedia Technologies and graduating the program in June 1996. In January 1998 he was admitted to the Graduate School of the University of Miami, where he was granted the MS degree in Music Engineering Technologies in December 1999. In January 2004 he began his work in the Ph.D. program at the Electrical and Computer Engineering department in the same university and was granted the degree of Doctor of Philosophy in Electrical and Computer Engineering in December 2009.