PHYSIOLOGICALLY-MOTIVATED FEATURE EXTRACTION METHODS FOR SPEAKER RECOGNITION

By
Jianglin Wang, B.S., M.S.

A Dissertation Submitted to the Faculty of the
Graduate School, Marquette University,
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy

Milwaukee, Wisconsin
December 2013

ABSTRACT

PHYSIOLOGICALLY-MOTIVATED FEATURE EXTRACTION METHODS FOR SPEAKER RECOGNITION

Jianglin Wang, B.S., M.S.
Marquette University, 2013

Speaker recognition has received a great deal of attention from the speech community, and significant gains in robustness and accuracy have been obtained over the past decade. However, the features used for identification are still primarily representations of overall spectral characteristics, and thus the models are primarily phonetic in nature, differentiating speakers based on overall pronunciation patterns. This creates difficulties in terms of the amount of enrollment data and complexity of the models required to cover the phonetic space, especially in tasks such as identification where enrollment and testing data may not have similar phonetic coverage. This dissertation introduces new features based on vocal source characteristics intended to capture physiological information related to the laryngeal excitation energy of a speaker. These features, including RPCC, GLFCC and TPCC, represent the unique characteristics of speech production not represented in current state-of-the-art speaker identification systems. The proposed features are evaluated through three experimental paradigms including cross-lingual speaker identification, cross song-type avian speaker identification and mono-lingual speaker identification. The experimental results show that the proposed features provide information about speaker characteristics that is significantly different in nature from the phonetically-focused information present in traditional spectral features. The incorporation of the proposed glottal source features offers significant overall improvement to the robustness and accuracy of speaker identification tasks.

ACKNOWLEDGEMENTS

Jianglin Wang, B.S., M.S.

Many individuals have contributed directly or indirectly to the success of this research. I wish to thank them for their patience, effort and their energy. I especially thank my dissertation supervisor at Speech and Signal Processing Laboratory of Marquette University, Dr. Michael Johnson. For almost six years now, Dr. Johnson has been a constant source of ideas and inspirations. Thank you for pushing me through my periods of pessimism and for helping me think like a scientist and innovate like an engineer. I appreciate all your kind support and suggestion on my life and research.

I also would like to express my gratitude to Dr. Richard Povinelli, Dr. James Richie, Dr. Jeffrey Berry and Dr. Frederick Frigo for serving on my committee and their support and helpful reviews. I would like to thank all my friends at the laboratory for their help and patience during the course of my work.

I thank my parents and young sister for their many years of support and their love throughout my life. Most of all I thank my wonderful wife Peng, who always accompanies with me in the whole journey.

Jianglin Wang

December, 2013

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Problem Definition and Motivation

The task of speaker identification and verification has received a great deal of attention from the research community in the past decade, and there have been substantial gains in accuracy as well as in channel and background robustness [1, 2]. However, the fundamental mechanism of state-of-the-art systems has remained phonetic rather than physiological in nature, and little progress has been made toward identifying individually unique speech characteristics that are independent of phonetic content. Automatic speaker recognition systems have mainly been explored in a single language or mono-lingual environment, where phonetic content is relatively consistent and such approaches are sufficient. Performance degrades for tasks such as cross-lingual speaker recognition, where there are multiple languages, both training and testing, since the phonetic features used in mono-lingual tasks are less useful in a cross-lingual or multi-lingual environment. This also limits application of speaker identification to other similar tasks, such as bioacoustic censusing, where individual identification is used to count and monitor animal populations within a species [3, 4]. In addition, the phonetic nature of traditional approaches requires larger amount of training data

than might otherwise be ready. Given these limitations, there is still a significant need for identification of unique non-phonetic and speaker-specific features that are more independent of phonetic content and representative of more underlying physiological characteristics of the speaker.

The most popular approach for current recognition systems is based on Gaussian Mixture Models (GMM) [5] or GMMs coupled with Support Vector Machines (GMM-SVM) [6, 7]. Verification is accomplished through likelihood comparison with appropriate cohort models, as with Universal Background Model (UBM) systems [8]. For the GMM and SVM classifiers, a number of techniques developed within the speaker recognition area have shown excellent performance. For example, many generative models, such as Eigenchannels, Eigenvoices and Joint Factor Analysis (JFA), have built on the success of the GMM-UBM approach. Inspired by the success of joint factor analysis in speaker recognition [9], Dehak proposed the i-vector approach which is to find a low dimensional subspace of the GMM supervector space, called the total variability space that represents both speaker and channel variability. Standard spectral features such as Mel Frequency Cepstral Coefficients (MFCCs) with energy and first and second order derivatives are used, typically projected to a lower dimension through Heteroscedastic Linear Discriminant Analysis (HLDA) [10]. A variety of feature transformation techniques are used, including Nuisance Attribute Projection and Latent Variable Analysis [11, 12]. Score normalization techniques such as TNorm

and ZNorm [13] are also incorporated. In the past few years it has become commonplace to build a large number of diverse system types and integrate them through various forms of intelligent score fusion. This is seen for example in the NIST speaker recognition challenges [14, 15], in which the focus is on environmental and channel variability. In all cases, though, the basic features remain primarily spectral in nature, relating to vocal tract parameters that are heavily correlated with phonetics.

Environmental, recording, and channel mismatches have also received a great deal of attention from the research community in the past decade, while spoken language mismatch has been neglected. In [16], authors presented that language mismatching scenarios between a target speaker and its world model cause critical degradations in the accuracy of speaker verification system, especially for Vietnamese and Mandarin against the native English target speakers. The work by [17] reported that the speaker verification system shows slight performance discrepancies if the model was trained on English and Chinese digits/sentences, and tested on English versus testing on Chinese. Research by Geoffrey and Faundez [18, 19] shows that there is a decreased accuracy in language-mismatched conditions, but they give no further insight on how to alleviate the problem. Other research by Hansen and Meng [20, 21] gives an approach to this problem, but it is necessary to know the possible languages that will be used and build a corresponding speaker model for each of them, and this is not always possible. In [22, 23], a multilingual phonetic string

approach, which is similar to the phone recognition and language modeling method used for language identification, is applied to speaker recognition. The advantage of these methods can provide language independence, but extensive training and testing materials are required, because speaker recognition systems with a small amount of data are not sufficient to eliminate the impact of language variation. In [24], Campbell et al. proposed to create corpora to support and evaluate speaker identification systems on language variation conditions. In [25], the problem of language variability was investigated on forty-nine bilingual speakers using the Spanish/Catalan corpus. The author proposed a method with a language-independent and language-dependent vector quantization codebook to address the problem of language variability. In [26], a series of listener tests was conducted to demonstrate that language familiarity plays a significant role in speaker identification. In [27] some short-term prosodic features were proposed to enhance the language robustness in speaker recognition systems, although the results are still far from being truly language independent.

In addition to these phonetic model and language mismatch issues, the significant amount of speech required for speaker model training has also limited the widespread use of speaker recognition technology in everyday applications. Reducing the amount of the required speech utterance to train speaker model has been investigated in a number of recent papers focused on Joint Factor Analysis (JFA) and SVM based

speaker recognition. These studies have shown that performance dramatically degrades with short utterances ($< 10s$) for both JFA and SVM approaches [28, 29]. Kanagasundaram found that for short duration ($< 2s$) text-independent speaker verification, i-vector approaches obtain comparable performance to the classical JFA approach but don't provide noticeable improvement [30]. Therefore, tasks with limited duration training data still represent a serious issue for text-independent speaker recognition. This is connected to the phonetic nature of speaker recognition models, because one of the reasons that so much training data is required is that the speech needs to sufficiently cover the entire phonetic space in order to build robust models.

From the above studies, it is clear that mismatched conditions and limited training data are not a well solved problem in speaker recognition. In the work introduced here, novel feature extraction techniques are used to create physiologically motivated measures to improve system robustness and accuracy across a wide range of such task configurations.

1.2 Purpose

The objective of this dissertation is to create novel feature extraction methods to capture speaker-discriminative features from speech utterances, applicable across a wide range of speaker recognition tasks. A number of new features are introduced and investigated, including some based on vocal excitation as opposed to vocal tract

information and some based on temporal variabilities within vocal tract characteristics represented. Following preliminary experimental investigations, three new vocal source features, Residual Phase Cepstral Coefficients (RPCC), Teager Phase Cepstral Coefficients (TPCC), and Glottal Flow Cepstral Coefficients (GLFCC) are selected for continued investigation and evaluation.

The experimental work in this dissertation is designed to test the potential influence of phonetic differences and evaluate whether the proposed features relate to physiological rather than phonetic distinctions. A Gaussian mixture model and universal background model (GMM-UBM) is used to create a baseline speaker recognition framework for this evaluation.

Three experimental paradigms have been introduced for evaluating phonetically-independent features for speaker identification. These tasks, bilingual speaker identification and cross song-type avian speaker identification, are each designed to assess the performance of the proposed features in conditions with mismatched training and test data. The third task, speaker identification on YOHO corpus, is to verify if the proposed features are also significant for single language condition. The first of these has multiple applications, such as forensic speaker recognition, which needs to be robust across language. The second of these, cross song-type bird recognition, addresses a key problem in bioacoustic censusing. Current methods for this task require prior separation by vocalization category, which severely limits its capability for many

species. The third task demonstrates application of the newly proposed features to text-dependent speaker recognition evaluation for a broad range of applications.

1.3 Dissertation Overview

The first chapter gives a brief overview of the dissertation and the motivation behind the research. The second chapter discusses background knowledge and related work in the field of speaker recognition. The third chapter reviews the state-of-the-art in features for speaker recognition and introduces the newly proposed feature extraction methods.

The fourth chapter details the experiments and results of implementing the new proposed features for cross-lingual speaker identification, cross song-type avian speaker identification and mono-lingual speaker identification. Recognition results based on current state-of-the-art baseline methods are also presented for comparison.

The fifth chapter summarizes the major work of this dissertation, highlights the contribution of the research, and proposes several different possible future research avenues in this area.

CHAPTER 2
SPEAKER RECOGNITION SYSTEMS

2.1 Overview

Speaker recognition is highly effective for human-computer interaction, identification and information retrieval. The required terminal equipment is simple, e.g., a good performance microphone. Current state-of-the-art recognition rates are very high; in some cases, automatic recognition performance is better than human. Speaker recognition can be applied to the use of computers and computer networks, access control of public security, voice authentication in telephone transactions, telephone monitoring in the military, access to network resources, and many other emerging applications. With the rapid development of Internet and network resources, speaker recognition can be widely applied for voice search in digital library and information retrieval.

2.2 Historical Achievements

Research in automatic speaker recognition has attracted a great deal of attention for the past five decades and is still considered an important area for speech signal

processing. The development of speaker recognition technology parallels the advancement of speech recognition technology.

Research for speaker recognition began in the 1960s, one decade later than that for speech recognition. Kersta at Bell Labs stimulated the research on speaker recognition by introducing the sound spectrograph from the Sonograph [31]. The spectrograms for different speaker could be easily distinguished by specially trained experts, however, the computer could not directly use the spectrogram for automatic speaker recognition.

In the mid 1960s, digital signal processing technology was employed in the area of speech recognition. Its significance was that the task of speech and speaker recognition could be implemented directly by computer, rather than requiring a trained expert. In 1969, Luck first implemented the Cepstrum for speaker recognition [32]. In the 1970s, Atal and Itakura proposed to apply the theory of linear predictive coding for speech signal processing [33, 34]. BS Atal later investigated the Linear Predictive Cepstrum Coefficients (LPCC) for speaker recognition in order to improve the recognition accuracy of the system.

In the 1980s, Steven B. Davis proposed mel-frequency cepstral coefficients (MFCC) [35]. MFCC features take into account human auditory mechanisms to obtain better recognition performance and noise robustness. This method is now widely used in

almost all speech and speaker recognition systems. At the same time, artificial neural networks (ANN) and hidden markov models (HMM) were successfully applied in speech recognition and became the core of speaker recognition methods [36, 37].

In the 1990s, Reynolds presented the Gaussian Mixture Model (GMM)-based speaker verification system used successfully in several NIST speaker recognition evaluations [5, 38]. As conducted by NIST in recent years, each speaker recognition evaluation on conversational telephone speech has involved a corpus with hundreds of speakers, thousands of conversation sides, and tens of thousands of individual test trails. Each evaluation test set is dependent on numerous data collection factors that affect evaluation performance [39]. GMMs promptly became the mainstream technology of the text-independent speaker recognition because it is simple, flexible, effective and robust. GMMs can be regarded as a single-state HMM which ignores the temporal pattern of speech. The state-of-the-art GMM system rapidly became the main stream technology for the text-independent speaker recognition due to its simplicity, flexibility and robustness, and further promoted the rapid development of speaker recognition. In 2000, Reynolds presented a GMM-based speaker verification system using a universal background model (UBM) for alternative speaker representation, and maximum a posterior (MAP) adaption method to derive speaker models from the UBM [40, 41]. The method of UBM-MAP made a significant contribution to speaker recognition, taking it from the laboratory to practicality. UBM-MAP reduces

the dependence on the training set for the statistical model of GMM and enhances the robustness of mismatch conditions for training and testing. In UBM-MAP, only small amounts of adaptive speech are needed for the training of speaker model.

Currently, various speaker recognition technologies are emerging, such as the application of large vocabulary continuous speech recognition (LVCSR) to text-independent speaker recognition [42], hybrid GMM and support vector machine (SVM) approaches to speaker recognition [7, 43], using prosodic and lexical information for speaker recognition [44], score normalization of TNorm and ZNorm for text-independent speaker recognition [13, 45] and latent factor analysis (LFA) for speaker recognition [46, 47]. However, the best text-independent speaker recognition systems are still based on GMM, in particular, the structure of UBM-MAP [48].

The recent advances in the domain of intersession variability compensations have been a breakthrough in improvements in error rates. One of these techniques is known as an i-vector extraction, which consists of mapping a sequence of vectors for a given speech utterance to a low-dimensional vector space [49–51], referred to as the total variability space based on a factor analysis technique. The *i*-vector extraction can be viewed as a probabilistic compression process in order to represent speech segments variability in a low dimensionality space. Hence, i-vectors convey speaker characteristic along with other information, such as acoustic environment, transmission channel or phonetic content of the speech segment.

2.3 Speaker Recognition Tasks

Speaker recognition doesn't focus on the semantic content of the speech signal, but extracts characteristics of a speaker's speech pattern to recognize the speaker identity. Speaker recognition not only seeks to represent the personal characteristics of a speaker, but also to emphasize the differences between individuals. In contrast, speech recognition and language recognition emphasize normalization of speaker variance. The fundamental goal of speaker recognition is to extract the speaker's characteristics from the speech waveform and use the extracted information for speaker identification or verification.

Speaker recognition systems can be classified into text-dependent and text-independent tasks based on the characteristics of the spoken text. A text-dependent recognition system is trained to recognize specific speakers from pre-defined utterances which can be either a fixed phrase or a randomly prompted phrase. The text-independent recognition system doesn't know the content of the speech utterance, which can be any words, phrases or sentences. In general, text-independent speaker recognition is more flexible and applicable, but more difficult to achieve, requiring more training and testing speech data. In contrast, text-dependent speaker recognition can achieve higher recognition accuracy with less training and testing data.

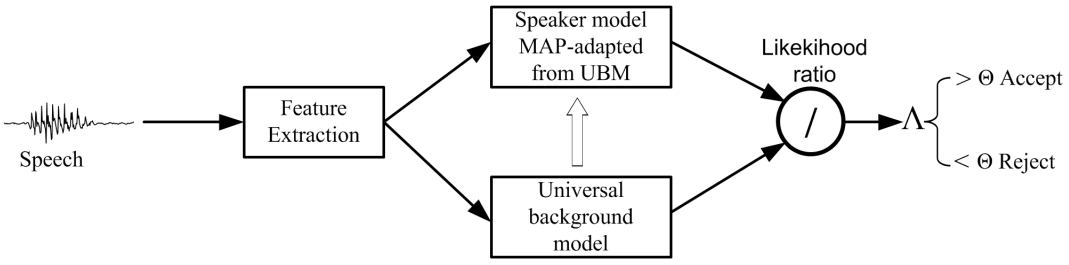


Figure 2.1: The structure of speaker verification.

Speaker recognition systems can also be divided into identification and verification tasks. Speaker verification is a binary classification task, determining whether the unlabeled input utterance belongs to a claimed speaker or not. Thus a speaker verification system either accepts the utterance as the claimed speaker or rejects it as an impostor, as shown in Figure 2.1. In contrast, the goal of speaker identification is to identify the unlabeled input utterance as belonging to one of a set of known speakers. It is an M-any classification task where the voice is compared against M speaker models, as shown in Figure 2.2. Speaker identification is usually more difficult than verification, since the number of decision alternatives of speaker identification is equivalent to the size of the enrolled speaker set M, and the speaker identification performance decreases as the size of the speaker increases.

In general, a speaker identification and verification system consists of three main components, including front-end processing, speaker modeling, and decision-making. The first step, whether for verification or identification, is front-end processing in

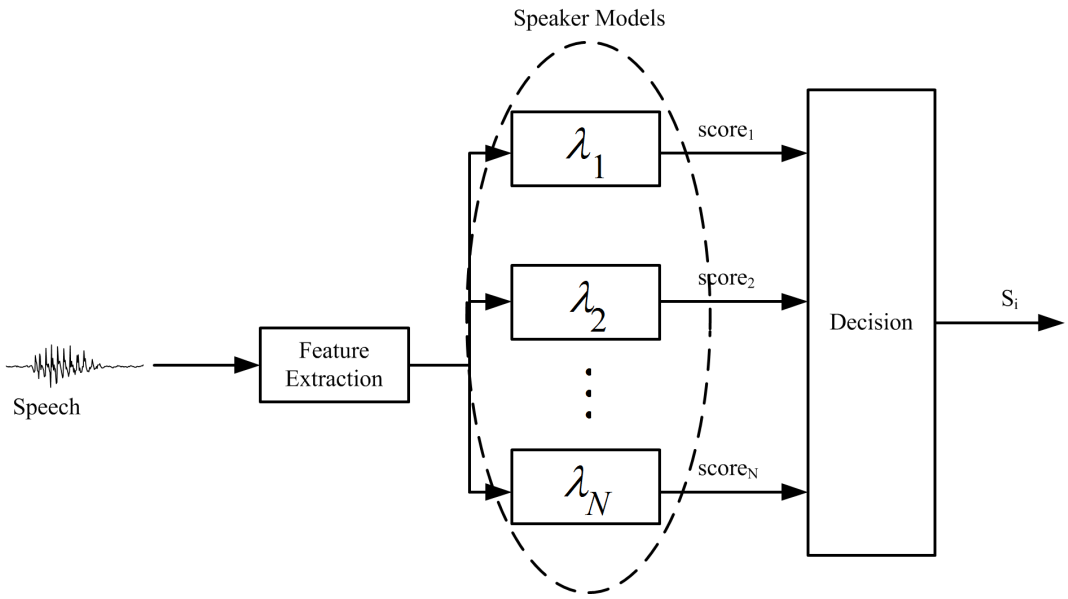


Figure 2.2: The structure of speaker identification.

order to extract from the speech signal features that convey speaker-dependent information. To train the speaker model in text-independent speaker recognition, generally a universal background model (UBM) with a large number of Gaussian mixture components is trained based on a large number of "imposter" speakers. Instead of training a speaker's model directly using the corresponding speaker's utterances, this method adapts a target speaker model from the UBM by maximum a posteriori (MAP) [40]. The difference between speaker verification and speaker identification lies in the stage of decision making. In a speaker verification system, the input speech of a claimed speaker is matched only against the claimed speaker model and universal background model to calculate the likelihood ratio. Then the ratio is compared to a threshold

to accept or reject the decision. In a speaker identification system, the test speech is matched against each speaker's model to calculate the corresponding likelihood scores, selecting the corresponding model using a maximum likelihood criterion.

2.4 Evaluation Metrics

Speaker verification systems need to compare the score of the test speech to the claimed speaker model and verify the result according to certain decision rules. The basic hypotheses are

- H_0: Y is from the hypothesized speaker S
- H_1: Y is not from the hypothesized speaker S.

The decision for these two hypotheses is a likelihood ratio test given by

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta \text{ accept } H_0 \\ < \theta \text{ reject } H_1 \end{cases}, \tag{2.4.1}$$

where $p(Y \mid H_i)$ is the likelihood of the hypothesis H_i given the observed speech segment Y, and θ is the threshold for accepting or rejecting H_0. The basic goal of a speaker verification system is to compute values for the two likelihood, $p(Y \mid H_0)$ and $p(Y \mid H_1)$.

Speaker verification has two types of errors. The false acceptance rate (FA) is the probability that an impostor is wrongly accepted as the true speaker. The false

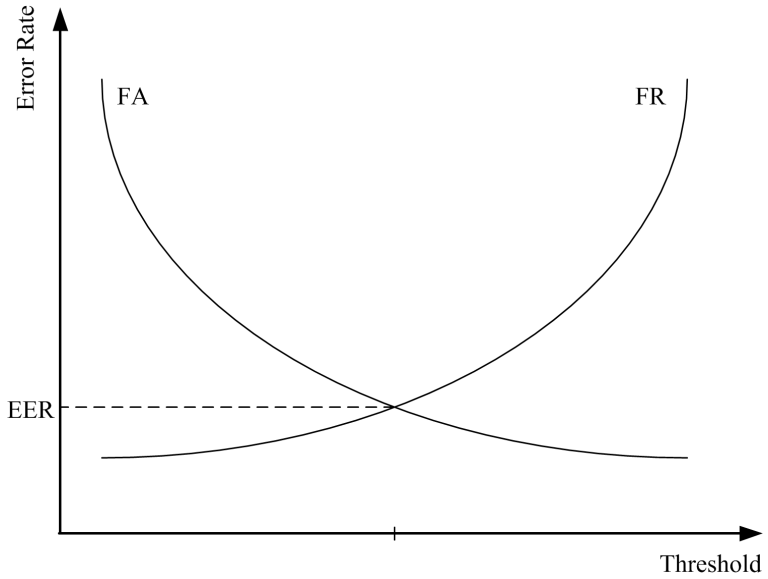


Figure 2.3: The impact of threshold on FA & FR.

rejection rate (FR) is the probability that the true speaker is rejected as an impostor. Different thresholds will generate different FA and FR rates in an inverse relationship. The relationship between the threshold and FA & FR is shown in Figure 2.3. Obviously, a higher threshold will make the system less likely to accept an impostor, but it also makes it more likely to reject the true speaker. In contrast, although a lower threshold will reduce the probability to reject the true speaker, it also increases the chance of accepting an impostor. This tradeoff can be also shown completely in the Receiver Operating Characteristic (ROC) curve, illustrated in Figure 2.4, that shows the FA & FR performance as the threshold is varied.

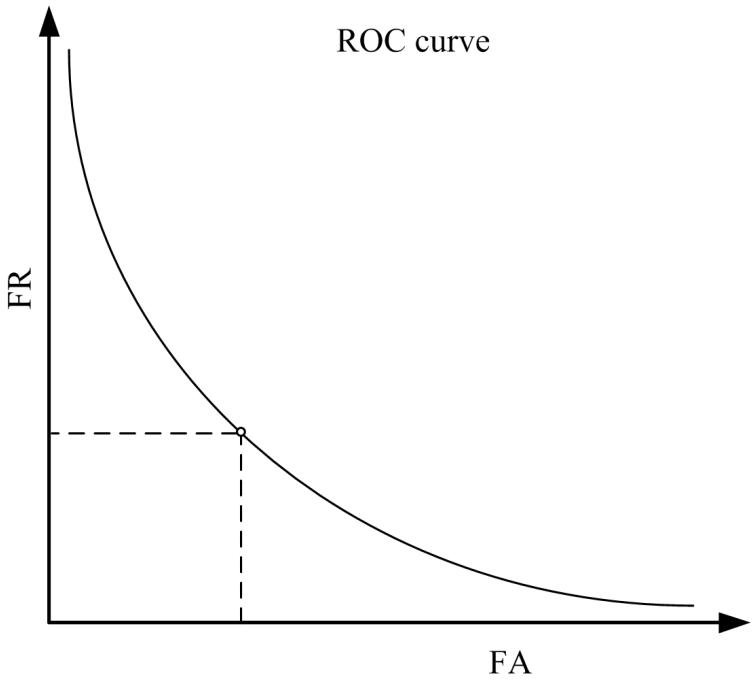


Figure 2.4: The ROC curve with EER identified.

2.4.1 Equal Error Rate (EER)

Figure 2.3 and Figure 2.4 show that FA decreases and FR increases as the threshold increases. The intersection point of the two curves is called the equal error rate, and the corresponding threshold value is called the EER. EER is often used as a speaker recognition system's evaluation criteria. It is simple and intuitive. However, EER is limited in that it doesn't address the tradeoffs between FA and FR costs.

2.4.2 Detection Cost Function (DCF)

To effectively evaluate the system performance, the National Institute of Standards and Technology (NIST) has defined an evaluation function-Detection Cost Function, which is as a weighted sum of FA and FR and computed as follows:

$$DCF = C_{FR} \cdot P_{FR} \cdot P_{Tar} + C_{FA} \cdot P_{FA} \cdot P_{Imp} \,, \tag{2.4.2}$$

where C_{FR} and C_{FA} are the cost of false rejection of a target speaker and false acceptance of an impostor. P_{FR} is the probabilities of false rejection, and P_{FA} is the probabilities of false acceptance. P_{Tar} and P_{Imp} are the prior probabilities of a target speaker and impostor speaker, respectively.

The DCF takes into account the costs of the two types of errors as well as the prior probability of the target speaker and impostor speaker. By incorporating such costs, the DCF is able to more flexibly meet the needs of different applications. The NIST evaluation uses $C_{FR} = 10$, $C_{FA} = 1$, $P_{Tar} = 0.01$, and $P_{Imp} = 0.99$, which emphasizes system security.

2.5 Features for Speaker Recognition

In pattern recognition, one of the first tasks is to extract parameters to represent the feature information. In speaker recognition, the purpose of feature extraction is to

remove redundant and irrelevant information to efficiently capture the characteristics of each speaker. In contrast to speech recognition, where the features need to represent phonetically relevant characteristics of the speech utterance, the features for speaker recognition focus on capturing the distinctive features of an individual to effectively distinguish speakers from each other. However, in practice this is not how current speaker recognition systems work. Current speaker verification and identification systems focus on uniqueness of pronunciation patterns and phonetics rather than physiologically speaker-distinctive vocal characteristics. The dominant features used in speaker recognition are adopted directly from those used for speech recognition, and primarily capture vocal tract spectrum characteristics of speaker. Although this is effective, since pronunciation patterns are unique, it is also somewhat non-intuitive considering the opposite nature of these two tasks.

Speech researchers have been looking for novel features for speaker recognition since the 1950s without much success. Feature extraction is still a bottleneck to improve the accuracy of speaker recognition algorithms. In particular, the physiological significance of the vocal source mechanism in speech production as features for speaker recognition has been significantly under-investigated. This dissertation strives to demonstrate the possible advantages to be gained by effectively combining vocal source with vocal tract features to improve the performance of speaker recognition. Established feature extraction method will be introduced in Section 2.5, while

the newly proposed features will be introduced in Chapter 3.

2.5.1 MFCC

Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used in most speech and speaker recognition systems. These approximate the perceptual model of the human auditory system by warping the linear frequency axis to match the Mel-scale cochlear frequency map [52]. This filter spacing is approximately linear up to 1 kHz and logarithmic at higher frequencies as shown in Figure 2.5. The reason for this warping is that human ears, for frequencies lower than 1 kHz, hear tones with a linear scale instead of logarithmic scale for the frequencies higher than 1 kHz. The warping function in equation (2.5.1) transforms linear frequencies to Mel-frequencies, where Mel is the perceptual frequency and f is the real frequency in Hz.

$$Mel(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{2.5.1}$$

MFCCs are usually derived using a triangular shaped Mel-scale filterbank as illustrated in Figure 2.6. It has been found that the energy in a critical band of a particular frequency influence the human auditory system perception [53]. This critical band bandwidth varies with the frequency, where it is linear below 1 kHz and logarithmic above. Mel filterbanks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and fewer filters in the high frequency regions [54].

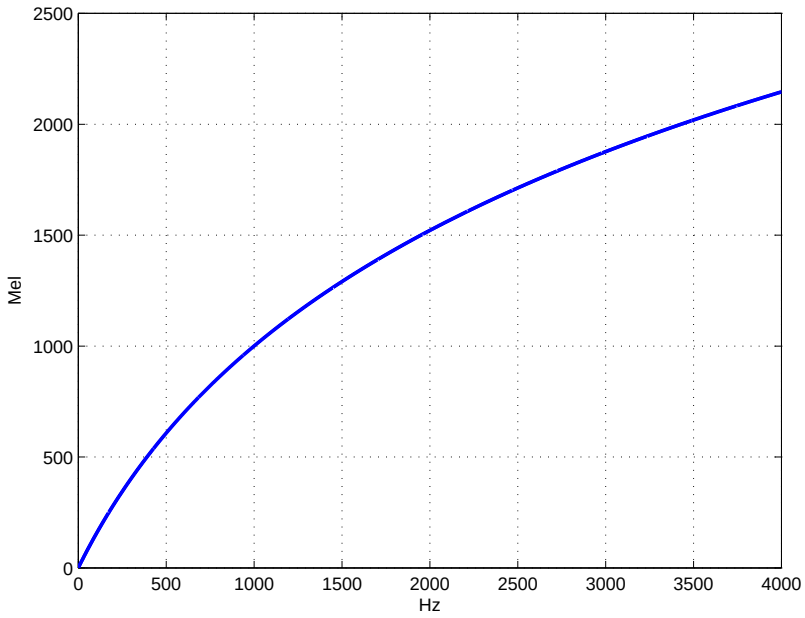


Figure 2.5: Relationship between the Hz frequency scale and the mel-scale.

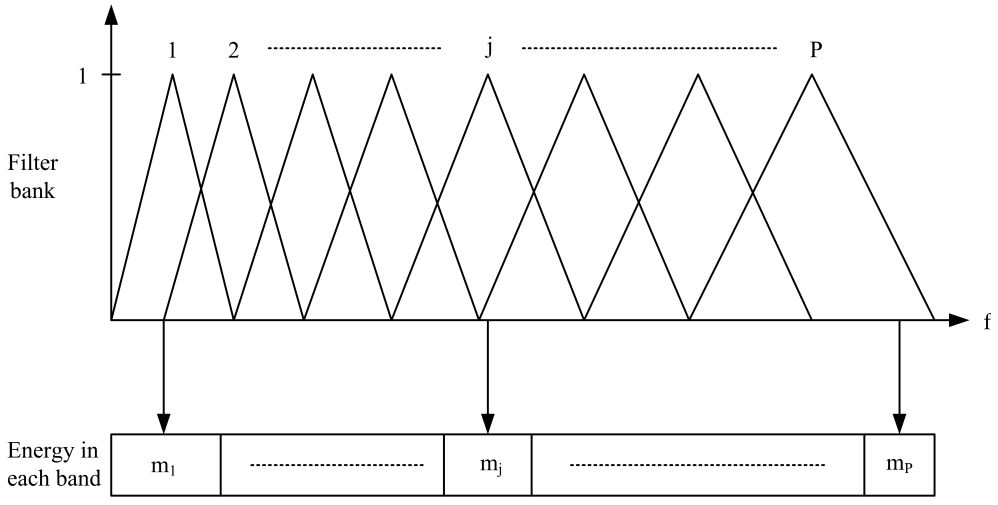


Figure 2.6: Mel-scale filter bank.

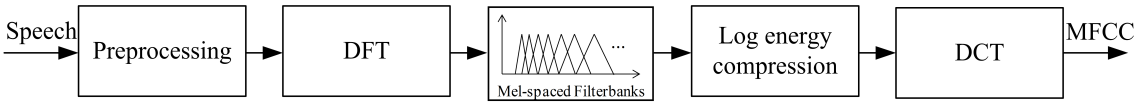


Figure 2.7: MFCC feature extraction process.

The feature extraction approach to MFCC is shown in Figure 2.7. The major computation steps can be summarized as follows.

1. Preprocessing: This step contains the process of pre-emphasis and windowing. The pre-emphasis boosts the energy in the high frequencies using a first order filter. The process of windowing is used to divide the speech signal into successive overlapping frames with minimum effect of the edge discontinuity.

2. The discrete Fourier transform (DFT) converts the windowed speech segment from the time domain into the frequency domain.

3. Mel-spaced filterbanks are used to group the frequency energy into bins with spacing that mimics the frequency response of the auditory system.

4. The discrete cosine transform (DCT) decorrelates the spectral coefficients and transforms the data to the cepstral domain. MFCC features are largely decorrelated and can be modeled with diagonal Gaussian distributions.

Although MFCCs have widely been applied into speech and speaker recognition, MFCC features also have limitations for certain applications. One of the benefits

of MFCCs for speech recognition is to eliminate the speaker-specific information for different speakers [55, 56], and capture the common information of words, phrases, sentences and semantics. In contrast, MFCC for speaker recognition uses phonetic characteristics to model speaker-specific information, without truly representing differences due to other factors. This limitation illustrates that traditional vocal tract features such as MFCC are not suitable for tasks such as our target applications, such as cross-lingual speaker identification and mismatched song-type individual identification as shown in Chapter 4, because those tasks contain little phonetic information across training and test sets.

2.5.2 Greenwood Function Cepstral Coefficients (GFCCs)

For the avian species individual identification task, MFCCs are generalized, with the frequency warping component adjusted according to the perceptual model of the target species - in our case, Ortolan buntings. This leads to the Greenwood Function Cepstral Coefficients (GFCC) [57, 58]. In the mammalian auditory system, the perceived frequency range differs from that of humans. Greenwood found that many mammals perceive frequency on a logarithmic scale along their cochlea [59, 60].

Figure 2.8 shows a block diagram of the process for calculating Greenwood function cepstral coefficients. The first step is to partition the vocalization signal into frames and then multiply each frame with a hamming window. The Greenwood filterbank

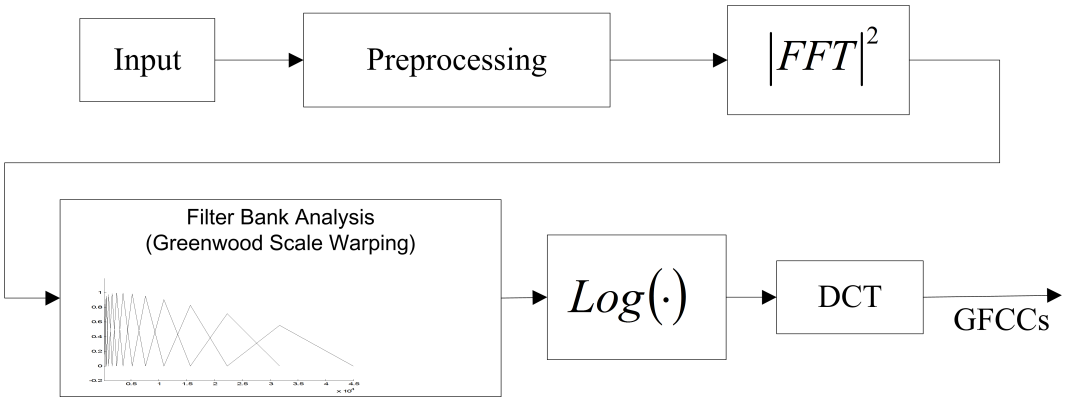


Figure 2.8: GFCC extraction block diagram.

for ortolan bunting vocalizations is adjusted to the range from 400 to 7400 Hz to fit the f_{min} and f_{max} of this species. A discrete cosine transform (DCT) is used to convert the filter bank energy into the cepstral domain.

2.5.3 LPC Coefficients

Linear prediction (LP) analysis is a powerful signal processing technique that has been widely used in the analysis of speech signals. The theory of linear prediction is closely linked to modeling of the vocal tract system, as linear prediction can be shown to be equivalent to auto-regressive modeling of the signal spectrum [61]. In the LP model, the n^{th} speech sample can be predicted by the $(n-1)^{th}$ to $(n-p)^{th}$ samples of the speech wave. The predicted value of the n^{th} speech sample is given by

$$\hat{s}(n) = \sum_{k=1}^{p} a(k)s(n-k), \tag{2.5.2}$$

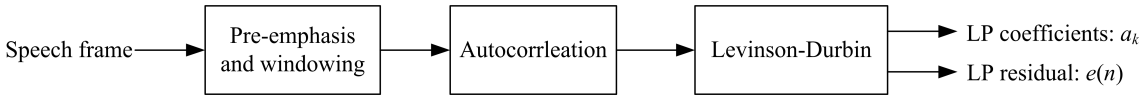


Figure 2.9: LPC coefficient computation.

where $a(k)$ are the predictor coefficients and $s(n)$ is the n^{th} speech sample. The prediction coefficients are determined by minimizing the mean square prediction error, defined as

$$E(n) = \frac{1}{N}\sum_{n=0}^{N-1}\left(s(n) - \hat{s}(n)\right)^2, \tag{2.5.3}$$

where summation is taken over all N samples. The set of coefficients $a(k)$ minimizing the mean-squared prediction error are obtained as the solution of the set of linear equations

$$\sum_{k=1}^{p} a_k R_{xx}[n-j, n-k] = R_{xx}[n, n-j],\ \forall j = 1, \ldots, p. \tag{2.5.4}$$

Thus, the prediction coefficients $a(k)$ are calculated by solving the recursive equation (2.5.4) using the Levinson-Durbin algorithm. Typical steps in implementing linear prediction are summarized in Figure 2.9. Reflection coefficients, log area coefficients and LP residual are also obtained as a byproduct of LP analysis. A detailed introduction to linear prediction can be found in [61].

2.5.4 Log Area Ratio

Log area ratio (LAR) coefficients are well-known spectral measures, typically derived from the linear prediction coefficients [61], which characterize the vocal tract of a

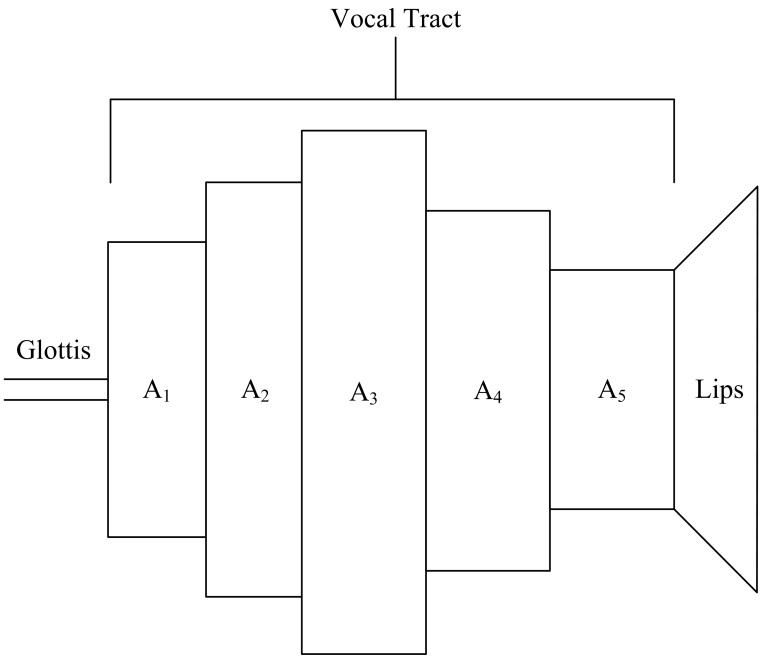


Figure 2.10: Acoustic tubes speech production model.

speaker. Based on LP analysis, the vocal tract model of a speaker can be regarded as a non-uniform acoustic tube which is represented by a concatenation of cylindrical tubes of equal length. The cross-sectional area of each section represents the average cross-sectional area of the corresponding portion of the vocal tract. As boundary conditions, it is assumed that the glottis connected to the first tube has zero area, while the lips connected to the last tube have infinite area. Figure 2.10 illustrates the acoustic tube model of speech production.

The relationship between the LAR coefficients and the LPC coefficients is:

$$LAR_i = \log\left(\frac{A_i}{A_{i+1}}\right) = \log\left(\frac{1+\alpha_i}{1-\alpha_i}\right),\ A_{p+1} = 1, \tag{2.5.5}$$

where α_i represent the parcor coefficients. If we denote $\alpha_i^{(k)}$ as the i^{th} LPC for a k^{th} pole linear prediction model, then $\alpha_i = \alpha_i^{(i)}$, $i = 1, \cdots, p$.

2.5.5 Nuisance Attribute Projection

Nuisance attribute projection (NAP) is an effective technique to compensate the supervectors, consisting of the stacked means of the mixture components [62], by removing the dimensions of unwanted session variability before the SVM training [49, 62]. These dimensions are generally determined through a data-driven approach using a large background database. The within-class variation can be used a model of the inter-session variation by examining the differences between examples of the same speaker in the background population.

There are two assumptions for this NAP approach. The first assumption is that the undesired inter-session variability can be thoroughly estimated in a high dimensional feature space using the covariance matrix. The second assumption is that the corresponding variability lies in a low dimensional subspace spanned by the eigenvectors of the covariance matrix. Based on these two assumptions, the variability can be suppressed by estimating the low dimensional subspace and then be removed. This can be accomplished by applying the transform

$$x' = P_n x = (I - U_n U_n^T)x, \tag{2.5.6}$$

where I denotes the $n \times n$ identity matrix and U_n represents the n^{th} direction of the

variability subspace base U defined by NAP. P_n stands for $n \times n$ projection matrix. U_n is trained to capture the principal direction of within-class variability using several recordings of various speakers. Thus, this approach tries to find the vectors u that maximize the below criterion

$$J(u) = u^T S_w u, \tag{2.5.7}$$

where S_w is the within-class scatter matrix of the training data, which is equivalent to determining the set of eigenvectors with the largest eigenvalues satisfying

$$S_w u = \lambda u. \tag{2.5.8}$$

2.5.6 Latent Factor Analysis (LFA)

Latent Factor Analysis (LFA) is an approach to modeling session variability for text-independent speaker verification, in which a constrained offset of the speaker's GMM supervectors is used to represent the effect of session differences [63]. The method summarizes the session variability in the high dimensional supervector space into a low dimensional eigenchannel factor vectors. In the GMM-LFA framework, the speech signal can be regarded as a normal speech signal corrupted by channel variability. In this approach, the speaker- and channel-dependent mean supervector, $M_{k,c}$, is decomposed into a speaker dependent component and a session dependent vector Uy, defined as

$$M_{k,c} = M_k + Uy, \tag{2.5.9}$$

where M_k is the session-independent GMM supervectors and y is defined to be independent and Gaussian with zero mean and unit variance. If we consider M_k as the trained speaker model and $M_{k,c}$ as the model adapted using the test utterance, equation (2.5.9) formulates the channel differences between the training and testing conditions. The data from multiple speakers is used to train the eigenchannel matrix. In the first step, for each speaker k, $k = 1, \ldots, K$ and all utterances j, $j = 1, \ldots, J_k$, an universal background model (UBM) is adapted to obtain a supervector M_{kj}. In the second step, the true supervector of the corresponding speaker is estimated by averaging all the supervectors from this speaker,

$$\tilde{M}_k = \sum_{j=1}^{J_k} \frac{M_{kj}}{J_k},\ M'_{kj} = M_{kj} - \tilde{M}_k. \tag{2.5.10}$$

Following this, all the speakers' state variability supervectors M'_{kj} are concatenated into a within variability matrix S with ND rows and J columns ($J = \sum_{k=1}^{K} J_k$),

$$S = [M'_{11}, \cdots, M'_{1J_1}, \cdots, M'_{K1}, \cdots, M'_{KJ_K}]. \tag{2.5.11}$$

The eigenchannel matrix U is given by the R PCA eigenvectors of the within-speaker covariance matrix $(1/J)\, SS^t$.

2.6 Speaker Modeling

The feature extraction approaches described above transform the raw speech signal into feature vectors in which speaker-specific information is extracted and statistical

redundancies suppressed. By using feature vectors, a speaker model is trained through the classical modeling methodology. This section describes some of the popular models, such as GMM-UBM and GMM-SVM, in text-independent speaker recognition.

2.6.1 Gaussian Mixture Models-Universal Background Models

2.6.1.1 Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are the state-of-the-art modeling approach in text-independent speaker recognition, and are widely accepted in the speech community [5]. The Gaussian mixture model uses a mixture of multivariate Gaussians to model the probability density function of observed variables. That is, for a GMM with N Gaussians, the likelihood of x given the mixture model, λ, is given by

$$p(x_i \mid \lambda) = \sum_{i=1}^{M} w_i p_i(x_i), \tag{2.6.1}$$

where x_i is the feature vector, and w_i is the weight of each Gaussian component under the constraint $\sum_{i=1}^{M} w_i = 1$. $p_i(x_i)$ is the i^{th} Gaussian density function, given by

$$p_i(x_i) = \frac{1}{(2\pi)^{d/2} |\sum_i|^{1/2}} \exp\left\{ \frac{-(x_i - \mu_m)^T \sum_M^{-1} (x_i - \mu_m)}{2} \right\}, \tag{2.6.2}$$

where d is the dimension of the feature vector, M is the number of Gaussian density, μ_i is the mean of the i^{th} Gaussian density function, $\sum_i$ is the i^{th} covariance matrix and w_i is the weight of the i^{th} Gaussian density function. The model parameters of

GMM are $\lambda = \{M, w_i, \mu_i, \sum_i\}$. Since the number of Gaussian density functions M is often preselected, the estimated model parameters are $\lambda = \{w_i, \mu_i, \sum_i\}$.

GMMs describe the statistical distribution of feature vectors of a sound unit, without regard to temporal sequence information. Because of this, GMMs are able to reduce the influence between the text content and timing on speaker recognition performance. GMMs provide an implicit segmentation of phonetic units within the feature space, without labeling the sound classes. Overall the GMM represents the speaker-dependent vocalization system's configurations.

2.6.1.2 GMM-UBM

The GMM-UBM with MAP adaption is the dominant method of modeling speakers in text-independent speaker recognition [40]. The UBM is a speaker-independent GMM trained by a large amount of data from different speakers and channel/microphone conditions. A UBM doesn't represent the distribution of a particular speaker, but describes the average distribution of speakers whose speech is distorted by channel and background noise, so that the model encompasses certain aspects of the variabilities encountered in the unknown test data.

Typically, the speaker model trained by the speaker enrollment utterances will have 64-256 mixture components. Higher numbers of mixture will require more utterances, which is often difficult to obtain for a single speaker. A better approach is to

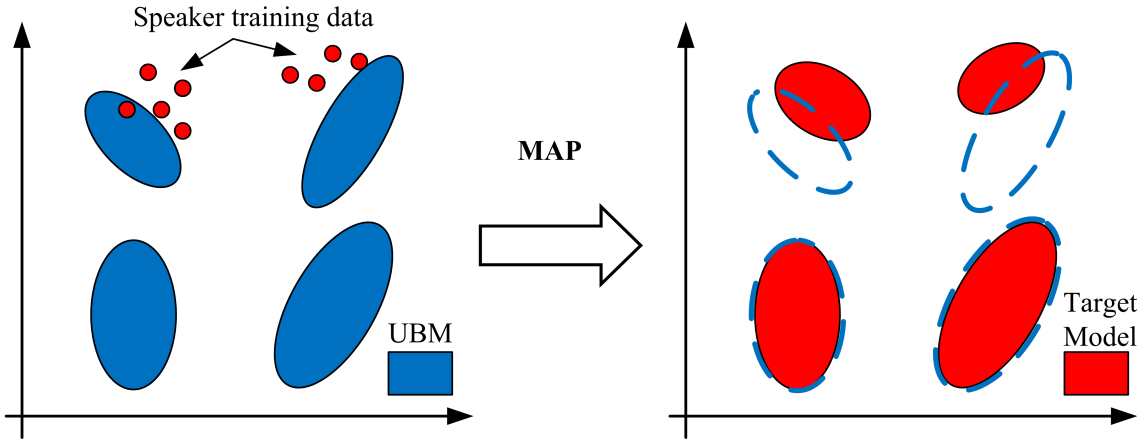


Figure 2.11: The structure of a GMM-UBM.

derive a speaker model using adaption of UBM parameters. This approach employs speaker's enrollment utterance to form a Maximum Posteriori (MAP) estimate [64]. The strategy for adapting target the speaker model is based on the similarity between the training data of the target speaker and the UBM, using MAP adaption to adjust the UBM to the speaker. Distributions within UBM which are far from the target speaker undergo relatively little change as shown in Figure 2.11. To adapt to the target speaker's training data $O = \{o_1, o_2, \ldots, o_r\}$, the algorithm first computes the probabilistic alignment of the training vectors and the UBM mixture components,

$$Pr(i \mid x_t) = \frac{\omega_i p_i(x_t)}{\sum_{j=1}^{M} \omega_i p_i(x_t)}. \tag{2.6.3}$$

The expectation maximization (EM) algorithm [65] iteratively learns the model

parameters from the data to compute the weight, mean and variance.

$$n_i = \sum_{t=1}^{T} Pr(i \mid x_t) \tag{2.6.4}$$

$$E_i(x) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i \mid x_t) x_t \tag{2.6.5}$$

$$E_i(x^2) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i \mid x_t) x_t^2 \tag{2.6.6}$$

These statistics are used to update the UBM statistics and create new adapted parameters, including adapted weights, means, and variance given by

$$\hat{\omega}_i = [\alpha_i^{\omega} n_i / T + (1 - \alpha_i^{\omega}) \omega_i] \gamma, \tag{2.6.7}$$

$$\hat{\mu}_i = \alpha_i^{m} E_i(x) + (1 - \alpha_i^{m}) \mu_i, \tag{2.6.8}$$

$$\hat{\sigma}_i^2 = \alpha_i^{v} E_i(x^2) + (1 - \alpha_i^{v})(\sigma_i^2 + \mu_i^2) - \hat{\mu}_i^2, \tag{2.6.9}$$

where γ is the scale factor to ensure all adapted mixture weights sum to unity, and $\{\alpha_i^{\omega}, \alpha_i^{m}, \alpha_i^{v}\}$ are the m^{th} adaption scale factors controlling the balance between old and new estimates, calculated as

$$\alpha_i^{\rho} = \frac{n_i}{n_i + \gamma^{\rho}}, \ \rho \in \{\omega, m, v\}, \tag{2.6.10}$$

where γ^{ρ} is a fixed relevance factor for parameter $\rho(\gamma = 16)$. When adaption data has low probability for the mixing component i, α_i becomes low. In this case, estimated parameters are more affected by UBM's parameters than by a speaker's statistics.

In contrast, when probability of adaption data is high, the estimated parameters are less affected by the UBM's parameters.

When a target speaker's GMM is adapted from the UBM, it is possible to select whether weights, means and variances of the target speaker's GMM are all adapted, or alternatively whether only means are adapted. Numerous speaker verification experiments [40] have shown that adapting model variances rarely improves accuracy significantly, so it is common to adapt only mean vectors. Based on this observation, this dissertation only adjusts the means of the target speaker's GMM, while the weights and variances are unchanged.

2.6.2 Gaussian Mixture Models-Support Vector Machines

GMM-UBM is a commonly used generative modeling methodology for text-independent speaker recognition. Recently, a discriminative modeling methodology, support vector machines (SVMs), has also been successfully applied to speaker recognition. Based on the SVM's discrimination ability between two classes [66, 67], a combination scheme using GMM and SVM has shown a significant performance applied in a text-independent speaker recognition task over the standard approach using only GMMs. In this section, the GMM-SVM modeling technique is described.

2.6.2.1 Support Vector Machines

Support vector machines are a discriminative approach, well suited to speaker verification. An SVM [68] is a two-class classifier constructed by the weighted sum of a kernel function $K(\cdot,\cdot)$ as follows:

$$f(x) = \sum_{i=1}^{L} \alpha_i t_i K(x, x_i) + d, \tag{2.6.11}$$

where x is the input data, L is the number of support vectors, α_i and d are training parameters, the vectors x_i are support vectors obtained from the training set by an optimization process [69], and t_i are the ideal outputs (either 1 or -1) depending upon whether the corresponding support vector is in class 1 or class 2, respectively. A classification decision is based upon whether the value $f(x)$ is above or below a threshold.

The kernel function $K(\cdot,\cdot)$ is constrained to have certain properties (the Mercer condition) so that $K(\cdot,\cdot)$ can be expressed as

$$K(x, y) = b(x)^t b(y) \tag{2.6.12}$$

where $b(\cdot)$ is the mapping that converts each data vector x from the input feature space into a high-dimensional SVM expansion space.

The SVM optimization process creates a hyperplane in the expansion space that can effectively separate between the two classes with maximum margin. This hyperplane is defined by support vectors x_i that are identifying boundary points from the

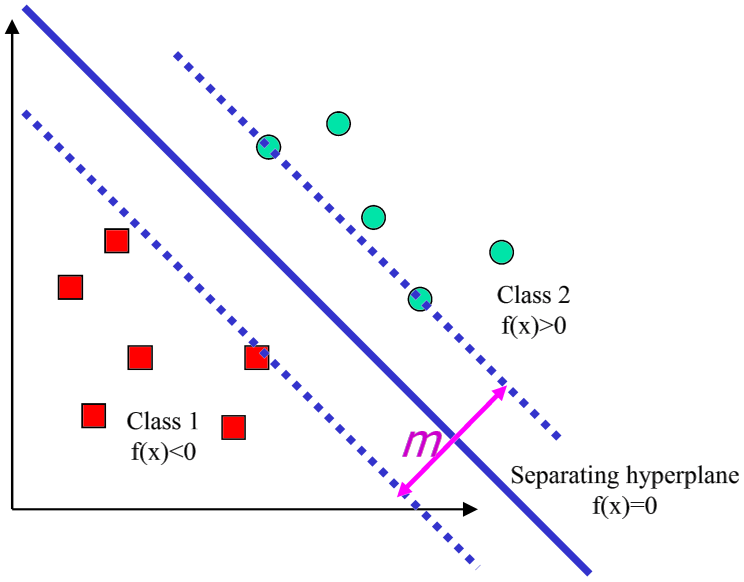


Figure 2.12: Principle of support vector machine.

training data. Each input vector x is classified to class 1 or 2 according the sign of $f(x)$ at the evaluation stage. The principle of the support vector machine is illustrated in Figure 2.12. As shown in this figure, the decision boundary should be as far away from the data of both classes as possible in order to maximize the margin m.

2.6.2.2 GMM-SVM

A GMM-SVM system for speaker verification [70] is a combination of the generative power of GMM-UBM with the discriminative properties of SVM, as shown in Figure 2.13. The GMM-UBM system serves as a means of feature extraction for the SVM system. The SVM classifier is used to model the target speaker characteristics and to score the test utterances. Given the SVM of target speaker s, the verification score

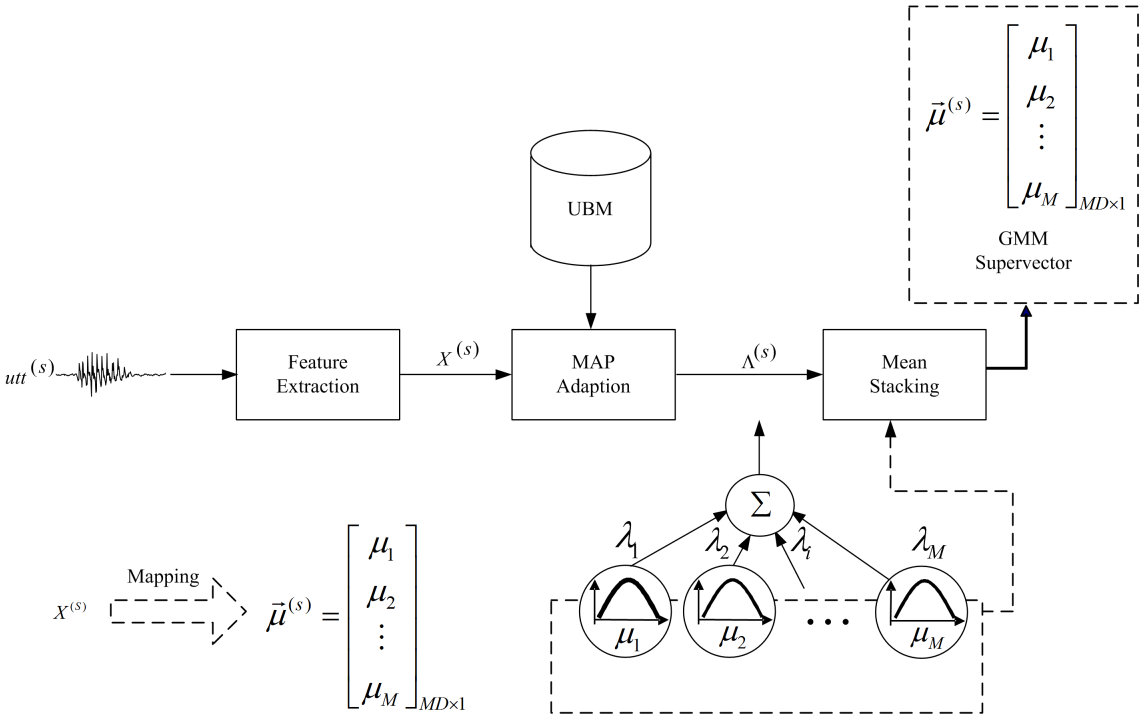


Figure 2.13: GMM-SVM for speaker verification.

of $X^{(c)}$ is given by

$$S_{GMM-SVM}(X^{(c)}) = \alpha_0^{(s)} K(X^{(c)}, X^{(s)}) - \sum_{i \in \text{SV from bkg}} \alpha_i^{(s)} K(X^{(c)}, X^{(b_i)}) + d^{(s)}, \quad (2.6.13)$$

where $\alpha_0^{(s)}$ is the Lagrange multiplier corresponding to the target speaker, $\alpha_0^{(s)}$ are Lagrange multipliers corresponding to the background speakers, and $X^{(b_i)}$ is the utterance of the i^{th} background speaker. The most popular kernel $K(\cdot, \cdot)$ in speaker verification is the GMM-supervector kernel [70]:

$$K(X^{(c)}, X^{(s)}) = \sum_{j=1}^{M} \left(\sqrt{\lambda_j} \sum\nolimits_j^{-\frac{1}{2}} \mu_j^{(c)} \right)^T \left(\sqrt{\lambda_j} \sum\nolimits_j^{-\frac{1}{2}} \mu_j^{(s)} \right), \quad (2.6.14)$$

where λ_j are the mixture weights, $\sum_j$ are the covariances of the UBM, and $\mu_j^{(c)}$ and $\mu_j^{(s)}$ are the j^{th} mean vectors of the GMM for the speaker s and claimant c.

2.6.3 I-vector Analysis

The GMM-UBM and GMM-SVM techniques have shown reliable performance for text-independent speaker recognition system [38, 71, 72]. Presently, the i-vector technique, originating from joint factor analysis (JFA) [73], has improved performance further and is quickly becoming a dominant approach. Similar to eigenvoices, the means of the speaker independent system are stacked together to form a supervector m of dimension Nd, where N refers to the number of Gaussian densities in the model and d is the feature vector dimension. Given an utterance i, the adapted mean supervector $M(i)$ is obtained by the following equation

$$M(i) = m + Tw(i), \tag{2.6.15}$$

where T is called the total variability matrix of size $Nd \times R$ estimated at training time from a large set of utterance (containing the speaker and channel variability simultaneously), and $w(i)$ is the resulting i-vector of dimension R corresponding to the utterance i in the subspace spanned by the columns of T [9]. Since $R \ll Nd$, the i-vector $w(i)$ can be estimated from a single utterance, leading to an adapted recognition model.

For each observation χ, the aim is to estimate the parameters of the posterior

probability of w

$$p(w \mid \chi) = N(w, w_\chi, L_\chi^{-1}). \tag{2.6.16}$$

The i-vector is the MAP point estimate of the variable w. It maps most of the relevant information from a variable length observation χ to a fixed and low dimension vector. The point estimate of the i-vector for a given observation data is given

$$w_0 = L_0^{-1} T^T \sum^{-1} \bar{F}, \tag{2.6.17}$$

where the supervector $\bar{F}$ is obtained by concatenating $\bar{F}_c$ vectors ($\bar{F} = [\bar{F}_1; \ldots; \bar{F}_C]$). The covariance matrix w is given as L^{-1}, which is computed as

$$L_0 = I + \sum_{c=1}^{C} N_c T_c^T \sum^{-1} T_c. \tag{2.6.18}$$

Parameters of the total variability matrix are estimated through EM algorithm. In the expectation step, accumulators are calculated over all K training utterances and for all C mixture components as

$$\bar{C} = \sum_{k=1}^{K} \bar{F}^k w_k^T \tag{2.6.19}$$

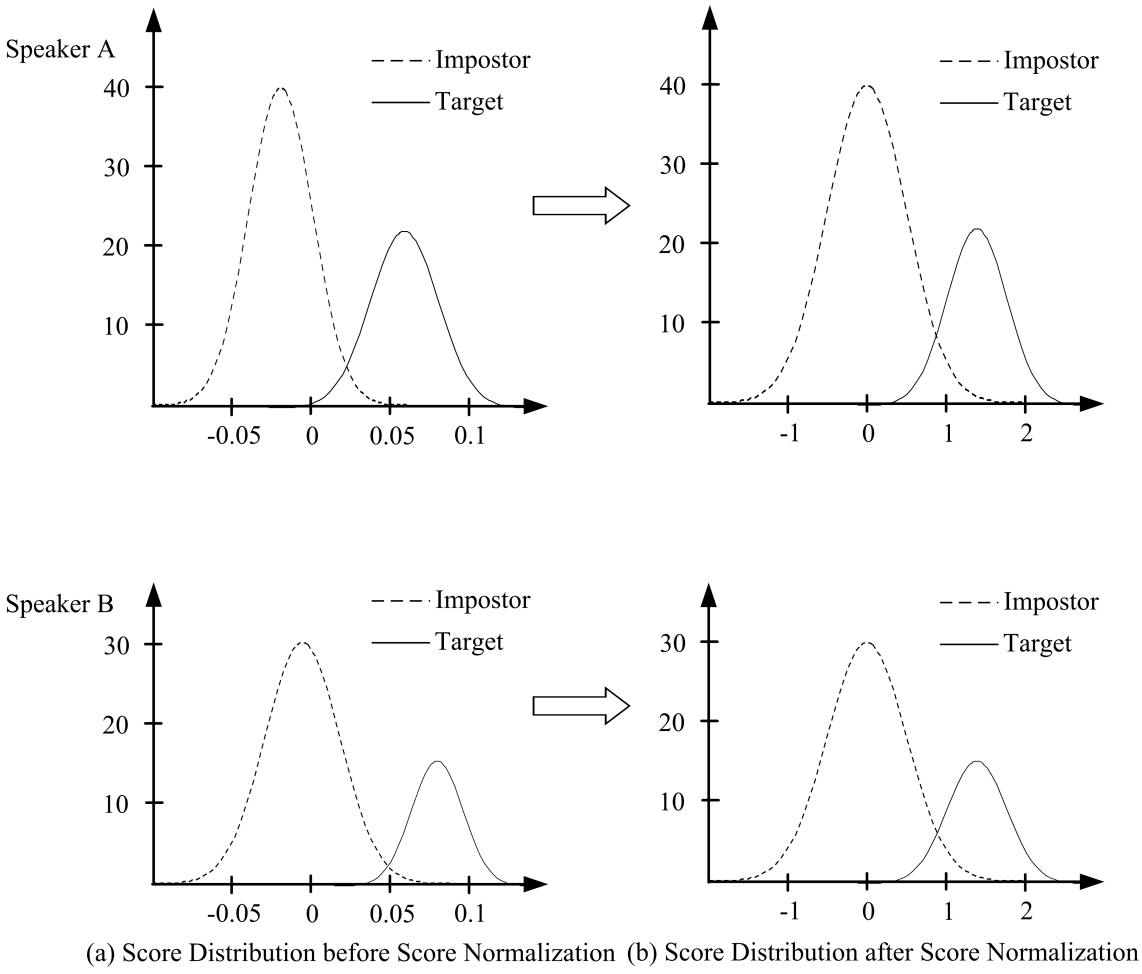
$$\bar{A}_c = \sum_{k=1}^{K} N_c^k (L_k^{-1} + w_k w_k^T). \tag{2.6.20}$$

In the maximization step, new estimate of the total variability matrix is calculated as

$$\bar{T}_c = \overline{CA}^{-1}. \tag{2.6.21}$$

2.7 Score Normalization

The determination of the likelihood ratio threshold involves two types of data: target speaker and impostor speaker. For many applications, each target speaker's data is very limited and the impostor data is extensive. If a separate threshold for each target speaker is configured, the performance and robustness of the threshold will deteriorate. Therefore, it is desirable to have a global threshold for all target speakers, which can be accomplished using a score normalization technique [13]. The aim of score normalization is to transform scores from different speakers into a similar range so that a common verification threshold can be chosen as shown in Figure 2.14. In the left column of Figure 2.14(a), the score distributions of each target speaker's model to the target and impostor data still have bias. This will cause a difference between the global threshold for all speakers and the true threshold for each individual target speaker, and make the global threshold unacceptable. Therefore, the output score needs to be normalized in order to transform the distribution of the output scores to the similar range so that a global decision threshold can be used for all speakers, as shown in the right column of Figure 2.14(b). Empirically, the score distributions of the target and impostor speakers approximate Gaussian distributions. Thus, straightforward score normalization method can be used to normalize the score into a normal distribution with mean 0 and variance 1. After normalization, the



(a) Score Distribution before Score Normalization (b) Score Distribution after Score Normalization

Figure 2.14: The bias of LLR score ranges for different speaker models.

global threshold can be easily determined.

Another reason for score normalization is to compensate for channel, environmental, and intra-speaker variabilities. These variabilities are caused by different environmental conditions, different microphones or headsets, different transmission conditions, and varying phonetic content across utterances. These variabilities can lead to mismatch between training and test, which can in turn cause score mismatch.

Hence, score normalization can also be used to minimize these variabilities at a score level. Common normalization techniques include using a set of impostor models to map score distributions, using methods such as zero normalization (Z-Norm) and test normalization (T-Norm). These will be discussed in the following sections.

2.7.1 Z-Norm

The Z-Norm method normalizes the score distribution using target-specific statistics

$$S_{ZNorm} = \frac{\log(P(\lambda \mid O)) - \mu_i}{\sigma_i}, \tag{2.7.1}$$

where μ_i and σ_i are respectively the mean and standard deviation of the scores, and $\log(P(\lambda \mid O))$ is the target speaker model against a set of impostor utterances. As seen in Figure 2.15, this approach compensates for inter-speaker variability, with normalization parameters estimated from scoring a set of imposter utterances through each speaker model.

The advantage of the Z-Norm method is that the normalization parameter calculation can be done offline, with the parameters calculated in the training stages. There is almost no additional computation during recognition. However, differing background and channel conditions between training and testing utterances will reduce the normalization performance.

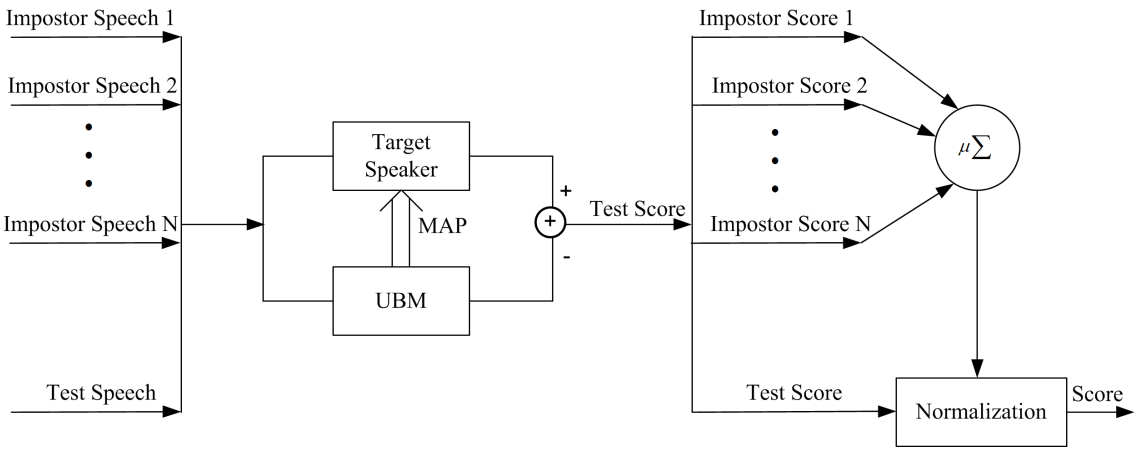


Figure 2.15: The Z-normalization technique.

2.7.2 T-Norm

T-Norm is a popular score normalization method in which the normalization parameters are estimated using scores derived at test time from a set of imposter speaker models. As shown in Figure 2.16, a set of imposter speaker models are scored in parallel with the target speaker model. The mean and standard deviation of the imposter scores are used to adjust the target speaker score as

$$S_{TNorm} = \frac{\log(P(\lambda \mid O)) - \mu_i}{\sigma_i}, \tag{2.7.2}$$

$$\mu_i = \frac{1}{n}\sum_{i}^{n} \log \Lambda_{I_i}, \tag{2.7.3}$$

$$\sigma_i = \sqrt{\frac{1}{n-1}\sum_{i}^{n} (\log \Lambda_{I_i} - \mu_i)}. \tag{2.7.4}$$

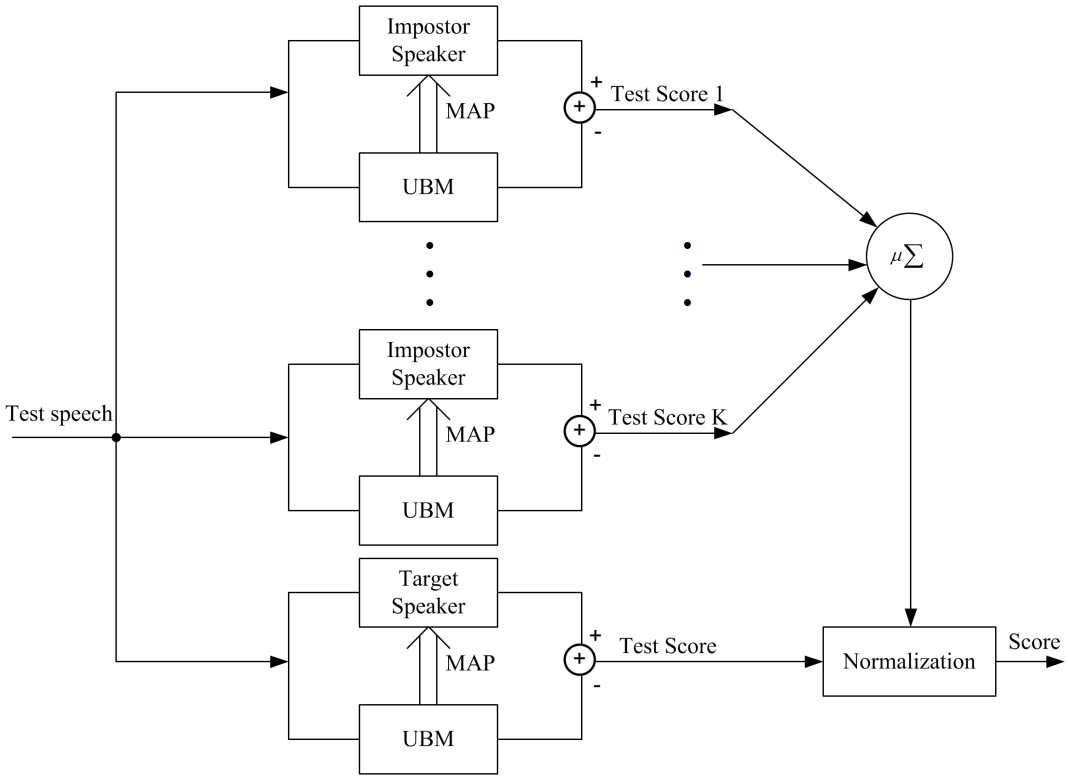


Figure 2.16: The T-normalization technique.

Compared to Z-Norm, T-Norm only requires the test utterance without need of additional imposter utterances, and thus avoids mismatched conditions between the training and testing utterance. However, T-Norm is an online method in which each test utterance is scored against a set of impostor models, increasing the calculation time of parameters.

CHAPTER 3

NOVEL SOURCE-BASED FEATURES FOR SPEAKER RECOGNITION

3.1 Introduction

The purpose of feature extraction is to parameterize a speech waveform into a sequence of feature vectors that are effective for modeling and recognition. In general, differences in speech characteristics among different speakers can be broadly categorized according to two main aspects, differences in vocal organs and differences in pronunciation habits. Organic physiological differences are caused by variations of sizes and shapes of the vocal tract components, including the larynx, pharynx, tongue, teeth, and the oral and nasal cavities. These organic differences show up in the frequency structure, reflecting both the characteristics of the spectral envelope caused by the vocal tract shape as well as the vibration pattern of the vocal fold. Beyond this, differences in learned pronunciation habits lead to variations in acoustic-phonetic characteristics as well as unique dynamics of the vocal tract during coarticulation and formant transition. To extract informative features of a speech signal for speaker recognition, it is important to fully understand the mechanism of speech production.

Based on the above separation of speech production systems, features for speaker recognition can be divided into two distinct types, vocal tract and vocal source related features. The former represent the shape of the spectral envelope derived from a spectral analysis of the speech waveform resonances. The latter capture the time/frequency characteristics of the glottal pulse waveform of a speaker's larynx. This chapter focuses on the mechanisms of speech production and the corresponding feature extraction approaches for extracting both vocal tract and vocal source related features for speaker recognition, with the goal of identifying under-utilized information in the vocal signal to create novel features for physiologically-driven speaker identification.

3.2 Speech Production Model

Speech production mainly results from the activity of three functionally distinct systems: the subglottal lungs, the larynx, and the supra-laryngeal vocal tract. Based on this, this process can be divided into three stages: respiration, phonation and articulation. Respiration provides the primary source of air during speech. Phonation then changes the steady airflow from the lungs into a quasi-periodic pulsating signal which then moves into the vocal tract. Articulation modulates the signal from the larynx into its temporal and spectral structure, generating different types of sounds. The schematic representation of speech production is shown in Figure 3.1. This section

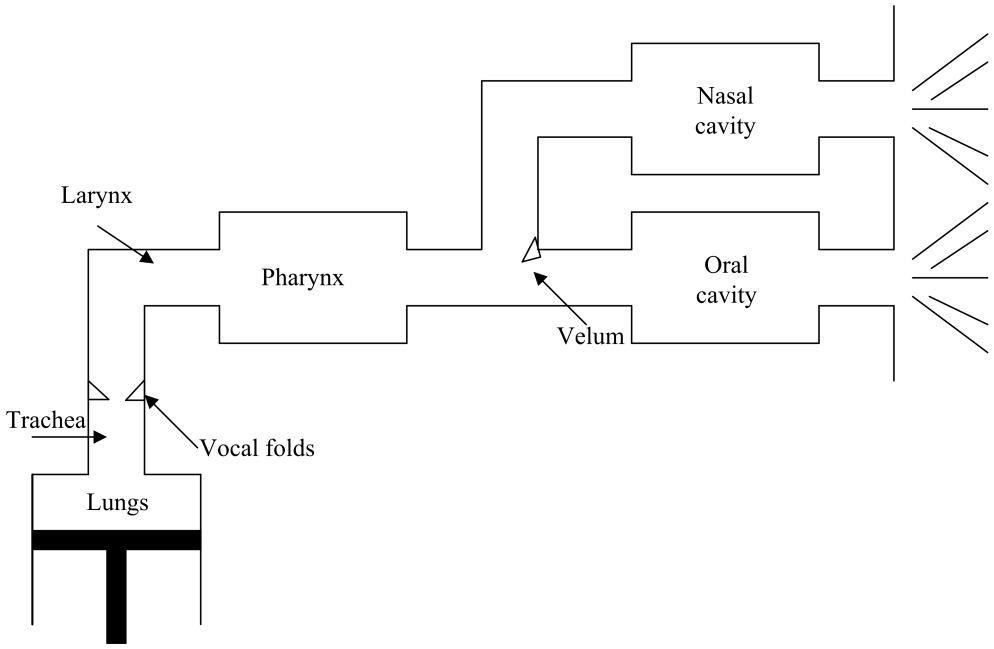


Figure 3.1: Schematic representation of speech production.

will give an introduction to the three distinct phases of speech production [74, 75].

3.2.1 The Sub-glottal Respiratory System

The sub-glottal component produces an airstream which provides the generative power of the speech production process. During inspiration, muscular force is used filling the lungs. The lungs expand, causing the air to flow from the mouth to the lungs with the glottis relatively open. During expiration, this energy will be spontaneously released, as the lungs contract, pushing the air from the lungs toward the mouth.

3.2.2 The Larynx

The larynx is used to convert an airstream into audible sounds. This refers to producing acoustic energy which serves as an input to the vocal tract. The larynx, particularly the action of the vocal folds, determines the phonation types, whose major types are voiceless, whisper and voiced. During whispering and voiceless phonation, the vocal folds are apart from each other, and the airstream passes through the open portion, the glottis. The difference between whispering and voiceless phonation is a function of the degree of the glottal opening. In voiceless speech, the glottal area is larger and the airstream is slightly turbulent when it enters the vocal tract. In whispering, the glottal area is smaller. This will create a more turbulent noise airstream.

Voicing phonation is more complex than whispering and voiceless speech. The adductor muscles are activated to produce sound, providing resistance to exhaled air from lungs. Air then bursts through the closed vocal folds. The vocal folds move outward and the glottis opens. As the air rushes through the vocal folds, the pressure between the vocal folds drops, sucking them back together. This is known as the *Bernoulli Effect*, as shown in Figure 3.2. This vibration is repeated hundreds or even thousands of times per second, producing what we perceive as vocal pitch. This periodically repeated opening and closing of the glottis is the response to subglottal air pressure from the trachea. The rate of opening and closing of the vocal folds in the larynx is called the *fundamental frequency* and abbreviated F_0. An example of

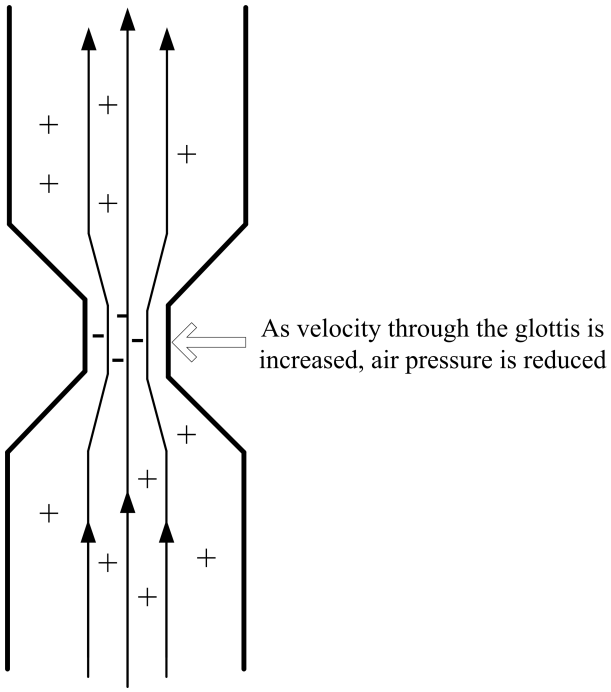


Figure 3.2: The Bernoulli effect.

the spectrum of such glottal air flow is shown at the leftmost column of Figure 3.3. Note that there is energy at the fundamental frequency (F_0=100 Hz and 200 Hz) and at the harmonics of the fundamental, and that the amplitude of the harmonics gradually decreases.

Fundamental frequency differs among males, females and children. This results from anatomical differences. Females usually have a high F_0 than males because females have smaller vocal folds than males. Children have even smaller vocal folds. F_0 also contributes directly to the perception of pitch (the semi-musical rising and falling of voice tones) in speech. It is noted that F_0 is defined only for voiced phonation

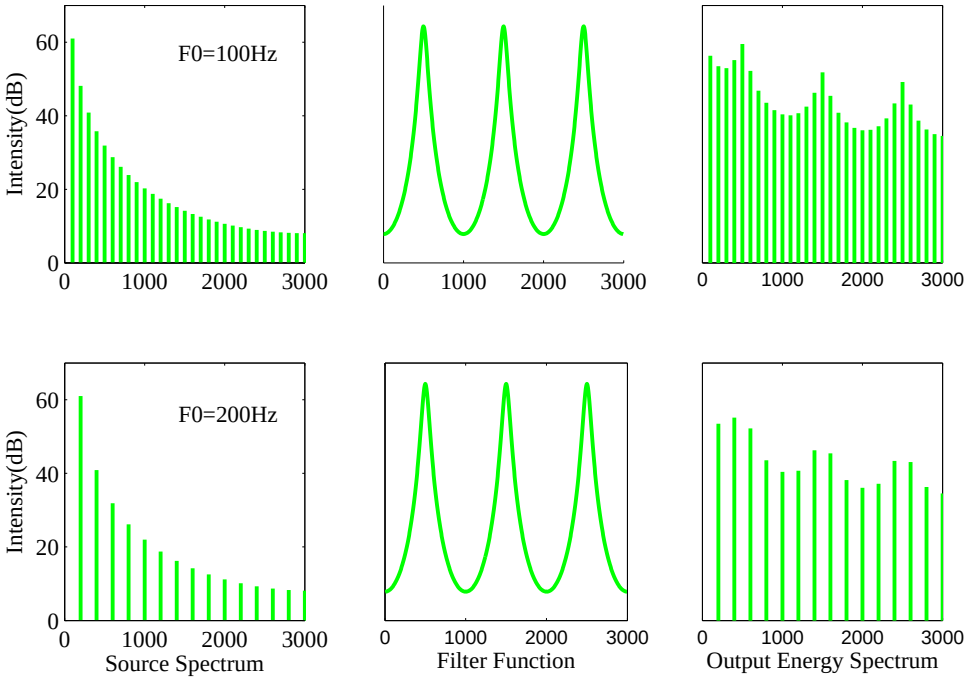


Figure 3.3: Acoustic output energy spectrum as a combination of laryngeal source excitation coupled with vocal tract function.

and is undefined for all unvoiced phonation.

3.2.3 The Supralaryngeal Vocal Tract

The action of the vocal folds classifies speech as voiced or unvoiced, while the vocal tract composed of the oral and the nasal cavities and airways transforms those sounds into intelligible speech. The vocal tract can be viewed as a time-varying acoustic filter that amplifies and filters the sound energy and shapes its frequency spectrum. Those frequencies at which there is maximum energy and the air vibrates with maximum

amplitude are called *formants*, and they are determined, in part, by the overall shape, length and volume of the vocal tract. The detailed shape of the filter function is determined by the whole vocal tract serving as an acoustically resonant system that includes losses due to the radiation at the lips. An idealized filter function is shown in the center of Figure 3.3. This idealized filter function has formant frequencies at 500, 1500 and 2500 each with a bandwidth of 100Hz. The rightmost column of the figure shows the energy spectrum resulting from filtering the laryngeal source spectrum at the leftmost figure with the filter function shown in the center of the figure.

3.2.4 The Linear Time-invariant Source-filter Model

As described above, the speech production system is usually modeled as two separate and independent processes: the sound generation in the larynx (source) and the acoustic filtering of the speech sounds in the vocal tract (filter). In summary, the glottal source excitation is produced by the vibratory behavior of the vocal folds and is modulated by the resonances of the vocal tract and radiated through the lips and/or nostrils to form the speech signal. This process can be modeled by a source-filter model as shown in Figure 3.4 [76, 77], in which a vocal tract model $V(z)$ and a radiation model $R(z)$ are excited by a glottal excitation signal $u(n)$.

Glottal excitation $u(n)$ includes voiced and unvoiced excitation, according to the action of the vocal folds. Unvoiced excitation is modeled as random Gaussian noise.

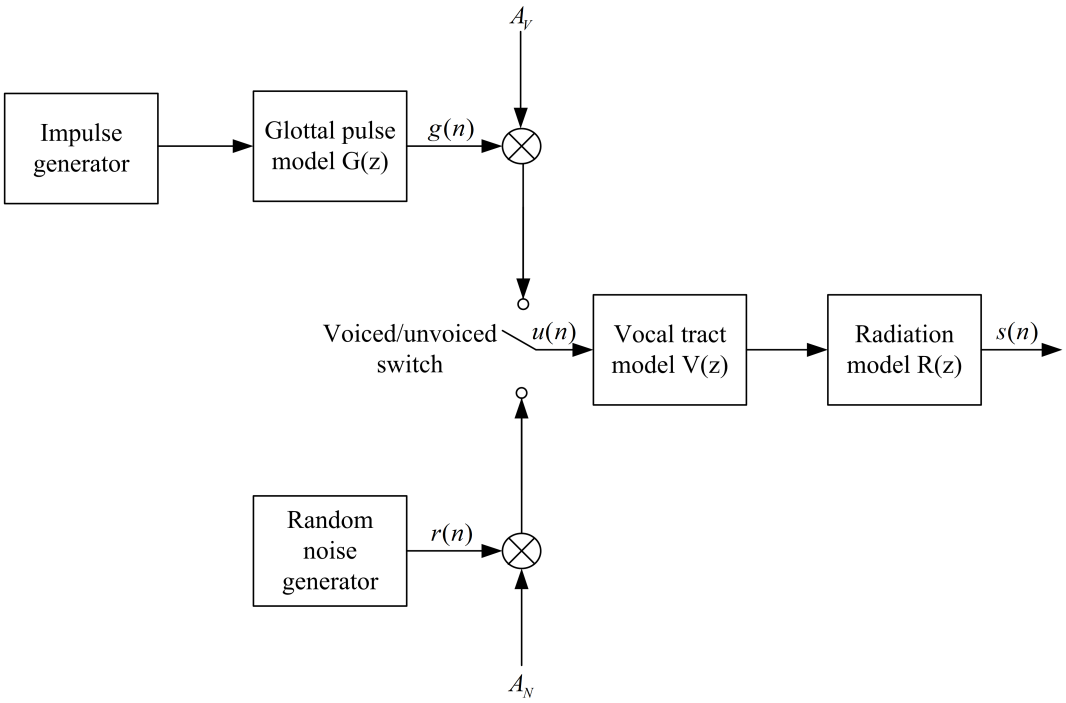


Figure 3.4: Acoustic model for speech production.

Voiced excitation is regarded as an impulse sequence exciting a time-varying filter $G(z)$ to produce a glottal source waveform $g(n)$. Generally, $G(z)$ can be considered as an all-pole model, e.g., a two-pole model,

$$G(z) = \frac{1}{[(1 - g_1 z^{-1})(1 - g_2 z^{-1})]}. \tag{3.2.1}$$

The vocal tract, consisting of both the oral and nasal airways, can be represented as a concatenation of N lossless tubes with constant cross-section areas and equal length, as discussed in Section 2.5.4 [78]. In source-filter modeling, the vocal tract

can be represented by an all-pole model with transfer function

$$V(z) = \frac{1}{1 - \sum_{i=1}^{P} a_i \cdot z^{-i}}. \tag{3.2.2}$$

The poles appear as conjugate pairs, with each pair generating a formant in the speech spectrum. The lip radiation model, which models the coupling of the vocal tract to the surrounding air volume, is approximated as a first order difference

$$R(z) = R_0(1 - z^{-1}). \tag{3.2.3}$$

Thus, the acoustic characteristics of speech are usually modeled as a sequence of source, vocal tract filter, and radiation characteristics

$$S(z) = A_V G(z) V(z) R(z). \tag{3.2.4}$$

For different speakers, the excitation source and vocal tract vary significantly. The physical size and shape of a speaker's larynx, vocal folds and vocal tract determine the range of the sounds that can be produced by the speaker. Since each speaker has their own vocal characteristics and most of them remain essentially unchanged, features based on vocal source characteristics are highly informative for speaker recognition.

3.2.5 Nonlinear Speech Production Models

The linear source-filter model introduced in the previous section assumes that the airflow propagates in the vocal tract as a plane wave. This pulsatile flow is considered

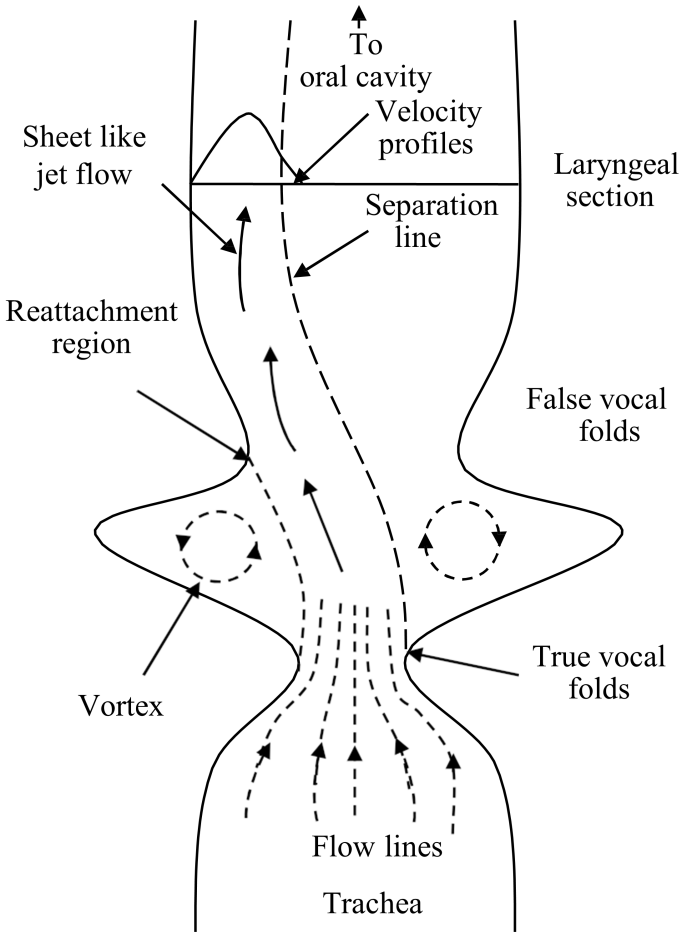


Figure 3.5: Nonlinear model of sound propagation along the vocal tract.

the source excitation of speech production [79]. However, studies by Teager [80] show that the air flow is actually separate and that concomitant vortices are distributed throughout the vocal tract, as depicted in Figure 3.5 [81]. Teager suggested that the actual source of speech production is nonlinear vortex-flow interactions. This observation is also supported by the theory of fluid mechanics [82].

Based on this, human speech production can also be modeled nonlinearly. To

characterize speech production in this way, the airflow pattern in the vocal tract is modeled decomposing the speech production model's characteristics for both vocal fold movement and vocal tract structure. In nonlinear modeling, the resulting airflow properties serve to excite those models which a listener will perceive as a specific speaker's voice [80, 81]. Although the airflow pattern shown in Figure 3.5 is closer to the real speech production process, there has been no single mathematical model to quantify this nonlinear speech production process, because complete solutions of airflow require accurate boundary conditions versus time. Teager [80] developed an energy tracking operator called the Teager Energy Operator (TEO) to reflect the instantaneous energy of the nonlinear vortex-flow interaction. Kaiser [83] introduced a simple and elegant form of this TEO as

$$\Psi_c[x(t)] = \left(\frac{d}{dt}x(t)\right)^2 - x(t)\left(\frac{d^2}{dt^2}x(t)\right) = [\dot{x}(t)]^2 - x(t)\ddot{x}(t), \tag{3.2.5}$$

where Ψ is the Teager energy operator, and $x(t)$ is a continuous speech signal.

The operator in the discrete-time form is

$$\Psi[x(n)] = x^2(n) - x(n+1)x(n-1), \tag{3.2.6}$$

where $x(n)$ is the sampled speech signal.

This TEO is typically applied to a band-pass filtered speech signal since its purpose is to reflect the energy of this nonlinear energy flow within the vocal tract for a single resonant frequency. Thus, the corresponding TEO profile can be used to decompose

a speech signal into its amplitude modulation (AM) and frequency modulation (FM) components within a certain frequency band by

$$f(n) \approx \frac{1}{2\pi T} \arccos\left(1 - \frac{\Psi[y(n)] + \Psi[y(n+1)]}{4\Psi[x(n)]}\right), \tag{3.2.7}$$

$$\mid a(n) \mid \approx \sqrt{\frac{\Psi[x(n)]}{\left[1 - \left(1 - \frac{\Psi[y(n)]+\Psi[y(n+1)]}{4\Psi[x(n)]}\right)^2\right]}}, \tag{3.2.8}$$

where $y(n) = x(n) - x(n-1)$ is the time domain difference signal, $\Psi[\cdot]$ is the TEO operator as shown in (3.2.6), $f(n)$ is the FM component at sample n, and $a(n)$ is the AM component at sample n.

An alternative nonlinear speech signal model has been proposed by Maragos [84] as

$$s(t) = \sum_{m=1}^{M} r_m(t), \tag{3.2.9}$$

where

$$r_m(t) = a_m(t) \cos\left(2\pi\left(f_{cm}t + \int_0^t q_m(\tau)d\tau\right) + \theta\right) \tag{3.2.10}$$

is a combined AM and FM structure representing a speech resonance at the m^{th} formant with a center frequency $F_m = f_{cm}$. In the above equation (3.2.10), $a_m(t)$ is the time-varying amplitude and $q_m(\tau)$ is the frequency modulation signal at the m^{th} formant.

3.3 Vocal Source Measures

3.3.1 LP Residual

Using the Linear Predictive Coding analysis discussed in Section 2.5.3, the linear prediction (LP) residual is defined as the prediction error $e(n)$ obtained from the difference between the predicted speech sample $\hat{s}(n)$ and the current sample $s(n)$ [61]. This is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a(k)s(n-k). \tag{3.3.1}$$

In the frequency domain, equation (3.3.1) can be represented as,

$$E(z) = S(z) + \sum_{k=1}^{p} (a_k S(z))z^{-k}. \tag{3.3.2}$$

By this we obtain

$$A(z) = \frac{E(z)}{S(z)} = 1 + \sum_{k=1}^{p} (a_k)z^{-k}. \tag{3.3.3}$$

Hence, we can obtain the LP residual signal by filtering the speech signal with $A(z)$ as indicated in Figure 3.6.

After LP analysis has estimated the vocal tract information from the speech signal, inverse filtering leaves a whitened LP residual signal that possesses abundant information about the excitation source of the speaker, including speaker-unique characteristic information in both the time and frequency domains. Some useful features, such

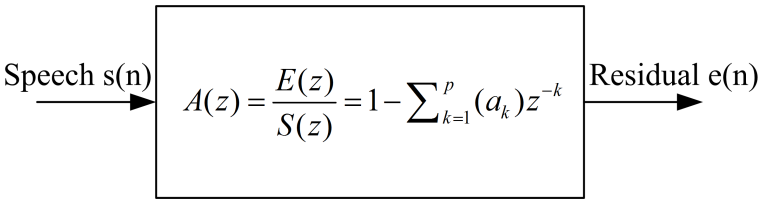


Figure 3.6: Computing the LP residual by inverse filtering.

as pitch, residual energy and residual cepstrum, have been previously investigated for speaker recognition [85–88]. Features related to excitation signals have contributed relatively good performance. However, these features are mostly based on power spectra of the excitation, where phase information of the residual signal has been removed. It has been shown that the phase spectrum can also contribute significantly to speech perception and speech recognition, perhaps even as much as the magnitude spectrum [89, 90]. Here, the question of whether phase spectrum of residual signal provides any useful information for speaker recognition will be investigated, as discussed in detail in Section 3.4.

3.3.2 Fundamental Frequency

Fundamental frequency (F_0) represents the rate of vibration of the vocal folds in the larynx during the production of speech. Average fundamental frequency contains significant speaker specific information, and several studies have demonstrated that fundamental frequency relates directly to anatomical differences [91].

Fundamental frequency can be described by the following equation [92, 93]

$$F_0 = \frac{1}{2L_m}\sqrt{\frac{\sigma_c}{\rho}}, \tag{3.3.4}$$

where L_m is the length of the vocal cords, σ_c is the longitudinal tension of the cords divided by the cross sectional area of the vibrating tissue and ρ is the density of vocal cord. The relationship between F_0 and the length of the vocal folds is inversely proportional. A speaker with a long vocal fold corresponds to a low F_0, while a speaker with a short vocal fold has a high F_0.

There are several approaches to estimating the fundamental frequency, for example [94], one analysis method involves determination of the average length of several successive periods. This approach is implicitly implemented by algorithms that use short-term analysis, such as auto-correlations. Another reliable approach is to use the cepstrum, a Fourier analysis of the logarithmic amplitude spectrum of the signal and turning the spectrum inside-out. An estimate of the fundamental frequency is obtained by searching for a peak from the cepstrum.

3.3.3 Jitter

Jitter is the relative evaluation of the period-to-period variability of the pitch within a frame [95]. The terminology was given by Lieberman [96] when he displayed speech waveforms on an oscilloscope and showed that no two periods were exactly alike. The fundamental frequency appeared "jittery" because of this period variation.

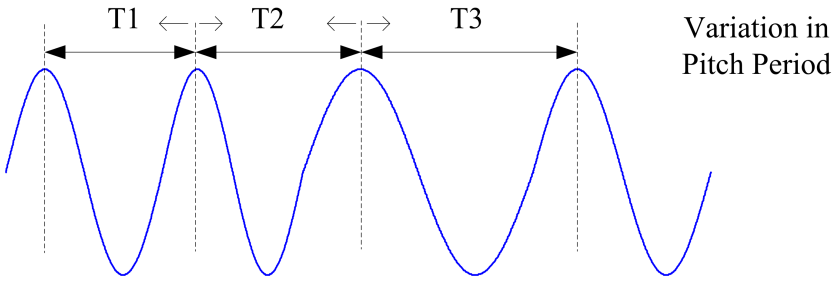


Figure 3.7: Jitter measurement.

The jitter in a frame is defined as

$$\text{Jitt}_i = \frac{\left|T_0^{(i)} - T_0^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} T_0^{(i)}}, \tag{3.3.5}$$

where $T_0^{(i)}$, $i = 1, 2, \ldots, N$ is the extracted pitch period of the i^{th} frame and N is the total number of voiced frames in the utterance, as shown in Figure 3.7.

3.3.4 Shimmer

Shimmer is the relative evaluation of the period-to-period variability of the peak-to-peak amplitude [97]. It is defined as the short term variation in the amplitude of the vocal-fold vibration due to perturbation in the amplitude of the glottal pressure waveform. Shimmer reflects the transient change of the utterance's energy.

The shimmer of each frame is calculated by

$$\text{Shim}_i = \frac{\left|A^{(i)} - A^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} A^{(i)}}, \tag{3.3.6}$$

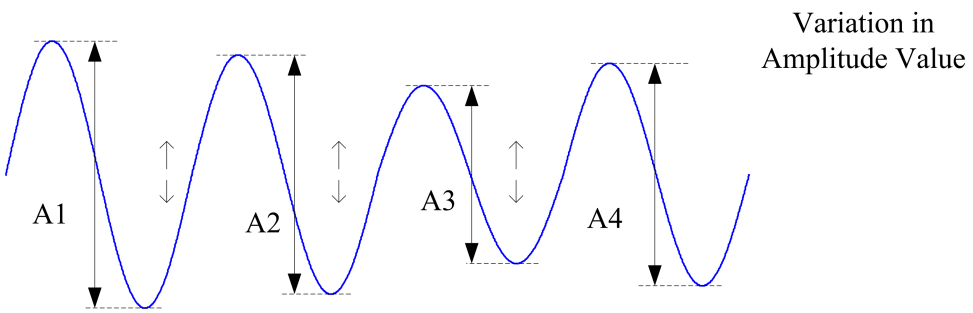


Figure 3.8: Shimmer measurement.

where $A^{(i)}$, $i = 1, 2, \ldots, N$ is the extracted peak-to-peak amplitude and N is the total number of voiced frames in the utterance, as shown in Figure 3.8.

3.3.5 Epoch Location

The epoch location of the source waveform is the instant of glottal closure of the vocal tract system during speech production. Although the excitation source for voiced speech is a sequence of glottal pulses, the most significant excitation takes place around the instant of glottal closure, defined as the *epoch*. The accurate estimation of the epoch locations within a glottal pulse is useful for many speech analysis situations. For instance, the measurement of epoch locations is beneficial to accurately estimating the fundamental frequency. After the glottal closure, the glottal airflow becomes zero. As a result, the supralaryngeal vocal tract system can be acoustically decoupled from the trachea [98]. Thus, the speech signal within the closed region represents the free resonances of the vocal tract system. The characteristics of the vocal source can

also be determined by an analysis of the speech signal within a glottal pulse using the epochs. In [99], the excitation features extracted from the regions around the epoch locations provide complementary speaker-specific characteristics to the spectral features.

These glottal closure instants throughout the glottal cycles play important roles in a speaker's ability to vary source characteristics [100]. The calculation of epoch locations relies on the residual signal derived by performing linear prediction analysis of the speech signal. The energy of the residual signal is computed in small frames ($\sim 1 - 2ms$), and the point with maximum energy is hypothesized as the instant of significant excitation [34].

3.3.6 Glottal Flow

Glottal flow is the airflow arising from the trachea and passing through the vocal folds. The air flow through vibrating vocal folds serves as the excitation of the speech production mechanism. Glottal flow characteristics are highly speaker specific. The principal features of the glottal flow can be described by a number of parameters given in Figure 3.9. The parameters describing the glottal volume velocity waveform are T_0 (duration of the pitch period), t_1 (beginning of the airflow), t_2 (instant of the maximum glottal flow amplitude), t_3 (moment of the glottal closure and maximum change of glottal flow) and t_4 (instant of complete glottal closure) [101]. As shown in

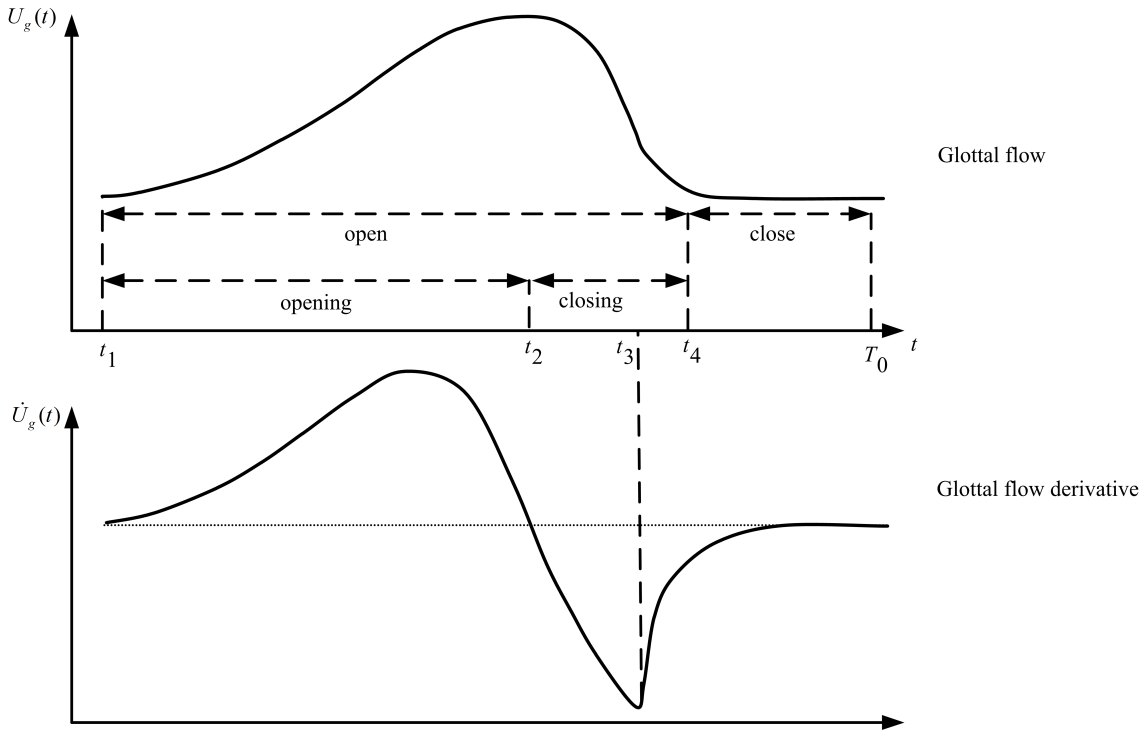


Figure 3.9: Schematic description of a glottal flow U_g and its derivative.

Figure 3.9, the glottal flow increases from t_1 to t_2, and decreases from t_2 to t_4 because of the opening and closing motions of the vocal folds.

Two important representations of the glottal flow are the open quotient (OQ) and the speed quotient (SQ), which are defined as

$$OQ = \frac{t_4 - t_1}{T_0}, \tag{3.3.7}$$

$$SQ = \frac{t_2 - t_1}{t_4 - t_2}, \tag{3.3.8}$$

where OQ is the ratio between the amount of time the vocal folds are open and the entire pitch period duration, and SQ is the ratio of rise and fall time of the glottal

flow. The open quotient represents the duty ratio of the glottal airflow. This is highly correlated to physiological constraints because a change of the duty ratio substantially changes the spectrum of an excitation. The speed quotient indicates the asymmetry characteristics of the glottal pulse. Based on the research of Rothenberg [102], the glottal flow is skewed to the right, that is, the decrease of the airflow is faster than its increase. Therefore, the glottal flow plays an important role in the analysis of speech and is different for every individual phonation type.

3.3.7 Envelope and Fine Structure

The time waveform of signals can be mathematically described as a slowly varying envelop (modulation) with rapidly varying fine-time structure (Carrier) [103], as shown in Figure 3.10. Envelope, reflecting changing amplitude of a signal, traces the crests and troughs of a periodic signal. Fine structure reflects spectral components of sounds in the sound waveform and periodicity. From this perspective, all waves can be considered as being generated by multiplying an envelope function against a carrier signal. The calculation of a signal envelope and fine structure can be derived by the Hilbert transform as described in Section 3.4.2.1. In this work, the envelope and fine structure is used to measure the vocal source characteristics.

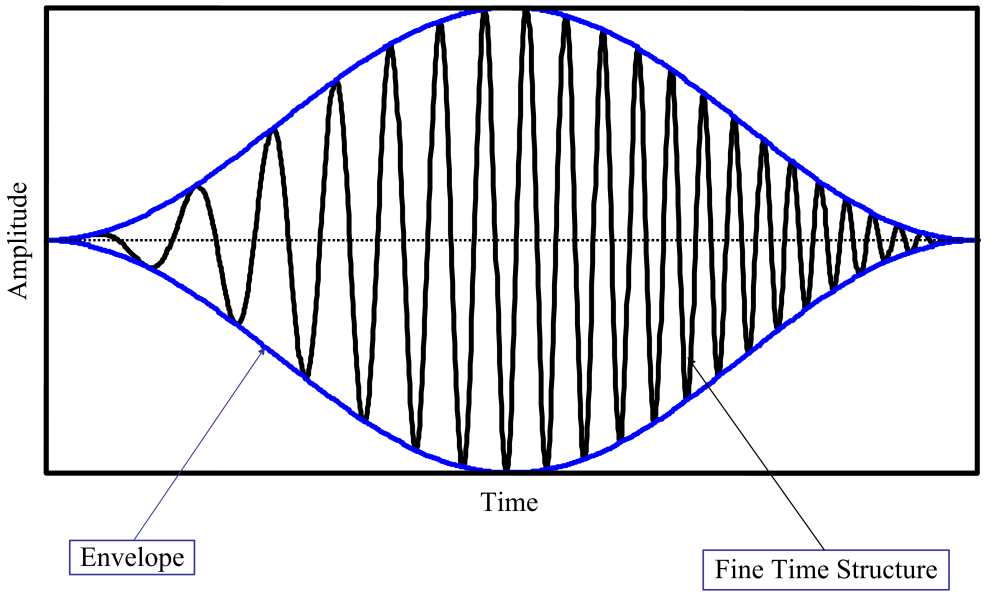


Figure 3.10: The envelope and fine structure of a signal.

3.4 Unused Information Residing in the Vocal Tract and Vocal source

The articulators are the parts of the vocal tract which are actively controlled to generate different types of sounds, including primarily the tongue, teeth and lips. Features extracted from the vocal tract carry important characteristic information for speech recognition. Different speakers have their own styles of vocal tract articulation, which generate unique variations in the speech spectrum. The habit of articulation is uniquely associated with individual speakers, and features derived from the spectrum of speech have thus proven to be effective in speaker recognition.

The vocal source contains much of the unique physiological properties of a speaker's speech production. The vocal source signal represents the musculature and tension of the vocal folds, and the associated glottal pulse parameters, including the rate of the closing phase and the degree of the opening. The vibratory pattern of the vocal folds not only produces a voicing source for speech production, but also characterizes unique personal patterns for each speaker. The quasi-periodic motion of relevant vocal organs brings periodicity to the excitation signal, while the pulse-like epoch shapes vary among speakers as well. These characteristics are unique to a given speakers speech production system. Hence, features derived from the vocal source can also provide significant information for speaker recognition.

Feature extraction methods for capturing the vocal tract characteristics of speakers have been investigated for many years. These features for speech recognition can accurately characterize the vocal tract configuration of a speaker, and can achieve good performance in current speech and speaker recognition system. However, the usefulness of vocal source excitation related features is still significantly under-investigated. Thus a detailed exploration to create the speaker-specific information present in the vocal source excitation is necessary, which is the motivation for this research.

The above sections have introduced the mechanisms of linear and non-linear speech production models and the unique information residing in vocal tract and source. In the following sections, novel proposed features related to both vocal tract and vocal

source are introduced with the goal of utilizing this unique information.

3.4.1 Proposed Vocal Tract Features

3.4.1.1 Spectral Shimmer

The definition of shimmer in Section 3.3.4 can be modified to calculate the shimmer of spectral features. Spectral shimmer features represent the short term variation in the amplitude of spectral features.

LAR shimmer measures fluctuations of the peak-to-peak value of the log area ratio within each frame, given by

$$\text{ShimLAR}_i = \frac{\left|LAR^{(i)} - LAR^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} LAR^{(i)}}, \tag{3.4.1}$$

where $LAR^{(i)}$ is the log area ratio in the i^{th} frame.

Similarly, LPC shimmer measure fluctuations of the peak-to-peak value of linear prediction coefficients within each frame, given by

$$\text{ShimLPC}_i = \frac{\left|LPC^{(i)} - LPC^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} LPC^{(i)}}, \tag{3.4.2}$$

where $LPC^{(i)}$ is the linear prediction coefficients in i^{th} frame.

Cepstral shimmer measures fluctuations of the peak-to-peak value of mel-frequency

cepstral coefficients (MFCC) within each frame, given by

$$\text{ShimCEP}_i = \frac{\left|CEP^{(i)} - CEP^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} CEP^{(i)}}, \tag{3.4.3}$$

where $CEP^{(i)}$ is the mel-frequency cepstral coefficients in i^{th} frame.

3.4.1.2 Average Harmonic Amplitude Difference

It is known that the amplitude difference of the first two harmonics (HA1-HA2) is related to the glottal open quotient [104]. Research by Chen [105] showed that the measurement of HA1-HA2 contains discriminative information for gender classification. This definition is extended here as a new feature, the Average Harmonic Amplitude Difference (AHAD), defined as the following function of the amplitudes of the harmonics of the fundamental frequency of a given frame

$$AHAD = \frac{\sum_{i=1}^{N} |HA_i - HA_{i+1}|}{N-1}, \tag{3.4.4}$$

where HA_i, $i = 1, 2, \ldots, 6$, are the first six harmonic amplitude values of the fundamental. Figure 3.11 illustrates the definition of the harmonic amplitude difference.

Another new feature related to the glottal open quotient is the Harmonic Amplitude Shimmer (ShimHA), defined as the fluctuations of the peak-to-peak value of harmonic amplitude within each frame

$$\text{ShimHA}_i = \frac{\left|HA^{(i)} - HA^{(i+1)}\right|}{\frac{1}{N}\sum_{i=1}^{N} HA^{(i)}}, \tag{3.4.5}$$

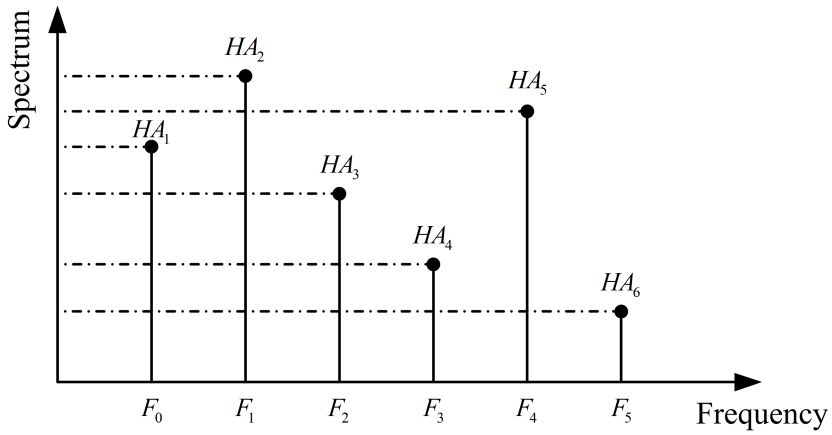


Figure 3.11: Harmonic amplitude difference.

3.4.2 Proposed Vocal Source Features

Many feature extraction methods have been developed to represent the vocal tract configurations of a speaker including those discussed in Chapter 2 and Section 3.4.1. However, features related to excitation patterns are rarely used in automatic speech or speaker recognition. In particular, the usefulness of excitation-based features for automatic speaker recognition has not been thoroughly studied. Here, several novel features based on this idea will be introduced.

3.4.2.1 Residual Phase Cepstrum Coefficients (RPCC)

The dominant features for speech and speaker recognition are extracted from the magnitude spectrum of the speech signal, which represents vocal tract characteristics. However, the human ear is not "phase-deaf". Many phase changes are clearly

audible. The importance of phase information in human speech recognition has been well-established. Research in [106] investigated the role of phase information for human perception of intervocalic plosives. The authors concluded that the short-time amplitude spectra cannot exclusively specify a stop consonant. Moreover, their results indicate that the perception of voicing in stops relies strongly on phase information. It is shown in [90] that the phase spectrum can contribute as much as the magnitude spectrum to speech intelligibility if the shape of the window function is properly selected. In [107], the authors analyze the effects of uncertainty in the phase of the speech signals on the word recognition error rate of human listeners. Their results indicate that a large amount of phase uncertainty has a significant effect on the human speech recognition rate. Therefore, the phase is very important in automatic speech and speaker recognition.

The LP residual, directly related to vocal source excitation, contributes little to linguistic distinguishing of phonemes. However, the source excitation signal carries abundant speaker discriminative characteristics. Unfortunately, it is difficult to thoroughly exploit the useful information from the amplitude spectrum of LP residual for speaker recognition, as illustrated in Figure 3.12, because the LP residual signal has a flat magnitude spectrum and only the pitch harmonics can be observed. This also suggests that significant information from LP residual is contained in its phase spectrum.

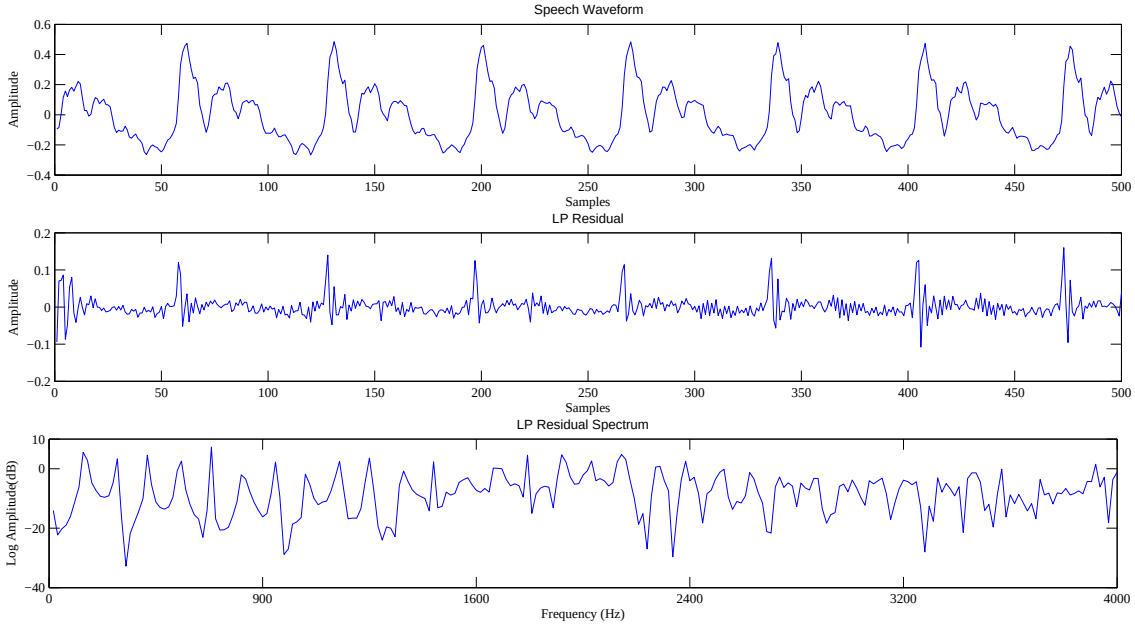


Figure 3.12: Amplitude spectrum of the LP residual.

The definition of residual phase is the cosine of the phase function of the analytic signal [108]. The analytic signal is derived from the LP residual of a speech signal. In this dissertation, the residual phase coefficients are modified by using a discrete cosine transformation to de-correlate the feature vectors, since the HMM recognition system usually assumes independence between the feature vector components. The calculation of RP is

$$r(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k), \tag{3.4.6}$$

where p is the order of prediction and a_k are the linear prediction coefficients obtained from LPC analysis.

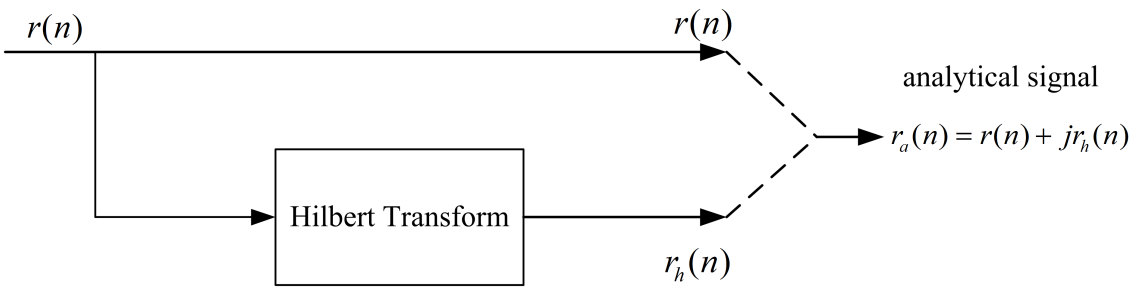


Figure 3.13: Calculation of analytical signal from a real signal.

The analytical signal $r_a(n)$ is given by

$$r_a(n) = r(n) + jr_h(n), \tag{3.4.7}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_h(n) = \begin{cases} IDFT[-jR(\omega)], & 0 < \omega < \pi \\ -IDFT[-jR(\omega)], & -\pi < \omega < 0 \\ 0, & \omega = 0, \pi \end{cases} \tag{3.4.8}$$

where $R(\omega)$ is the discrete Fourier transform of $r(n)$ and $IDFT$ denotes the inverse discrete Fourier transform. The generation of an analytical signal from a real signal is shown in Figure 3.13.

The cosine of the phase information is calculated by the following equation:

$$\text{ResidualPhase} = \frac{R_e(r_a(n))}{|r_a(n)|} \tag{3.4.9}$$

The above calculation of residual phase can be explained as the calculation of the fine structure of the LP residual signal, as shown in Figure 3.14. The residual signal is transformed into an envelope (right upper panel) and temporal fine structure

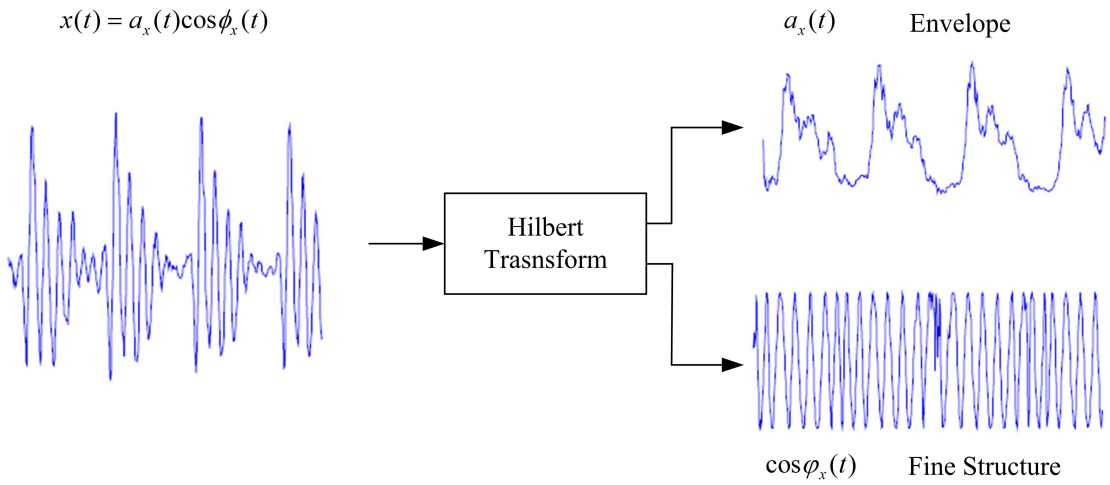


Figure 3.14: The Hilbert transform.

(right lower panel). The original residual signal can be reconstructed by multiplying the envelope with the fine structure. The envelope effectively codes speech. The fine structure contains temporal pitch and timbre information and is the primary source of interaural timing information (differential arrival time of a sound between the two ears). From this point of view, the residual phase provides useful information about the fine structure of the LP residual for speaker recognition.

The rest of the processing is similar to the calculation of the cepstrum of residual phase. The final DCT step is to de-correlate the feature vectors,

$$RP = DCT(\cos(\theta_n)). \tag{3.4.10}$$

The feature extraction process of residual phase cepstrum coefficients (RPCC) is

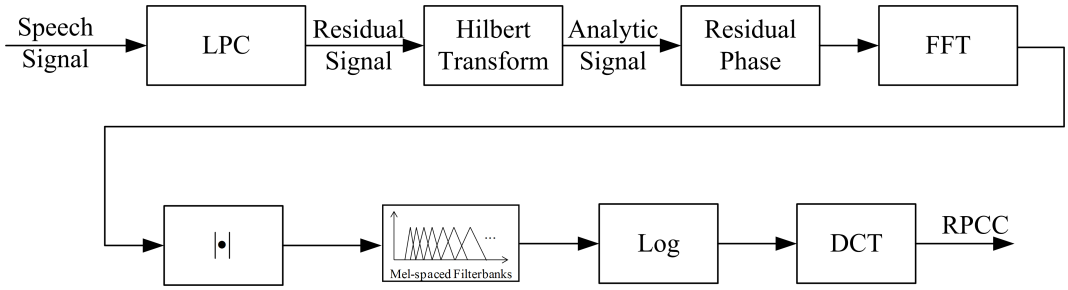


Figure 3.15: Calculation of residual phase cepstrum coefficients.

shown in Figure 3.15. The residual phase was introduced in [108] and directly implemented as a complementary feature to MFCCs in a speaker recognition system. Here, a mel-spaced cepstral analysis is first implemented on the residual phase. The magnitude spectrum of the residual phase is computed and warped to the Mel frequency scale, followed by the usual log and DCT to obtain RPCC. Compared to residual phase, this novel feature represents the spectral shape of residual phase more compactly and provides significant information about a speaker with fewer parameters.

3.4.2.2 Teager Phase Cepstrum Coefficients (TPCC)

Most features for speaker recognition are derived from a linear source-filter speech production model, assuming that the airflow propagates in the vocal tract as a plane wave. According to research by Teager as discussed in Section 3.2.5 [80], however, the air flow actually consists of separate and simultaneous vortices distributed throughout the vocal tract. The Teager energy operator (TEO) provides an advantage over

Fourier analysis methods in capturing the characteristics of nonlinear systems, because it measures the energy in the system that generated the signal rather than the energy of the signal itself. The underlying concept of TEO is that while the vocal tract articulators do move to configure the vocal tract shape, the resulting airflow properties serve to excite those models, which a listener will perceive as a particular phoneme. Hearing can then be viewed as the process of detecting the energy. In this nonlinear approach, the energy required by a speech production system to generate a signal at high frequencies is expected to be much higher than the energy required by the system to generate a signal at low frequency. This energy is potentially related to the pitch information of different speakers.

The relevant amplitude-frequency modulation patterns in speech resonances are investigated by the research of Teager [80]. Work by Maragos [84] also introduces this method as an appropriate tool for speech analysis. TEO is applicable to analyze and estimate the characteristics of the existing amplitude and frequency modulation patterns in a vocal excitation signal. This is intuitively a favorable approach for investigation of the vibration characteristics yielded by the vocal folds for potential speaker-specific features.

The Teager energy operator (TEO), described in Section 3.2.5, was introduced by Teager and extended by Kaiser [83]. It measures the energy required to generate a signal in the human speech production system. In this dissertation, features derived

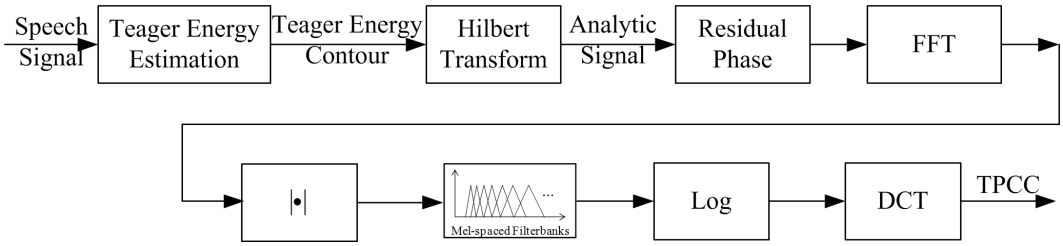


Figure 3.16: Calculation of Teager phase cepstrum coefficients.

from the TEO are proposed to reflect properties of the speech production process that are not covered by features derived from the linear model of speech production. Figure 3.16 shows a block diagram of the proposed Teager Phase Cepstrum Coefficients (TPCC). In this feature extraction method, the excitation energy to generate the speech signal is calculated through the Teager energy operator. Then the fine energy structure is obtained by a Hilbert transformation, and the cepstrum of the fine energy structure is computed and warped to the Mel frequency scale followed by the usual log and DCT, to obtain TPCC.

3.4.2.3 Glottal Flow Cepstrum Coefficients (GLFCC)

As introduced in Section 3.3.6, phonation type strongly influences the properties of a generated sound, with a unique glottal flow characteristic for every individual phonation type [79]. Videos of vocal fold vibration [109] show large variations in the movement of the vocal folds from one speaker to another. For some individuals the vocal folds never close completely (soft voice) and in other cases vocal folds close

completely and rapidly (hard voice). The manner and speed of vocal fold closing during voicing is also speaker dependent. The closure of vocal folds for some individuals shows a zipper-like pattern, while others close along the length of the vocal folds about the same time. The spectral content of the glottal source is determined by the speed of glottal closure, since a fast glottal closure acts like an impulse and generates a source with a wide bandwidth. A slow glottal closure will cause less energy to be present at higher harmonics compared to the fundamental.

Additionally, the configuration of the area of the opening shows differences for different individuals [101, 110]. The glottal opening for some individuals is approximately equal in width along the length of the glottis, such as pressed phonation. This leads to both ends of the vocal folds being held close together, creating an opening shaped like a football. For some individuals, a more triangle shaped opening will occur, according to their own anatomical vocal fold structure.

The duration of vocal fold opening and closing, the glottal closing instants (GCIs) and glottal opening instants (GOIs), and the shape of the glottal flow vary significantly across speakers. These variations correspond to the variations in the glottis, and then are reflected in the glottal flow. Therefore, the glottal flow contains speaker distinctive information and features derived from glottal flow are expected to be useful for speaker recognition.

The accurate estimation for the glottal flow has been a target of speech research

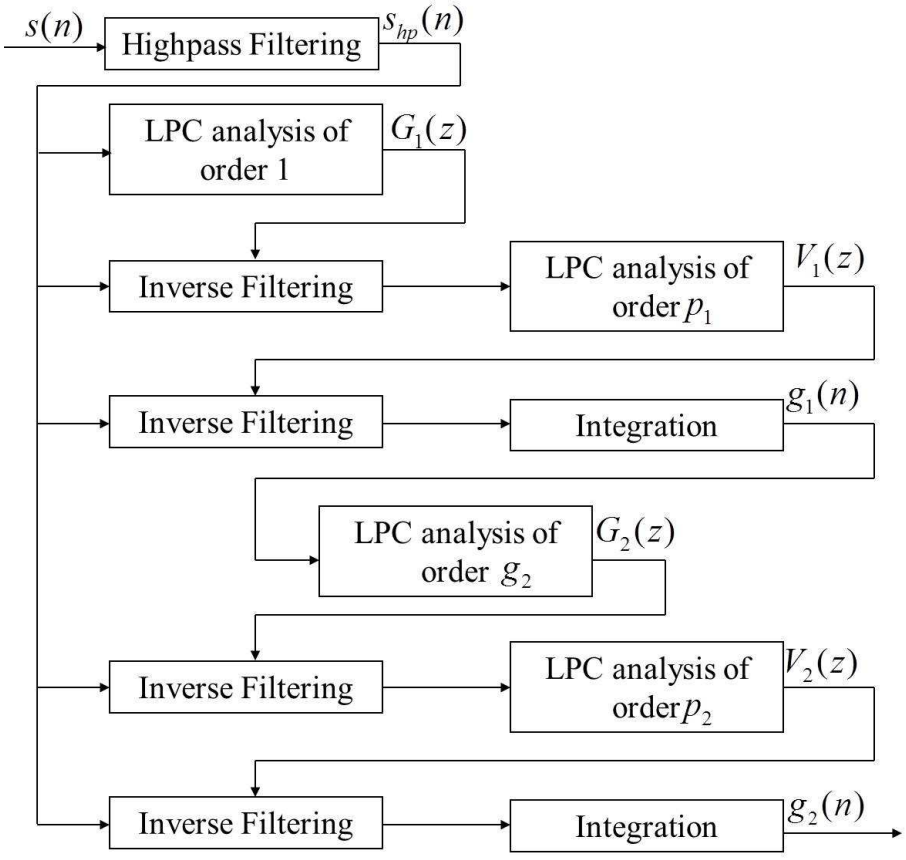


Figure 3.17: Structure of the IAIF algorithm for computing glottal flow.

for several decades, and many different methods have been developed. Among these methods, Pitch Synchronous Iterative Adaptive Inverse Filtering (PSIAIF) [111] has been shown to be an efficient method in the estimation of the glottal flow. A flow diagram of PSIAIF method is shown in Figure 3.17.

The block diagram of the IAIF algorithm is displayed in Figure 3.17, detailed as follows:

1. High-pass filtering of the speech signal $s(n)$ removes the low frequency fluctuations captured by the microphone.

2. LPC analysis of order 1 models the gross effect of the glottal source by $G_1(z)$.

3. The estimated glottal effect is eliminated by filtering $s_{hp}(n)$ through $G_1(z)$.

4. The first estimate of the vocal tract is computed by applying LPC analysis of order p_1 $(8 \sim 14)$.

5. The effect of the vocal tract is eliminated from the signal $s_{hp}(n)$ by inverse filtering.

6. The lip radiation effect is removed by integration in order to obtain the first estimate of the glottal excitation $g_1(n)$.

7. LPC analysis of order p_2 $(2 \sim 4)$ models the glottal source by $G_2(z)$.

8. Filtering of the speech signal by the inverse of $G_2(z)$ leads to a signal containing the effect of the vocal tract.

9. The final model of the vocal tract is obtained by applying LPC analysis of order p_1 $(8 \sim 14)$.

10. The effect of the vocal tract is eliminated from speech by filtering $s_{hp}(n)$ through $V_2(z)$.

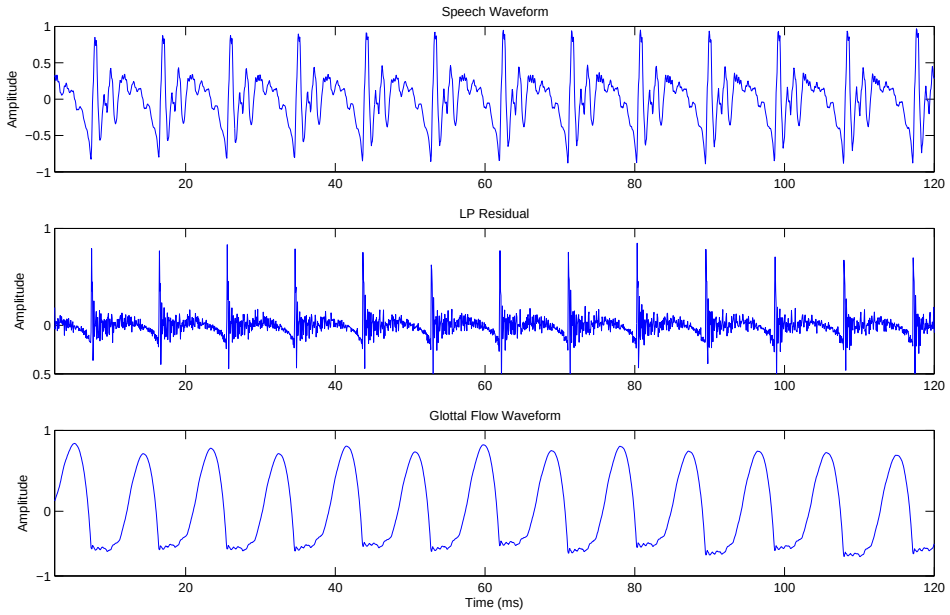


Figure 3.18: A diagram of glottal inverse filtering.

11. The final glottal waveform estimate $g_2(n)$ is obtained by cancelling the lip radiation effect by integration.

The Iterative Adaptive Inverse Filtering (IAIF) algorithm is used to estimate the glottal waveform of speech signal by filtering the original speech signal using an inverse model of the vocal tract filter, modeled as an all-pole system [111]. An example of glottal inverse filtering is shown in Figure 3.18. The upper figure is a speech signal, the middle is the LP residual and the lower is the IAIF output.

The GLFCC feature introduced here performs mel-spaced cepstral analysis on glottal flow as shown in Figure 3.19. The IAIF method helps in separating the

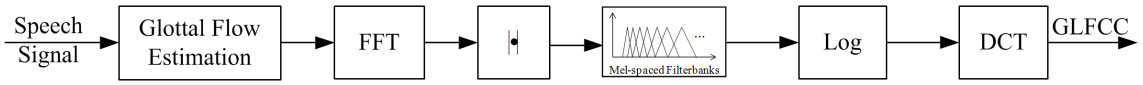


Figure 3.19: Computation of glottal flow cepstrum coefficients.

source and filtering related information. The magnitude spectrum of the glottal flow, similar to the process of RPCC feature extraction, is computed and warped to the Mel frequency scale followed by the usual log and DCT to obtain the GLFCC.

CHAPTER 4
SPEAKER RECOGNITION EXPERIMENTS

We have described in Chapter 3 the details of several feature extraction methods for speaker recognition. We also argued that features captured from the LPC residual and excitation waveform can be expected to be useful features, capturing unique characteristics of a speaker. In this chapter, speaker identification systems are developed to evaluate the effectiveness of the proposed features. Section 4.1 describes a cross-phoneme experiment paradigm designed to test for features that identify physiological rather than phonetic characteristics of speakers. Section 4.2 describes the experimental results of a cross-language speaker identification using those same novel feature sets. Section 4.3 describes the experimental results of a cross song-type avian individual identification experiment, again using these novel feature sets. Section 4.4 assesses the effectiveness of proposed feature set in mono-language environment.

4.1 Preliminary Cross-phoneme Speaker Identification Experiments

This preliminary experiment describes a unique cross-phoneme speaker identification experiment, using deliberately mismatched phoneme sets for training and testing. The

purpose is to identify features that represent broad individually unique characteristics rather than those that represent phonetic differences. The proposed features for this experiment include log area ratio (LAR), fundamental frequency, jitter, shimmer, shimmer of the LAR (ShimLAR), shimmer of the LPC coefficients (ShimLPC), harmonic amplitude difference (HAD), average harmonic amplitude difference (AHAD), RPCC, TPCC, GLFCC and some statistical feature calculations. Of these, RPCC, TPCC, GLFCC, HAD, AHAD, ShimLAR, and ShimLPC have not been previously investigated for speaker identification. The baseline features used for comparison are MFCCs.

4.1.1 Data Corpus

The TIMIT corpus contains recording of phonetically-balanced prompted English speech [112]. It was recorded using a close-talking microphone at a 16 kHz sampling rate with 16 bit resolution. This corpus contains a total of 6300 sentences, consisting of 10 phonetically-rich sentences spoken by each of 630 speakers representing 8 major divisions of American English. All sentences were manually partitioned into the phoneme level. This speech corpus has been a standard database for speech community for several decades and is still widely used for both speech and speaker recognition experiments.

In order to support the goal of evaluating phonetically-independent features for

speaker identification, a cross-phonetic experimental paradigm has been designed. The data used in this experiment was extracted from the TIMIT database. In this experiment, using the well-known vowel triangle in the F1/F2 feature space, vowel sets with minimal spectral similarity were selected for training and testing sets. The phoneme set is divided primarily on the basis of the overall vowel height (i.e. primarily correlated with formant F1). For training low vowels {/ae/, /aa/, /ah/, /eh/, /ao/} are used, while the phoneme set for testing includes the high vowels {/iy/, /ih/, /er/, /ow/, /uh/, /uw/}. A subset of 25 speakers within the same dialect region from the TIMIT corpus is used for evaluation, leading to an overall data set consisting of 1621 phoneme utterances. Phonemes with fewer than 700 samples are discarded in order to ensure accuracy of fundamental frequency related and statistical features.

4.1.2 Experimental Setup

In this speaker identification system under investigation, each speaker enrolled in the system is represented by 128 Gaussian mixture components in the training phase, as discussed in Section 2.6.1.1. In the testing phase, the log-likelihood scores of the incoming sequence of feature vectors are computed from each speaker model by

$$L(X, G_S) = \sum_{t=1}^{M} P(\vec{x}_t | \, G_S), \tag{4.1.1}$$

where $X = \{\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_M\}$ is the sequence of speaker feature vectors, and M is the total number of feature vectors. The GMM generating the highest $L(X, G_S)$ score is

selected as the most likely target speaker. The programming toolkit HTK 3.4.1 from Cambridge University is used for all training and testing implementations [113].

4.1.3 Experimental Results

Each individual feature was evaluated for identification accuracy, in comparison to a baseline using a 22-dimensional MFCC feature vector. Various combinations of feature sets were then implemented to assess the impact of using these features in combination.

The results of each individual feature are shown in Table 4.1, while the results of different feature combinations are shown in Table 4.2. Individually, the best overall feature is the LAR feature at 39.5% accuracy, followed by GLFCC at 26.9%, which is also the highest non-spectral feature.

The highest total accuracy of 56.1% is obtained by a combination of all the proposed features, with a feature dimension of 63. This compares to a 26.0% accuracy obtained using the baseline MFCC features, with an absolute accuracy increase of 28.1%. In all combinations, there is improvement in accuracy as a result of adding TPCC, RPCC and GLFCC.

Table 4.1: Classification accuracy on cross-phoneme experiment.

Feature Category	Individual Feature$^{(\text{Dimension})}$	Accuracy(%)
Vocal Tract Features	ShimLAR$^{(1)}$	6.0
	ShimLPC$^{(1)}$	6.3
	AHAD$^{(1)}$	9.0
	HAD$^{(6)}$	17.2
	MFCC(baseline)$^{(22)}$	26.0
	LAR$^{(22)}$	39.5
Vocal Source Features	Shimmer$^{(1)}$	8.9
	TPCC$^{(22)}$	11.5
	Jitter$^{(1)}$	14.6
	F0$^{(1)}$	16.6
	RPCC$^{(22)}$	22.4
	GLFCC$^{(22)}$	26.9

Table 4.2: Classification results for different feature combinations.

Combined Feature$^{(\text{Dimension})}$	Accuracy(%)
Jitter + Shimmer + F0 + ShimLPC & ShimLAR + AHAD + HAD + Statistic Features$^{(19)}$	29.3
Jitter + Shimmer + F0 + ShimLPC & ShimLAR + AHAD + HAD + Statistic Features + LAR$^{(41)}$	48.5
Jitter + Shimmer + F0 + ShimLPC & ShimLAR + AHAD + HAD + Statistic Features + LAR + TPCC$^{(63)}$	49.1
Jitter + Shimmer + F0 + ShimLPC & ShimLAR + AHAD + HAD + Statistic Features + LAR + GLFCC$^{(63)}$	55.0
Jitter + Shimmer + F0 + ShimLPC & ShimLAR + AHAD + HAD + Statistic Features + LAR + RPCC$^{(63)}$	56.1
Jitter + Shimmer + F0 + ShimLPC & ShimLAR + AHAD + HAD + Statistic Features + LAR + + RPCC + GLFCC + TPCC$^{(107)}$	54.1
Jitter + Shimmer + F0 + ShimLPC & ShimLAR + AHAD + HAD + Statistic Features + LAR + RPCC + TPCC + GLFCC + MFCC$^{(129)}$	46.9

4.1.4 Experimental Summary

It is interesting to observe that, despite the fact that training and testing sets were deliberately chosen to minimize phonetic similarity, the basic spectral measures of MFCCs and LARs still outperform excitation measures. One key observation is that LARs seem to be a substantially better spectral feature for the purposes of phonetically-independent speaker identification, and that the excitation-related measures are at least able to contribute to an overall increase in accuracy when combined with LAR and the proposed feature set. Of the excitation-related measures, TPCC, RPCC and GLFCC clearly have the strongest individual component because of their complimentary characteristics to other features. These features are selected based on their strong performance for further investigation.

4.2 Cross-language Speaker Identification Experiments

A great many speaker recognition tasks, such as forensic, military and intelligence applications, involve speech material from more than one language. It is difficult to reliably identify speakers if the training and testing utterances represent different language conditions [114]. This scenario deteriorates the performance of speaker recognition system unless properly compensated, since the current feature vectors are

composed of acoustic parameters that are derived from the resonance characteristics of vocal tract cavities. These traditional acoustic feature vectors mainly represent the characteristics of phonetic context. This causes significant concern because the cross-language speaker identification task has a phonetic mismatch between the training and testing data.

In order to address the degradation of speaker identification performance under a mismatched language scenario, one potential compensation approach is to focus on speaker-discriminative features such as vocal source features, in conjunction with traditional acoustic type features such as those introduced in Chapter 3, to rapidly capture the vocal source characteristics of the speaker. Hence, the goal of this cross-language experimental configuration is to evaluate the effectiveness of our proposed vocal source features for this purpose.

4.2.1 Data Corpus

The NIST SRE 2004 corpus considered some of the factors affecting performance of text-independent speaker recognition system when designing its speech corpus collection [115]. This allows investigation of the effects of language, transmission type, and handset type on the recognition performance. The evaluation was designed for all trials to involve the use of different handsets in the training and test segment data. The training and testing segments in this corpus are continuous conversational

excerpts, with no prior removal of silence intervals.

For the cross-language speaker identification experiment, bilingual speaker data were extracted from 2004 NIST SRE corpus. Since 2004, a special effort has been made to recruit bilingual speakers who can speak Arabic, Mandarin, Russian or Spanish in addition to English. Hence a considerable percentage of the calls from the bilingual speakers are in a language other than English. This corpus was originally collected to evaluate the effect of language, particularly differences between training and testing language, on speaker recognition systems, although the main task of the 2004 NIST SRE corpus involves speaker verification. In this work, the data of twenty-four bilingual speakers were extracted from this corpus and used as the basis for a cross-language speaker identification experimental framework, in order to evaluate the effectiveness of the discriminatively vocal source features proposed here.

4.2.2 Experimental Setup

The classification framework in this dissertation is based on a Gaussian Mixture Model and Universal Background Model (GMM-UBM) approach [40], commonly used in the speech processing community to perform speech and speaker recognition. This approach has advantages in its flexibility and robustness to duration and temporal

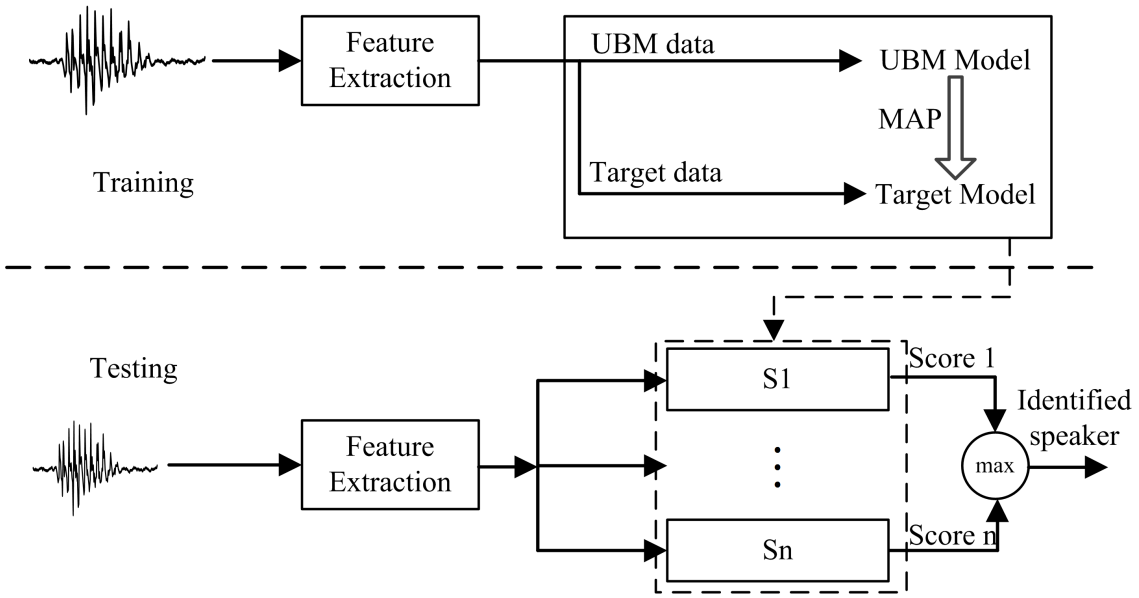


Figure 4.1: Block diagram of the GMM-UBM SlD system.

alignment differences between training and testing examples. The UBM is a speaker-independent GMM trained with speech samples from a large set of speakers to represent general speech characteristics. The hypothesized speaker model is derived from the UBM using Maximum A Posteriori (MAP) adaptation with the corresponding speech samples from a particular enrolled speaker. The strategy of adapting the target speaker model is based on the similarity between the enrollment data of target speaker and UBM, adjusting the UBM to the speaker training data. During adaptation, the distributions of the UBM which are far from the feature of target speaker remain almost unchanged. The block diagram of a speaker identification system based on GMM-UBM is showed as Figure 4.1.

In this experiment, a UBM is trained using data from all twenty-four non-English speakers in the NIST corpus, representing 8 Arabic speakers, 7 Mandarin speakers, 6 Russian speakers, and 3 Spanish speakers. The total number of samples for initial training is 262, with an additional 260 samples from the target speakers used for identification. There is an average of 8 speech samples per speaker, with an average length of about two minutes. Each target speaker's model is adapted from the global model using the individual English language speech samples, and the identification is performed using their alternative language speech samples.

For comparison, MFCCs are used as the baseline feature. The analysis window size is 12.5ms with an overlap of 6.25ms. Twenty MFCCs are calculated and an LPC order of 22 is used to calculate the residual phase. The LPC residual is used to calculate RPCC features as described in the previous section, with a matching RPCC dimension of twenty.

Two comparison experiments have been implemented. The first experiment focuses on evaluating the performance of each individual feature as a function of the number of mixture components. The second is to evaluate the effectiveness of the baseline feature in conjunction with the proposed source features, also against an increasing number of mixture components. By considering performance as a function of model complexity, this experiment is intended to reveal whether the proposed vocal source features have a more direct relationship to physiological distinctiveness, as

opposed to the relatively large complexity associated with phonetic differences.

4.2.3 Experimental Results

4.2.3.1 Accuracy of individual feature versus increasing number of mixtures

Figure 4.2 shows the classification accuracy of each individual feature as a function of an increasing number of mixtures. The proposed source features give higher performance for all low complexity models up to 8 mixture components. Above that, the MFCC features give better performance, suggesting that once there is sufficient model complexity the amount of total information in the MFCC features relevant to speaker is higher than that of RPCC and TPCC. GLFCC, in particular, performs better than MFCC across all numbers of mixtures. These experimental results support the idea that the proposed features are more compact with less dependence on phonetic content, needing a smaller number of model parameters to represent the information for each speaker.

4.2.3.2 Accuracy of combined features versus increasing number of mixtures

Figure 4.3 shows the performance of the combined feature versus the increasing number of mixture components. The results clearly show the deterioration of performance using MFCC features with a lower model complexity, since the MFCC features need

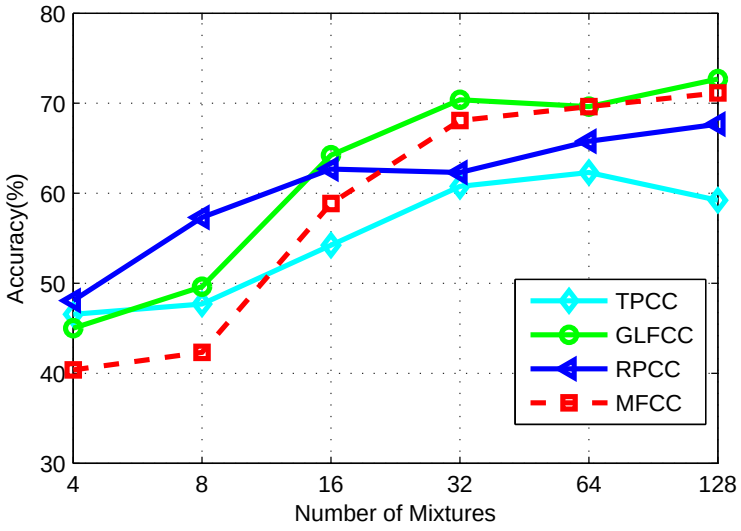


Figure 4.2: SID performance on a cross-lingual NIST task using individual features and varying number of Gaussian mixture components.

more coverage of phonetic feature space. Comparing Figure 4.2 and Figure 4.3, it is also clear that the vocal tract and source features appear to be complementary. This suggests that a speaker identification system which combines both vocal tract and source features will offer further improvements. This effect is most likely due to an under-representation of the intra-speaker variance using standard MFCC features with a lower model complexity.

The proposed vocal source features in conjunction with the traditional MFCC features outperform MFCC features alone, with maximal improvement at 8 mixtures converging to relatively small improvement with a larger model, as shown in Table

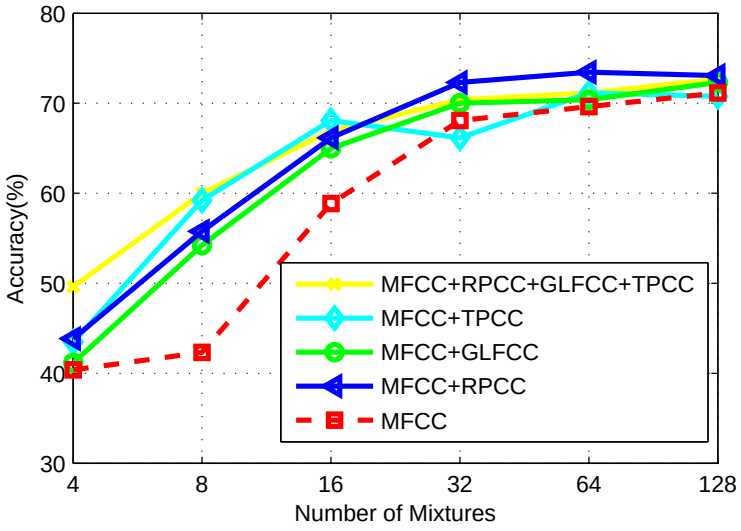


Figure 4.3: SID performance on cross-lingual NIST task using the combined features and a varying number of Gaussian mixture components.

4.3 and Table 4.4. From Table 4.3, we can see the large difference in system performance using all of the combined features under relatively low model complexity. This observation is significant for many practical applications of speaker identification where data is limited. A negligible deterioration of system performance is observed as the model complexity increases to 128 mixtures. It is interesting to note that the combination with the greatest increase in accuracy at 8 mixtures is given by the same combination of features with the least impact at 128 mixtures, the TPCC/MFCC combination.

Another important observation is that the performance difference between the baseline feature and combination feature set decreases gradually as the number of

Table 4.3: Accuracy improvement with 8 mixtures.

Feature Combinations	Accuracy Differential
MFCC+RPCC	13.5
MFCC+GLFCC	11.9
MFCC+TPCC	16.9
MFCC+RPCC+GLFCC+TPCC	17.7

Table 4.4: Accuracy improvement with 128 mixtures.

Feature Combinations	Accuracy Differential
MFCC+RPCC	1.93
MFCC+GLFCC	1.2
MFCC+TPCC	-0.38
MFCC+RPCC+GLFCC+TPCC	1.5

mixture components increases. This is due to the fact that MFCC features gradually become dominant components on the combined feature set with increasing model complexity, slowly building a more comprehensive and accurate speaker model based on phonetics. This demonstrates that combined vocal source/vocal tract feature sets are significantly more robust to data size and data variation.

4.2.4 Experimental Summary

The experimental results confirm that the proposed vocal source feature set provides discriminative information about speaker characteristics that is significantly different

in nature from the phonetically-focused information present in traditional speaker identification features such as MFCCs. These new features give better identification accuracy with lower model complexities, and also provide complementary information to the traditional acoustic feature that can improve overall system performance. The fact that these new features are less dependent on the phonetic content of the speaker makes them useful for practical speaker identification application, in which the computational model complexity and efficiency is an important indicator.

4.3 Individual Identification Experiments in Cross Song-type Avian Data

Current approaches to automatically analyzing and classifying animal vocalizations are based on integrating successful models and ideas from the field of speech processing and recognition into bioacoustics. One significant application in this domain is that of bioacoustic censusing based on individual identification [116–118]. The performance of this automated methodology largely depends on successful extraction of relevant call-independent features that give information about identity without dependence on knowledge of call repertoires. Traditional features based on cepstral coefficients [119] typically need a large amount of data to train an individually sophisticated identification model in order to cover the wide vocalization content of

the species, just as with humans. In addition, prior knowledge of the vocalization repertoire categorization is necessary for individual and species identification, but these are not well established or understood for all species. Hence, features need to be less dependent on the vocalization categories of the species to be useful for tasks of bioacoustic censusing, where speaker identification is used to count and monitor animal populations within a species.

In this work, the proposed vocal source feature extraction methods include RPCC, GLFCC and TPCC, which capture characteristics from speakers' excitation rather than vocal tract characteristics and are more compact across a wide range of vocalization conditions. The goal of these alternative features is to rapidly capture the characteristic physiological features of a species, requiring less complex models and greatly increasing the efficiency and performance for cross song-type identification tasks. The features proposed here are mainly based on the human speech production system. However, the vocal organ that birds use to produce the vast majority of vocalizations is different in location and anatomy structure from that of humans. The primary sound source mechanism in birds is called the syrinx, and contains two independently controlled sound sources located in the cranial end of each primary bronchus. Air from the main anterior air sacs is forced through the syrinx, where it vibrates the tympaniform membranes on either side of the syrinx. Sound is generated by the vibration of the air column as it passes through the narrow passageways of

the syrinx. Syringeal muscles control the fine details of the syrinx action during song production by changing the air pressure, the tension of the membranes and the shape of the syrinx. The syrinx contains two independent halves. Thus, birds can produce two complex song-types simultaneously. The two sources can also be coupled together to form a single complex sound. Although the avian anatomy is different from the human, the avian song production can also be considered as a source-filter model, similar to humans. On this basis, the proposed features can also be implemented as a species-distinctive feature for avian species.

4.3.1 Data Corpus

The ortolan bunting, *Emberiza hortulana*, is classified as an endangered migratory passerine distributed from Western Europe to Mongolia [120], which has undergone a major population decline both in individual number and in their distribution. The species occurs in easily accessible agricultural areas where they nest on raised peat bogs, clear-cut forest on poor sand, and cleared farmland and forest burn [121].

The ortolan bunting vocalizations in this study were collected from County Hedmark, Norway in May of 2001 and 2002. All recordings were sampled at 48 kHz with 16 bit quantization. This species vocalizes in a frequency range between 1.9 kHz and 6.7 kHz, and has a relatively simple song and a repertoire size of typically 2-3 song-types for each individual. Songs are categorized in terms of their syllables, song-types

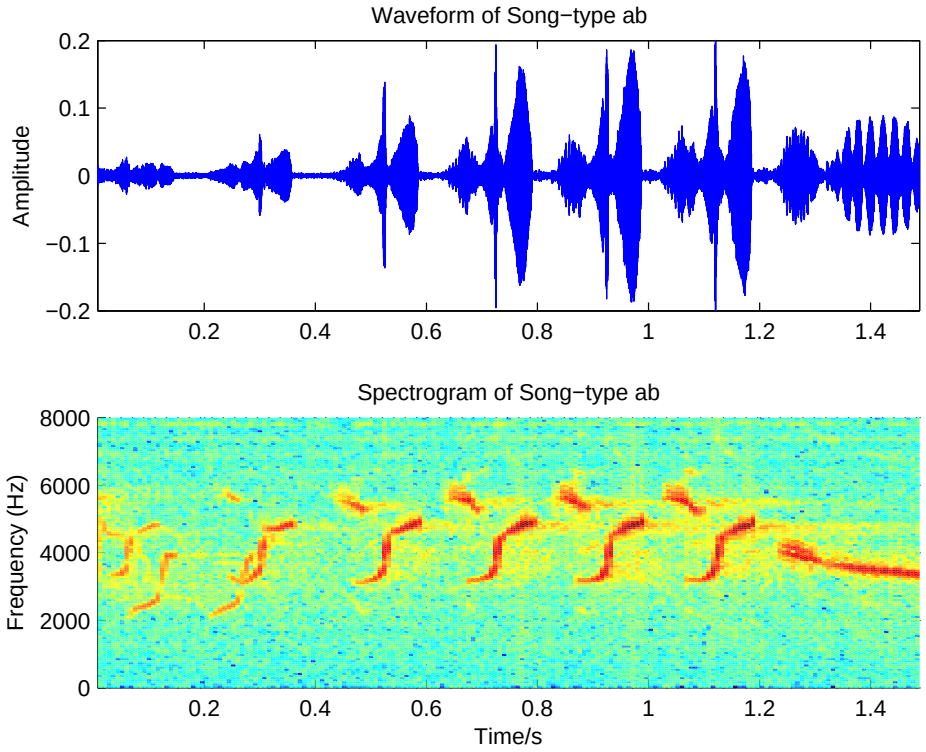


Figure 4.4: Waveform and spectrogram for song-type *ab*.

and song variants. In total, the data corpus contains 63 different song types and 234 different song variants, composed of 20 different syllables. In this dissertation, the most frequent song-types within the studied population were selected for the experiment, with the 'ab' call type used for training and the 'cd' call type used for testing. Example waveforms and corresponding spectrograms of the song-type 'ab' and 'cd' are shown in Figure 4.4 and Figure 4.5.

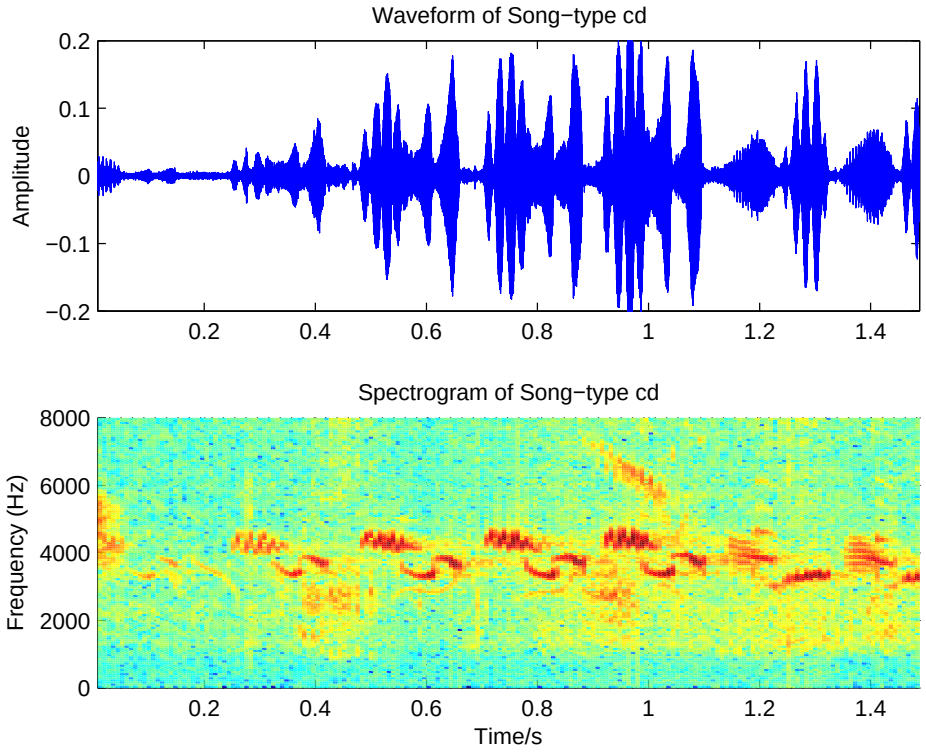


Figure 4.5: Waveform and spectrogram for song-type *cd*.

Table 4.5: Data for cross song-type speaker identification.

Bird ID	Song-type	No. of Songs for Training	Song-type	No. of Songs for Testing
1	ab	31	cd	57
2	ab	59	cd	45
3	ab	95	cd	35
4	ab	110	cd	80
5	ab	95	cd	41
Total		390		258

4.3.2 Experimental Setup

The goal of this experiment is to measure the impact of song-type on an avian speaker identification system and evaluate the proposed features for this task. The main idea of this experiment configuration is similar to cross-language human speaker recognition. The enrollment model is trained on one song type, while verification is done on another different song type. Table 4.5 shows the data used for this cross song-type individual identification. For this cross song-type speaker identification task, the call type 'ab' is used to train the statistical model of each individual, and the call type 'cd' is employed to test.

4.3.3 Experimental Results

4.3.3.1 Accuracy versus increasing number of mixtures (GFCC & RPCC)

Figure 4.6 shows classification accuracy as a function of an increasing number of mixtures using all available training data. This result also supports the idea that RPCC features are more compact, needing a smaller number of model parameters to represent the information for each speaker. The RPCC features show a good performance for all small model sizes up to 128 mixtures. Overall, the GFCC features give better performance, suggesting that once there is sufficient model complexity and training data the amount of total information in the GFCC features relevant to speaker identification is higher than that of RPCC. However, the combination of GFCC and RPCC shows the best performance. This suggests that RPCCs provide complimentary feature information on each individual's vocal source to the vocal tract related feature GFCC.

4.3.3.2 Accuracy with increasing amount of training time (GFCC & RPCC)

Figure 4.7 shows the classification accuracy of GFCC and RPCC features against an increasing amount of training time with 128 mixtures. Results give a similar result to the previous experiment in that RPCC features are more compact with less dependence on phonetic content, showing higher accuracy in the 1 to 15 second conditions. GFCC features show better performance with increasing amount of data, indicating

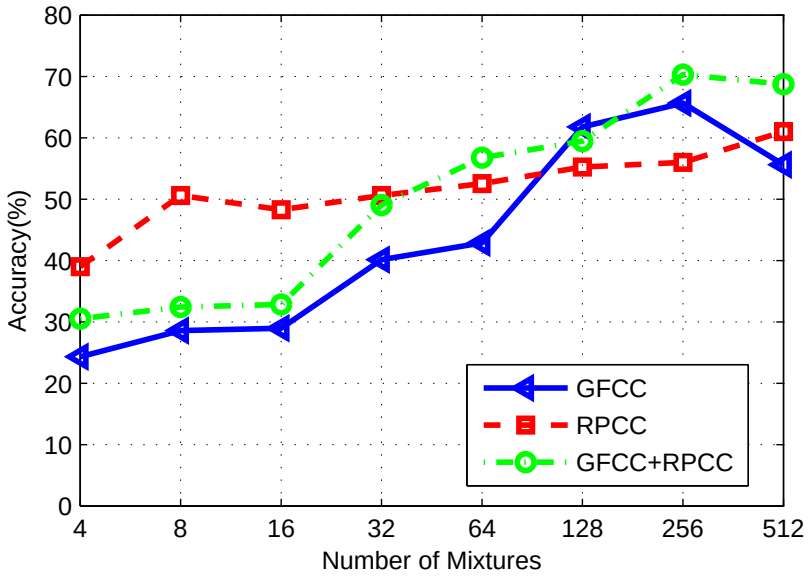


Figure 4.6: Accuracy versus increasing number of mixtures (GFCC & RPCC).

that the additional spectral information contained in the GFCC gives better overall discriminability once enough information is available to train the models.

4.3.3.3 Accuracy versus increasing number of mixtures (GFCC & TPCC)

A comparison of the classification accuracy between the baseline GFCC, and the proposed feature TPCC is shown in Figure 4.8. GFCCs show better performance than TPCCs, while the combination of GFCCs and TPCCs gives substantially better results than using them individually.

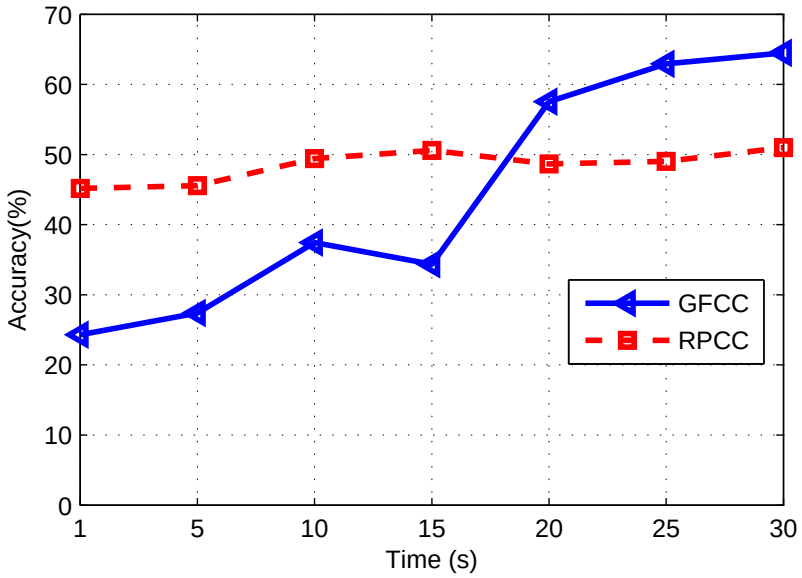


Figure 4.7: Accuracy versus duration of enrollment data (GFCC & RPCC).

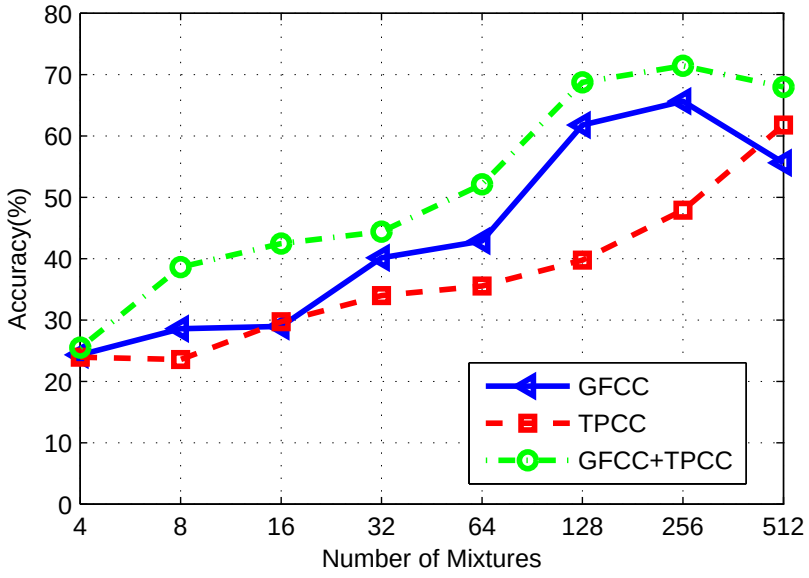


Figure 4.8: Accuracy versus increasing number of mixtures (GFCC & TPCC).

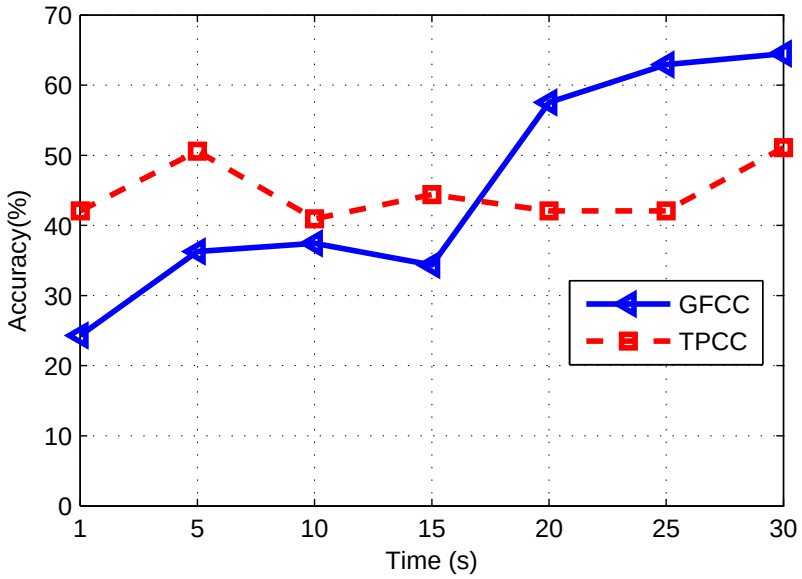


Figure 4.9: Accuracy versus duration of enrollment data (GFCC & TPCC).

4.3.3.4 Accuracy with increasing amount of training time (GFCC & TPCC)

The accuracy versus duration of enrollment data between GFCC and TPCC is shown in Figure 4.9. The accuracy of TPCC shows good performance even with a short duration (1-15 seconds data) of training data.

4.3.3.5 Accuracy versus increasing number of mixtures (GFCC & GLFCC)

The accuracy versus an increasing number of mixtures for GFCC and GLFCC features both individually and in combination is shown in Figure 4.10. The proposed GLFCC feature shows better performance with an increasing number of mixtures compared to the baseline GFCC features. GLFCCs, in particular, have good accuracy even with

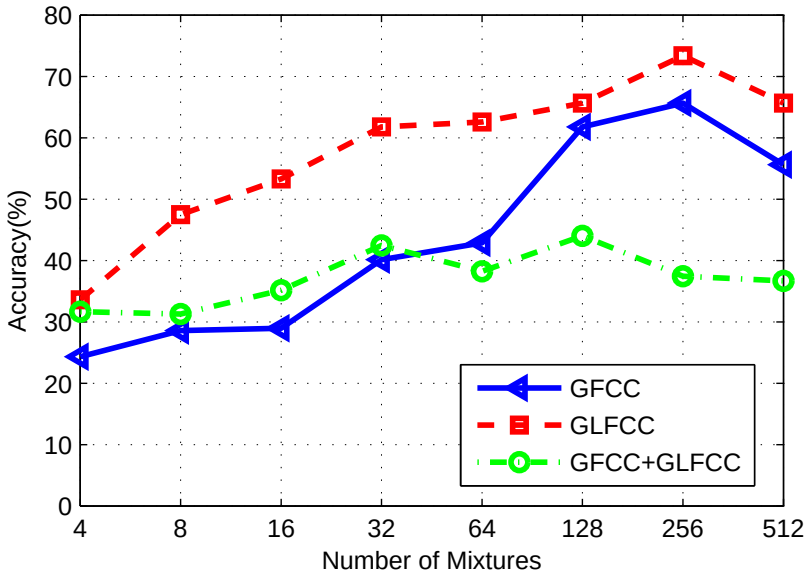


Figure 4.10: Accuracy versus increasing number of mixtures (GFCC & GLFCC).

lower model complexity (small number of mixtures).

4.3.3.6 Accuracy with the increasing amount of training time (GFCC & GLFCC)

Figure 4.11 shows the accuracy versus duration of enrollment data between GFCCs and GLFCCs. Results in Figure 4.11 above show a similar pattern to that of RPCC and TPCC features in Figure 4.7 and Figure 4.9, except that the overall accuracy of the GLFCC features is somewhat higher. This indicate that the proposed GLFCC features are more compact with less dependence on phonetic content, because the short duration of data is sufficient to capture the speaker-specific characteristics in order to achieve a good performance.

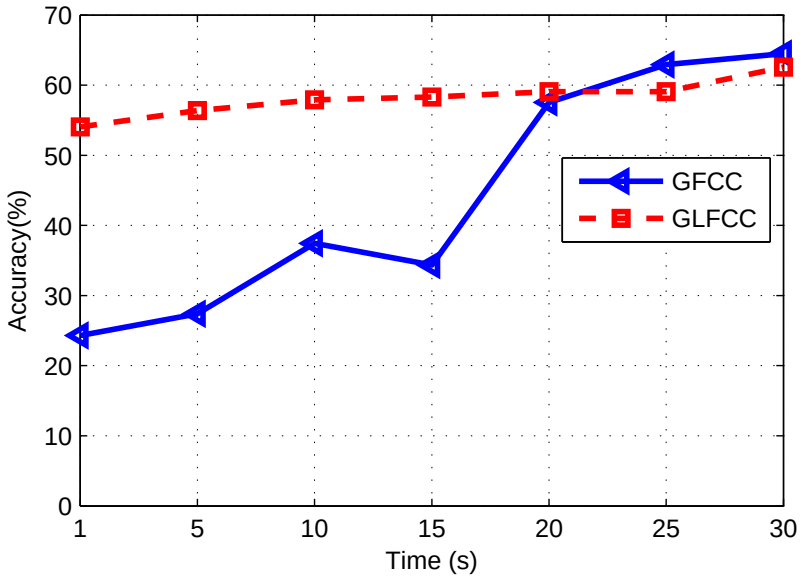


Figure 4.11: Accuracy versus duration of enrollment data (GFCC & GLFCC).

4.3.4 Experiment Summary

The experimental results confirm that the proposed features provide information about species characteristics that is significantly different in nature from the spectrally focused information present in traditional vocalization features such as GFCCs. Those new features give better results with smaller amounts of enrollment data and lower model complexities, and also provide complementary information that can improve overall system performance even for larger amounts of data. The fact that these new features are less dependent on the vocalization content of the species makes it useful for tasks of bioacoustic censusing.

4.4 Speaker Identification Experiments on YOHO Corpus

We have demonstrated that vocal source features can be shown to significantly improve identification accuracy in cross-condition scenarios, as these discriminative features have more power to differentiate speakers based on physiologically-motivated characteristics. In particular the proposed source features result in a better classification rate than the baseline features with lower model complexity. In this section, speaker identification performance is evaluated using a standard speaker identification task, the YOHO database, in order to further assess the usefulness of proposed features in a matched single language condition.

4.4.1 Data Corpus

The YOHO corpus was collected by ITT under a US government contract and was designed for speaker recognition systems in office environments with limited vocabulary [122]. This database was recorded using a telephone handset in a real office environment and sampled at an 8 kHz sampling frequency with 16 bits per sample. This corpus consists of 138 speakers each with 24 training utterances and 40 test utterances recorded in different sessions. The vocabulary consists of 56 two-digit numbers, ranging from 21 to 97 spoken continuously in sets of three (e.g., 32-56-68)

Table 4.6: YOHO corpus description.

No. of Speakers	138 (106 Males/32Females)
No. session/speaker	4 enrollments, 10 verifications
Type of speech	Prompted digit phrases
Microphones	Fixed high-quality in handset
Channels	3.8 kHz/clean
Acoustic environment	Office

in each utterance. In this work, all the utterances identified as enrollment data were used to train the model, and utterances in the verification data set were used for testing. Each enrollment session consists of 24 phrases and each testing utterance is a single example. There are about 6 minutes of speech used for training each speaker, and 2.4 seconds of speech for testing. The description of YOHO corpus is shown in Table 4.6.

4.4.2 Experimental Setup

A GMM-UBM framework speaker identification system is also used in this work, as shown in Figure 4.1. Initially, a universal background model is trained using the training utterances from all 138 speakers. Following this each speaker's model is adapted from the corresponding training utterance using the MAP adaption approach as described in Section 2.6.1.2. The identification experiments were conducted on the YOHO database as the number of mixtures is increased. This experimental

configuration is designed to evaluate if the proposed features can rapidly build an accurate model with lower model complexity.

The speech utterances were analyzed using 32ms frames with a 50% frame overlap, and twelve coefficients of each feature (MFCC, RPCC, GLFCC and TPCC) are derived from each frame. The first experiment uses individual features alone to assess their individual performance respectively, and then proposed features are appended to the baseline MFCC in order to evaluate their complimentary characteristics to the baseline feature.

4.4.3 Experimental Results

4.4.3.1 Accuracy of individual feature versus increasing number of mixtures

The accuracy versus increasing number of mixtures for the baseline feature MFCC and the proposed feature is shown in Figure 4.12. At low model complexity, RPCC and GLFCC features show better performance than the baseline feature MFCC. Of more significance is that GLFCC in particular performs better than the traditional MFCC features across all model configurations.

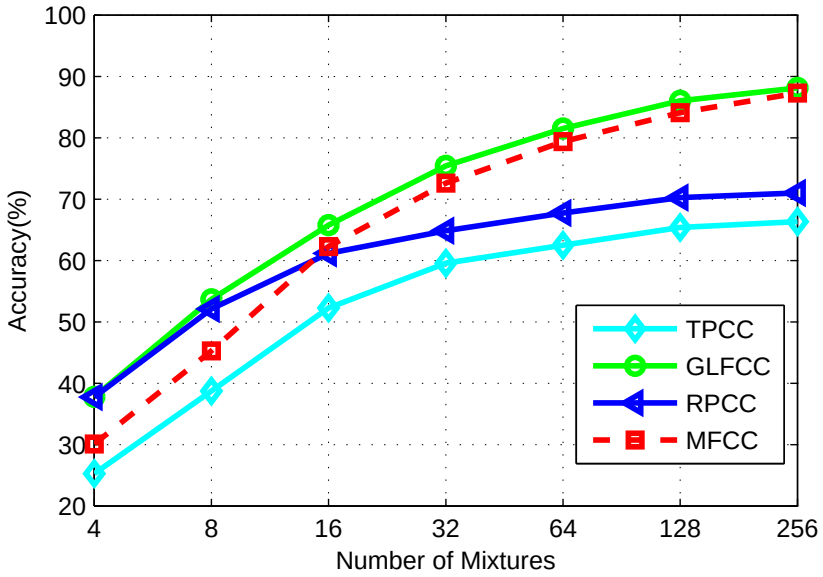


Figure 4.12: SID performance on YOHO with an increasing number of Gaussian mixture components.

4.4.3.2 Accuracy of combined features versus increasing number of mixtures

Figure 4.13 shows classification accuracy using MFCC features combined with the proposed source features versus increasing model complexity. The main trend visible in the figure is that appending more source features gives significantly better performance than using the traditional acoustic MFCC features alone. There is a significant difference in performance between the MFCC features and all combined features.

The performance across all model complexities clearly shows the robustness of

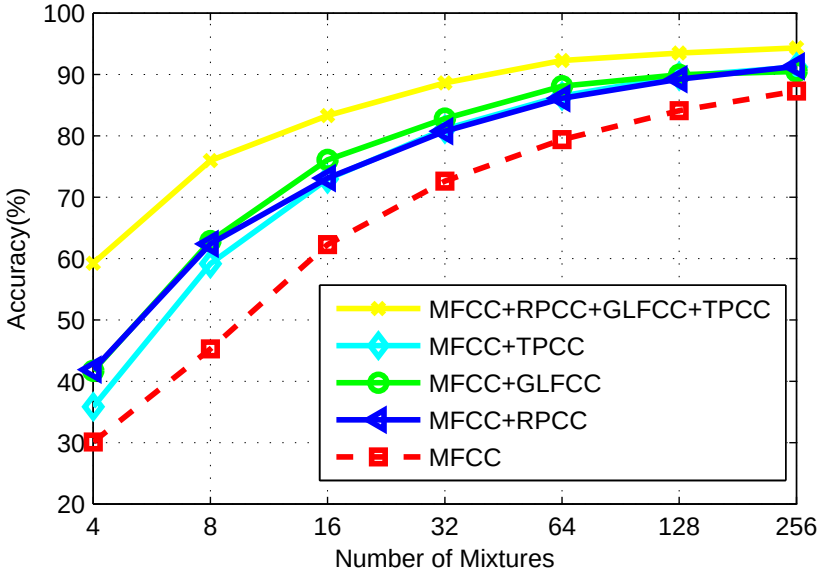


Figure 4.13: SID performance of combined features on YOHO with an increasing number of Gaussian mixture components.

combining the baseline features with the proposed vocal features. At low model complexities the improvement is very significant, for example, 60% (combined) vs. 30% (MFCC) at 4 mixtures and 75% (combined) vs. 45% (MFCC) at 8 mixtures. Table 4.7 shows the final system performance, which still indicates significant improvements.

Table 4.7: Accuracy of the final system with 256 mixtures.

Feature Combinations	Accuracy
MFCC + Proposed features	94.3
MFCC	87.3

4.5 Summary of All Experiments

The usefulness of vocal source features has been confirmed by several well-designed experimental paradigms. Taken as a whole across all 3 experimental paradigm investigated in this chapter, it is clear that the incorporation of the proposed vocal source features offers significant overall improvement to the robustness and accuracy of speaker identification tasks.

CHAPTER 5
CONCLUSION AND FUTURE WORK

This chapter summarizes the main contributions of this dissertation, highlights the major conclusions, and suggests several potential directions for further work.

5.1 Summary of Work

This dissertation has concentrated on front-end processing for speaker recognition applications with the aim of extracting speaker-discriminative feature parameters. By doing so, we have expanded the applicability of speaker recognition system towards challenging scenarios with mismatched conditions.

In this work a significant number of novel vocal tract and vocal source features have been introduced. New vocal tract features include primarily those based on spectral variability patterns such as LPC shimmer, LAR shimmer, cepstral shimmer, average harmonic amplitude difference (AHAD), and AHAD shimmer. New vocal source features include Residual Phase Cepstral Coefficients (RPCC), Teager Phase Cepstral Coefficients (TPCC), and Glottal Flow Cepstral Coefficients (GLFCC), all of which introduce different ways of capturing information related to the laryngeal excitation energy of a speaker.

The usefulness of these physiologically-motivated features for supplementing the vocal tract features in speaker identification has been evaluated by several carefully chosen experimental paradigms. Each experimental paradigm represents a practical application of speech technology that has mismatched phonetic characteristics and could be expected to benefit from more physiologically motivated features. For instance, cross language scenario represents how language difference factors deteriorate the performance of human speaker recognition systems, while the cross song-type individual avian identification experiment represents the limitation of missing prior knowledge of the vocalization repertoire categorization in bioacoustic censuing. Taken as a whole across all experimental results, we have seen that the proposed vocal source features offers significant overall improvement to the robustness and accuracy of speaker identification tasks. In particular, the improved performance with lower model complexity shows clearly that vocal source features outperform vocal tract features in terms of representative compactness. Furthermore, the performance across all model complexities demonstrates the robustness of combining the vocal tract features with the proposed vocal source features. Overall, these discriminative features significantly improve identification accuracy and have more power to differentiate speakers based on physiologically-motivated characteristics.

5.2 Summary of Contribution

This research has pushed forward relevant research in the following aspects:

First, a novel cross phoneme, language and song-type experimental paradigm has been introduced for evaluating phonetically-independent features for speaker recognition, using deliberately mismatched datasets for training and testing. By using this paradigm while increasing the duration of training data and mixture complexities we are able to evaluate the speaker-specific features in detailed ways. These experimental paradigms provide a new approach to identify features that represent broad individually unique characteristics rather than those that represent phonetic differences.

Second, it is shown in this research that the phase, glottal airflow and the energy contour of nonlinear speech production models in speech signal can be employed as speaker representative information that is effective for discriminating different speakers. Using these signal extraction methods to capture the physiological characteristics of the laryngeal source excitation will offer researchers new directions to extract useful features for speaker recognition.

Finally, the most important contribution of this work is that these novel vocal source features have been shown to signicantly improve identification accuracy even with cross-condition scenarios, as these discriminative features have more power to

better differentiate speakers based on their underlying physiological characteristics rather than the difference in pronunciation patterns. Moreover, the proposed source features have a better classification rate than the baseline features with lower model complexity. In practical applications, this lower model complexity offers several advantages, including improved computational efficiency as well as reduced demands for training and enrollment data, each of which has great significance in many real-world recognition, identification, and verification tasks. Over all, the proposed vocal source features have the potential to impact the practical applications in human speaker recognition and bioacoustic censusing.

5.3 Future Work

The research discussed in this dissertation suggests a number of avenues for future work as outlined below:

1. This work suggests that fusion of vocal source and tract information merits further research. Future work could involve investigation of more sources of information and the efficient fusion of features.

2. The effectiveness of utilizing vocal source excitation features to complement vocal tract features has been demonstrated in this work. To further improve the performance, it may be valuable to investigate new reliable and accurate

techniques for estimating vocal source signal characteristics.

3. Throughout the experiments, we have used the state-of-the-art GMM-UBM based speaker modeling because the main concentration of this work is to extract the speaker distinctive features rather than the novel modeling technique. The proposed features should be evaluated using other modeling techniques as well, in particular the Gaussian Mixture Model Support Vector Machines (GMM-SVM) since it is considered as the current dominant discriminative modeling technique. In addition, it may be important to investigate whether the i-vector approach, which has been very successful recently and is quickly becoming a dominant approach as well, can be integrated with these new features.

4. The currently state-of-the-art of spectral features contain a mixture of linguistic and speaker-related factors, which are not easily separated. Traditional spectral features such as MFCCs in reality work far better for a task such as speaker recognition than would be expected based on the information they represent. In order to understand better what makes up these individual characteristics in speech, it would be valuable to investigate this question in detail from both a practical and a theoretical perspective. This also has significant implications for automatic speech recognition, which benefits from removing such speaker-specific information from recognition features.

5.4 Conclusions

This dissertation has introduced several physiologically motivated features for speaker identification based on vocal source characteristics. These features, including RPCC, GLFCC and TPCC, represent information about unique aspects of speech production not represented in most state-of-the-art identification systems. The experimental results show competitive and in some case superior accuracy compared to baseline MFCC spectral features. With lower amount of training data and lower model complexity, these features give substantial improvement over MFCCs. The incorporation of these proposed glottal source features combined with traditional features offers significant overall improvement to the robustness and accuracy of speaker identification tasks. Since these proposed features are less dependent on the underlying phonetic content, they are broadly applicable to many speaker ID tasks, especially those with limited availability of enrollment data or mismatch between training and testing environments. Overall, this work enforces the need for investigation of vocal tract and vocal source features for robust speaker identification.

BIBLOGRAPHY

[1] J. Campbell, J.P., "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] N. Zheng, T. Lee, and P. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Proc. Letters*, vol. 14, no. 3, pp. 181–184, 2007.

[3] S. E. Anderson, "Speech recognition meets bird song: A comparison of statistics-based and template-based techniques," *The Journal of the Acoustical Society of America*, vol. 106, no. 4, pp. 2130–2130, 1999.

[4] T. M. Peake and P. K. Mcgregor, "Corncrake crex crex census estimates: a conservation application of vocal individuality," *Animal Biodiversity and Conservation*, vol. 24, pp. 81–91, 2001.

[5] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, pp. 91–108, 1995.

[6] C. H. You, K. A. Lee, and H. Li, "An SVM kernel with GMM-supervector based on the Bhattacharyya distance for speaker recognition," *IEEE Signal Proc. Letters*, vol. 16, pp. 49–52, 2009.

[7] S. Fine, J. Navratil, and R. Gopinath, "A hybrid GMM/SVM approach to speaker identification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01).*, vol. 1, 2001, pp. 417–420.

[8] A. Akula, V. Apsingekar, and P. De Leon, "Speaker Identification in Room Reverberation Using GMM-UBM," in *IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop. DSP/SPE 2009.*, 2009, pp. 37–41.

[9] N. Dehak, R. Dehak, N. B. P. Kenny, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space

for speaker verification," in *INTERSPEECH*, 2009, pp. 1559–1562.

[10] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, pp. 283–297, 1998.

[11] A. Solomonoff, W. Campbell, and I. Boardman, "Advances In Channel Compensation For SVM Speaker Recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05).*, vol. 1, 2005, pp. 629–632.

[12] B. Raj and P. Smaragdis, "Latent variable decomposition of spectrograms for single channel speaker separation," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.*, 2005, pp. 17–20.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, pp. 42–54, 2000.

[14] (2010) The nist year 2010 speaker recognition evaluation plan. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2010/NISTSRE10evalplan.r6.pdf

[15] (2008) The nist year 2008 speaker recognition evaluation plan. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/spk/2008/sre08evalplanrelease4.pdf

[16] R. Auckenthaler, M. Carey, and J. S. D. Mason, "Language dependency in text-independent speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01).*, vol. 1, 2001, pp. 441–444.

[17] X. Qing and K. Chen, "On use of GMM for multilingual speaker verification: An empirical study," in *Proceedings of ICSLP*, 2000, pp. 263–266.

[18] M. Faundez and A. Satue-Villar, "Speaker recognition experiments on a bilingual database," in *Proceedings of IV Jornadas en Tecnologias del Habla*, 2006, pp. 261–264.

[19] D. Geoffrey, "Multilingual text-independent speaker identification," in *Proceedings of MIST*, 1999.

[20] B. Ma and H. Meng, "English-Chinese bilingual text-independent speaker verification," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04).*, vol. 5, 2004, pp. 293–296.

[21] M. Akbacak and J. Hansen, "Language normalization for bilingual speaker recognition systems," in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. (ICASSP '07).*, vol. 4, 2007, pp. 257–260.

[22] Q. Jin, T. Schultz, and A. Waibel, "Phonetic speaker identification," in *ICSLP*, 2002, pp. 1345–1348.

[23] W. D. Andrews, M. A. Kohler, and J. P. Campbell, "Phonetic speaker recognition," in *INTERSPEECH*, 2001, pp. 2517–2520.

[24] J. P. Campbell, H. Nakasone, C. Cieri, D. Miller, K. Walker, A. F. Martin, and M. A. Przybocki, "The MMSR bilingual and cross-channel corpora for speaker recognition research and evaluation," in *Odyssey Speaker and Lang. Recog. Workshop*, 2004.

[25] A. Satue-Villar and M. Faundez-Zanuy, "On the relevance of language in speaker recognition," in *Eurospeech*, vol. 3, 1999, pp. 1231–1234.

[26] J. Goggin, C. Thompson, and L. S. G. Strube, "The role of language familiarity in voice identification," in *Memory and Cognition*, 1991, pp. 448–458.

[27] I. Luengo, E. Navas, I. Sainz, I. Saratxaga, J. Sanchez, I. Odriozola, and I. Hernaez, "Text independent speaker identification in multilingual environments," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, may 2008.

[28] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances," in *Interspeech*, 2008, pp. 853–856.

[29] M. McLaren, R. Vogt, B. Baker, and S. Sridharan, "Experiments in SVM based speaker verification using short utterances," in *Odyssey Speaker and Language Recognition Workshop*, 2010, pp. 83–90.

[30] A. Kanagasundaram, R. Vogt, D. B. Dean, S. Sridharan, and M. W. Mason, "i-vector based speaker recognition on short utterances," in *Interspeech 2011*, 2011, pp. 2341–2344.

[31] L. G. Kersta, "Voiceprint identification," *Nature 196*, vol. 196, no. 4861, pp. 1253–1257, 1962.

[32] J. E. Luck, "Automatic speaker verification using cepstral measurements," *Journal of Acoustical Society of America*, vol. 46, no. 4B, pp. 1026–1032, 1969.

[33] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[34] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of speech wave," *Journal of Acoustical Society of America*, pp. 637–655, 1971.

[35] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[36] L. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition.* Prentice Hall Press, 1993.

[37] L. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[38] D. Reynolds and R. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[39] A. Martin, D. Miller, M. Przybocki, J. Campbell, and H. Nakasone, "Conversational telephone speech corpus collection for the nist speaker recognition evaluation 2004," in *in Proceedings 4th International Conference on Language Resources and Evaluation*, 2004, pp. 587–590.

[40] D. A. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[41] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. of the European Conference on Speech Communication and Technology*, 1997, pp. 963–966.

[42] F. Weber, B. Peskin, M. Newman, A. Corrada-Emmanuel, and L. Gillick, "Speaker Recognition on Single- and Multispeaker Data," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 75–92, 2000.

[43] S. Fine, J. Navratil, J. R. N. Atil, and R. Gopinath, "Enhancing GMM Scores Using SVM "hints"," in *Proc. of the European Conference on Speech Communication and Technology*, 2001.

[44] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using prosodic and lexical information for speaker identification," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 141–144.

[45] A. Rosenberg, J. Delong, and C. Lee, "The use of cohort normalized score for speaker recognition," in *The Second International Conference on Spoken Language Processing (ICSLP)*, 1992, pp. 599–602.

[46] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[47] R. Vogt and S. Sridharan, "Experiments in session variability modelling for speaker verification," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP'06)*, vol. 1, 2006, pp. 897–900.

[48] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garca, D. Petrovska-Delacrtaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Advances in Signal Processing*, pp. 430–451, 2004.

[49] A. Solomonoff, W. Campbell, and I. Boardman, "Advances In Channel Compensation For SVM Speaker Recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05).*, vol. 1, 2005, pp. 629–632.

[50] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end

factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[51] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[52] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, no. 4, pp. 543–565, 2012.

[53] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[54] K. Prahallad, "A Practical Introduction, Topic: Spectrogram, Cepstrum and Mel-Frequency Analysis," in *Carnegie Mellon University and International Institute of Information Technology Hyderabad*, 2010.

[55] R. Zilea, J. Navratil, and G. N. Ramaswamy, "Depitch and the role of fundamental frequency in speaker recognition," in *2003 IEEE International Conference oAcoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 2, 2003, pp. II–81–4.

[56] K. Chen and A. Salman, "Extracting speaker-specific information with a regularized Siamese deep network," in *Advances in Neural Information Processing Systems*, 2011.

[57] P. Clemins, M. Trawicki, K. Adi, J. Tao, and M. Johnson, "Generalized perceptual features for vocalization analysis across multiple species," in *2006 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, 2006, pp. 1253–1256.

[58] X. Huang and A. Acero, *Spoken Language Processing*. Upper Saddle River, New Jersey, USA: Prentice Hall Press, 2001.

[59] D. Greenwood, "A cochlear frequency-position function for several species," *The Journal of the Acoustical Society of America*, 1990.

[60] D. D. Greenwood, "Critical bandwidth and the frequency coordinates of the basilar membrane," *The Journal of the Acoustical Society of America*, vol. 33, no. 10, p. 1344, 1961.

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

[62] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2006)*, vol. 1, 2006, pp. I–I.

[63] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, and C. Vair, "Compensation of nuisance factors for speaker and language recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1969–1978, 2007.

[64] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden markov models," *IEEE Transactions on Signal Processing*, vol. 39, no. 4, pp. 806–814, 1991.

[65] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[66] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[67] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*. ACM Press, 1992, pp. 144–152.

[68] N. Cristianini and J. Shawe-Taylor, *Support Vector Machines*. Cambridge University Press, 2000.

[69] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143–160, 2001.

[70] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *2006 IEEE International Conference on Acoustics, Speech*

and Signal Processing (ICASSP 2006), vol. 1, 2006, pp. I–I.

[71] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of Heterogeneous Speaker Recognition Systems in the STBU Submission for the NIST Speaker Recognition Evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[72] B. Pellom and J. Hansen, "An efficient scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Processing Letters*, vol. 5, no. 11, pp. 281–284, 1998.

[73] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms," Tech. Rep., 2005.

[74] G. J. Borden, K. S. Harris, and L. J. Raphael, *Speech Science Primer: Physiology, Acoustics, and Perception of Speech.* Lippincott Williams and Wilkins, 2003.

[75] D. O'Shaughnessy, *Speech Communications: Human and Machine.* Institute of Electrical and Electronics Engineers, 2000.

[76] J. L. Flanagan, *Speech analysis and perception, 2nd edition.* Springer-Verlag, 1965.

[77] H. W. Dudley, "The carrier nature of speech," *Bell Systems Technical Journal*, vol. 19, pp. 495–513, 1940.

[78] G. Fant, *Acoustic Theory of Speech Production.* The Hagues: Mounton, 1960.

[79] T. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice.* Prentice Hal, 2002.

[80] H. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[81] G. Zhou, J. Hansen, and J. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[82] A. J. Chorin and J. E. Marsden, *Mathematical Introduction to Fluid Mechanics*. Springer-Verlag, 1990.

[83] J. Kaiser, "Some useful properties of Teager's energy operators," in *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, vol. 3, 1993, pp. 149–152 vol.3.

[84] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, 1993.

[85] B. S. Atal, "Automatic speaker recognition based on pitch contours," *Journal of the Acoustic Society of America*, pp. 1687–1697, 1972.

[86] J. H. L. Liu and G. Palm, "On the use of features from prediction residual signal in speaker recognition," in *Proc. of European Conf. Speech Processing*, 1997, pp. 313–316.

[87] P. Thevenaz and H. Hugli, "Usefulness of LPC residue in text independent speaker verification," *Speech Communication*, pp. 145–157, 1995.

[88] H. Wakita, "Residual energy of linear prediction applied to vowel and speaker recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 270–271, 1976.

[89] G. Shi, M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1867–1874, 2006.

[90] K. K. Paliwal and L. Alsteris, "Usefulness of phase spectrum in human speech perception," in *Proc. Eurospeech*, 2003, pp. 2117–2120.

[91] W. Huang, J. Chao, and Y. Zhang, "Combination of pitch and MFCC GMM supervectors for speaker verification," in *International Conference on Audio, Language and Image Processing (ICALIP 2008)*, 2008, pp. 1335–1339.

[92] I. R. Titze, *Principles of voice production*. Prentice Hall Press, 1994.

[93] E. G. Hautamaki, *Fundamental Frequency Estimation and Modeling for Speaker Recognition*. University of Joensuu, 2005.

[94] J. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 5, pp. 367–377, 1972.

[95] X. Li, J. Tao, M. Johnson, J. Soltis, A. Savage, K. Leong, and J. Newman, "Stress and Emotion Classification using Jitter and Shimmer Features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2007)*, vol. 4, 2007, pp. IV–1081–IV–1084.

[96] P. Lieberman, "Some acoustic measures of the fundamental periodicity of normal and pathologic larynges," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 344–353, 1963.

[97] R. Wendahl, "Laryngeal analog synthesis of jitter and shimmer auditory parameters of harshness," *Folia Phoniatrica*, vol. 18, no. 2, pp. 98–108, 1966.

[98] D. E. Veeneman and S. BeMent, "Automatic glottal inverse filtering from speech and electroglottographic signals," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 369–377, 1985.

[99] A. Neocleous and P. A. Naylor, "Voice source parameters for speaker verification," in *Proc. Eur. Signal Process. Conf.*, 1998, pp. 697–700.

[100] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*. Prentice Hall Press, 1993.

[101] H. M. Hanson, "Glottal characteristics of female speakers," Ph.D. dissertation, Harvard University, 1995.

[102] M. Rothenberg, "Acoustic interaction between the glottal source and the vocal tract," *Vocal Fold Physiology*, pp. 305–328, 1981.

[103] Z. M. Smith, B. Delgutte, and A. J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature (London)*, 2002.

[104] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. Guiod, and S. L. Goldman, "Comparisons among aerodynamic, ectroglottographic, and acoustic spectral measures of female voice," *J. Speech Hearing*, 1995.

[105] G. Chen, X. Feng, Y. L. Shue, and A. Alwan, "On using voice source measures in automatic gender classification of children's speech," in *INTERSPEECH*,

2010.

[106] L. Liu, J. He, and G. Palm, "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, vol. 22, pp. 403–417, 1997.

[107] G. Shi, M. Shanechi, and P. Aarabi, "On the importance of phase in human speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1867–1874, 2006.

[108] K. Murty and B. Yegnanarayana, "Combining evidence from residual phase and mfcc features for speaker recognition," *IEEE Signal Processing Letters*, vol. 13, no. 1, pp. 52–55, 2006.

[109] B. T. Labs, *High speed motion pictures of the human vocal cords.* Bureau of Publication, 1937.

[110] M. Plumpe, T. Quatieri, and D. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 5, pp. 569–586, 1999.

[111] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, pp. 109–118, 1992.

[112] W. M. Fisher, G. R. Doddington, and K. M. Goudie-Marshall, "The DARPA Speech Recognition Research Database: Specifications and Status," in *Proceedings of DARPA Workshop on Speech Recognition*, 1986, pp. 93–99.

[113] S. Young, *the HTK Book (for HTK Version 3.4.1)*, 2009.

[114] R. Rose, E. Hofstetter, and D. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 245–257, 1994.

[115] (2004) The nist year 2004 speaker recognition evaluation plan @ONLINE. [Online]. Available: http://www.nist.gov/speech/tests/spk/2004

[116] K. Adi, K. Sonstrom, P. Scheifele, and M. Johnson, "Unsupervised validity measures for vocalization clustering," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)*, 2008, pp. 4377–4380.

[117] K. Adi and M. T. Johnson, "Feature normalization for robust individual identification of the ortolan bunting (emberiza hortulana L)," *The Journal of the Acoustical Society of America*, vol. 119, no. 5, pp. 3435–3435, 2006.

[118] K. Adi, T. S. Osiejuk, and M. T. Johnson, "Automatic songtype classification and individual identification of the ortolan bunting (Emberiza hortulana L) bird vocalizations," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2639–2639, 2004.

[119] P. Clemins and M. T. Johnson, "Generalized perceptual linear prediction (gPLP) features for animal vocalization analysis," *Journal of the Acoustical Society of America*, vol. 120, pp. 527–534, 2006.

[120] S. Cramp and C. M. Perrins, *The Birds of the Western Palearctic*. Oxford University Press, 1994.

[121] S. Dale and O. Hagen, "Population size, distribution, and habitat choice of the ortolan bunting emberiza hortulana in norway," *Fauna Norv. Ser. C*, vol. C, pp. 93–103, 1997.

[122] J. Campbell and H. Alan, *YOHO Speaker Verification (LDC94S16)*, 1994.