Speaker Recognition Using Complementary Information from Vocal Source and Vocal Tract

ZHENG Nengheng

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

November 2005

To my wife Suyan *and our daughter* Yoyo

&

in memory of my mother

Acknowledgements

Firstly, I would like to express my gratitude to my supervisor, Prof. P. C. Ching, for his advice, encouragement and guidance throughout this PhD study and research. I am also much obliged for his supporting me to conduct research in CUHK and to attend oversea academic activities.

I would also like to thank Prof. Tan Lee for his advice and suggestions in my research. Prof. Lee also helps me a lot in improving my English writing and presentation skills. Thanks also to Prof. Willam S-Y Wang for his instructive lectures on Language and Speech Science, Prof. Y. T. Chan for his insightful view of signal processing in his ADSP course, and Prof. W. K. Cham for his advice in teaching during my being a tutor of his courses. Thanks also go to Prof. X. G. Xia, Prof. W.-K. Ma and Prof. W.-K. Lo for their help. Special thanks to Dr. Frank Soong, for his valuable advice and suggestions in my research and life, including giving a name to my new born daughter.

Many thanks to my friends in DSP and Speech Technology Laboratory. Life with them is much more than happiness. Thanks to Mr. Arthur Luk for his technical support.

I wish to express my deepest gratitude to my parents for their love throughout my life, my brother and sister for their support in my Undergraduate and Master study. Finally, the deepest gratitude to my wife Suyan, who has been with me in the whole journey.

Abstract of thesis entitled:

Speaker Recognition Using Complementary Information from Vocal Source and Vocal Tract

Submitted by **ZHENG Nengheng**

for the degree of **Doctor of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong** in

November 2005.

This thesis investigates the feasibility of using both vocal source and vocal tract information to improve speaker recognition performance. Conventional speaker recognition systems typically employ vocal tract related acoustic features, e.g the Mel-frequency cepstral coefficients (MFCC), for discriminative purpose. Motivated by the physiological significance of the vocal source and vocal tract system in speech production, this thesis develops a speaker recognition system to effectively incorporate these two complementary information sources for improved performance and robustness.

This thesis presents a novel approach of representing the speaker-specific vocal source characteristics. The linear predictive (LP) residual signal is adopted as a good representative of the vocal source excitation, in which the speaker-specific information resides on both time and frequency domains. Haar transform and wavelet transform are applied for multi-resolution analyses of the LP residual signal. The resulting vocal source features, namely the Haar octave coefficients of residues (HOCOR) and wavelet octave coefficients of residues (WOCOR), can effectively extract the speaker-specific spectro-temporal characteristics of the LP residual signal. Particularly, with pitch-synchronous wavelet transform, the WOCOR feature set is capable of capturing the pitch-related low frequency properties and the high frequency information associated with pitch epochs, as well as their temporal variations within a pitch period and

over consecutive periods. The generated vocal source and vocal tract features are complementary to each other since they are derived from two orthogonal components, the LP residual signal and LP coefficients. Therefore they can be fused to provide better speaker recognition performance. A preliminary scheme of fusing MFCC and WOCOR together illustrated that the identification and verification performance can be respectively improved by 34.6% and 23.6%, both in matched conditions.

To maximize the benefit obtained through the fusion of source and tract information, speaker discrimination dependent fusion techniques have been developed. For speaker identification, a confidence measure, which indicates the reliability of vocal source feature in speaker identification, is derived based on the discrimination ratio between the source and tract features in each identification trial. Information fusion with confidence measure offers better weighted scores given by the two features and avoids possible errors introduced by incorporating source information, thereby improves the identification performance further. Compared with MFCC, relative improvement of 46.8% has been achieved.

For speaker verification, a text-dependent weighting scheme is developed. Analysis results show that the source-tract discrimination ratio varies significantly across different sounds due to the diversity of vocal system configurations in speech production. This thesis analyzes the source-tract speaker discrimination ratio for the 10 Cantonese digits, upon which a digit-dependent source-tract weighting scheme is developed. Information fusion with such digit-dependent weights relatively improves the verification performance by 39.6% in matched conditions.

Experimental results show that source-tract information fusion can also improve the robustness of speaker recognition systems in mismatched conditions. For example, relative improvements of 15.3% and 12.6% have been achieved for speaker identification and verification, respectively.

摘要

說話人識別是人機交互系統中的一個重要技術。基於系統信息安全的考慮，在人機交互系統中往往需要對用戶准入提供必要的身份認證。說話人識別就是根據用戶的語音信號中所攜帶的每個說話人所特有的特徵來進行身份認證。語音信號的產生包含了聲帶振動產生的激勵源信號（聲門波）以及聲道（包含咽腔、口腔和鼻腔）調製兩個部分。傳統的說話人識別系統通常僅利用了聲道調製有關的說話人特徵參數，例如 Mel-倒譜系數（MFCC）。前人的研究發現激勵源信號中包含有豐富的說話人的特徵信息。但是如何有效地提取源信號中的與說話人有關的特徵信息卻一直是一個未能解決的難題。本論文系統地探討了激勵源信號中所附帶的說話人特徵信息，進而提出了一種基於時頻分析的有效的特徵提取技術。在此基礎上，本文探討了如何高效地利用所提出的激勵源特徵信息以及傳統的聲道特徵信息，以期達到更高的識別率及更強的魯棒性。

本文分別以綫性預測（LP）殘差信號及綫性預測參數（LPC）代表激勵源信號和聲道調製系統。由於二者的正交性，由之推導出來的源特徵參數以及聲道特徵參數具有最大的互補性。爲了提取 LP 殘差信號的與說話人有關的時頻信息，本文分別採用了 Haar 變換和小波（wavelet）變換對 LP 殘差信號進行時頻分析，並進而推導出兩种激勵源特徵參數 HOCOR 和 WOCOR。特別地，由於採用了基音同步小波變換（pitch-synchronous wavelet transform）算法，WOCOR 不但有效地提取了與基頻相關的低頻特徵和與基音脈衝相關的高頻特徵，而且刻畫了這些頻率成分在一個基音周期內部和相鄰兩個基音周期之間的時間變化特性。實驗證明這些時頻特徵有效地刻畫了每個特定說話人的激勵源特性，因此可以用於說話人識別。初步的實驗結果表明，相對于傳統的僅利用聲道特徵參數 MFCC，同時利用激勵源特徵參數（HOCOR 或 WOCOR）和聲道特徵參數能顯著地提高說話人的識別率。例如，在訓練和識別條件匹配的實驗中，說話人辨認（speaker identification）錯誤率有 34.6%的相對減少，說話人確認（speaker verification）錯誤率則有 23.6%的相對減少。在條件不匹配的實驗中，辨認和確認的錯誤則分別減少了 15.3%和 12.6%。

爲了進一步提高識別率，本論文提出了一種新的信息融合（information fusion）算法以高效地利用源特徵參數和聲道特徵參數之間的互補性。在說話人辨認實驗中，我們提出以可變的，基於兩种特徵參數辨識力的置信測度（confidence measure）作爲信息融合的權值代替事先訓練好的不變的權值。由於引入了激勵源和聲道特徵參數在每一次辨認試驗中的相對辨識力，這種信息融合算法進一步提高了識別率。相對于僅利用 MFCC 參數，錯誤率減少了 46.8%。此外，本論文分析了激勵源和聲道特徵參數在說不同文本語音時對說話人的辨識能力。以粵語的 10 個數字為例，分析結果表明對於不同的數字，MFCC 和 WOCOR 對於說話人的辨識能力有顯著的不同。因此本論文訓練了一組基於文本的信息融合權值。在採用了該融合算法的說話人確認的實驗中，相對于 MFCC，錯誤率減少了 39.6%。

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Fundamental of Speaker Recognition

Speaker recognition is a branch of biometric authentication which refers to the automatic identity recognition of individuals using certain intrinsic characteristics of the person. Biometric authentication has been an important techniques for human-machine communication system in applications with security consideration. Besides the voice, there are many other physical and behavioral patterns, e.g. eyes, face, fingerprint, signature, etc., for biometric authentication. Practically, selection of a promising biometric pattern should take into account at least the following concerns: robustness, distinctiveness, accessibility, and acceptability [111]. Table 1.1 compares the four properties of some commonly adopted biometric patterns. The judgement of a *good* biometric pattern is very complicated and depends on the specifics of the applications.

Among all the biometric authentication technologies, speaker recognition is probably the most natural and economical one for human-machine communication systems due to (1) speech data collection is much more convenient than other patterns; and (2) more importantly, speech is the dominant mode of information exchange for human beings and it tends to be the dominant mode for human-machine information exchange. The development of speech processing technology has boosted many applications of speaker recognition, especially in the following areas:

Table 1.1: Comparison of some biometric patterns for identity recognition

Biometric patterns	Iris	Face	Fingerprint	Voice
Distinctiveness	high	high	high	moderate
Robustness	high	high	moderate	moderate
Accessibility	low	moderate	moderate	high
Acceptability	moderate	high	moderate	high

Distinctiveness: the existence of wide differences in the pattern among the population.
Robustness: repeatable, not subject to large changes.
Accessibility: easily presented to an sensor.
Acceptability: perceived as non-intrusive by the users.

- access control to physical facilities or data networks;
- telephone credit card purchases or other bank transaction;
- information retrieval, e.g. customer information for call centers and audio indexing;
- remote monitoring;
- forensic voice sample matching.

Speaker recognition can be divided into verification and identification tasks. The verification task is to decide whether or not an unlabeled voice belongs to a claimed speaker (Figure 1.1.a). There are only two possible decisions: either to accept the voice as belonging to the claimed speaker or to reject it as belonging to an impostor. The identification task is to classify the unlabeled voice as belonging to one of the registered speakers (Figure 1.1.b). The number of decision alternatives in speaker identification is the same as the population size N, and generally the performance is inversely proportional to N. Therefore, it is usually a more difficult task than verification with large N. Speaker identification as defined is also called *closed-set* identification. Its contrast, the *open-set* identification encompasses a possibility that the unlabeled voice belongs to none of the registered speakers. Therefore, the number of decision alternatives is $N + 1$

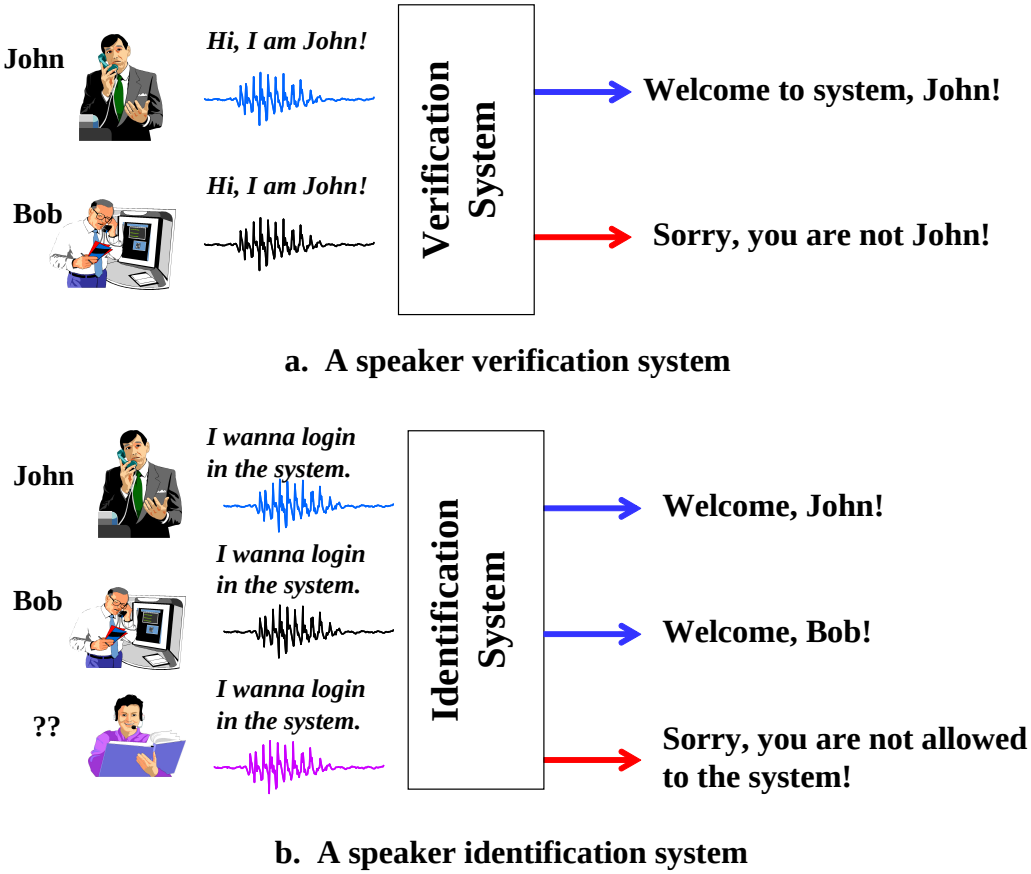


Figure 1.1: Speaker verification and identification systems

which includes a decision that the voice belongs to an unknown speakers (Figure 1.1.b). The open-set identification is a combination of identification and verification.

Speaker recognition can also be divided into text-dependent and text-independent recognitions. In text-dependent recognition, the system knows exactly the spoken text which could be either fixed phrase or prompted phrase. In text-independent recognition, the system does not know the text of the spoken utterance, which could be user selected keywords or conversational speech. With the knowledge of spoken text, the system can exploit the speaker individuality associated with specific phonemes or syllables. Thus a text-dependent system generally performs better than the text-independent system. However, it requires highly cooperatives of the speakers and can be used only for applications with strong control over user input. The text-independent system is more user-friendly and more applicable but, without the knowledge of the spo-

ken text, also more difficult to achieve high performance. In text-independent applications, a speech recognizer which provides the correct text knowledge can improve the speaker recognition accuracy [31]. Although the text-independent task has been accepted as a good platform for evaluating the general technologies for speaker recognition (e.g. the NIST annual evaluations [113]), many commercial and industrial applications focus more on the text-dependent, or text-constraint speaker recognition.

1.2 Historical Achievements in Speaker Recognition Technology

Researches on speaker recognition have been undertaken for more than 40 years and it continues to be an active area of spoken language processing. The development in speaker recognition technology is closely concomitant with the advancement in speech and signal processing and computer technology.

Speaker recognition by human was broadly studied in the 1960s. The motivation of these studies was to learn how human recognizes speakers and the reliability of human in recognizing a speaker [13][102]. The most significant work which stimulated the further research on speaker recognition by machine was done by Kersta who introduced the *spectrogram* (where he noted as *voiceprint*) as a means of personal identification [51].

In the 1970s, attentions had been turned to speaker recognition by computer and cames the so called *automatic speaker recognition*. Speaker recognition systems in this era usually only dealt with a small population (< 20 speakers) [7][64]. Fourier transform, linear predictive and cepstral analysis techniques have been applied for generating feature parameters. Long-time average of these parameters were used as the speaker references.

In the 1980s, more complicated statistical pattern recognition methods had been investigated, e.g. the Dynamic Time Warping (DTW) [30][46] and Vector Quantization (VQ)[101], for large scale speaker recognition systems (> 100 speakers). The contribution of static and dynamic features for speaker recogni-

tion was also investigated [41][100].

Since the 1990s, the available of large scale speech database (e.g. the YOHO corpus [17]) has boosted studies on more complicated models for speaker representation. These models include the stochastic models (e.g. Hidden Markov Model (HMM) [23][27], Gaussian Mixture Model (GMM) [89]), neural networks (e.g. Multilayer Perceptron (MLP) [61], Radial Basis Function (RBF) [76]) and support vector machines (SVM) [96][110], etc. Among these modeling techniques, the GMM has been recognized to be the most effective in characterizing the density distribution of the speech data and has been the dominant modelling technique for speaker recognition. As for feature extraction, cepstral coefficients incorporating the auditory model, i.e. the Mel-frequency cepstral coefficients (MFCC) and their dynamic coefficients have been the dominant feature parameters. Besides, various score normalization techniques have also been investigated for robust speaker recognition [9][27][88][85]. A system with MFCC parameters, GMM modeling and universal background model (UBM) for score normalization has been reported to achieve best results and have been widely accepted as the baseline for comparing new technologies [88]. To foster interaction among researchers in speaker recognition, a benchmark evaluation program has been carried out by NIST for different research communities to demonstrate their technology advancements [113]. In this practice, common test data and evaluation process are used so that different technologies and systems become comparable.

General and detailed overviews of speaker recognition can be referred to [8][18][30][40][73][77][80] [86][91][93].

1.3 Challenge to the State-of-the-art Speaker Recognition

The State-of-the-art speaker recognition systems work very well in laboratory experiments or under some specific applications with sophistically designed training and operation conditions. Experimental results showed that automatic

speaker recognition in such ideal environments performs as good as, or even better than, the recognition by human beings [97]. However, as an application-oriented technique, the performance of current speaker recognition systems in real-world applications is far from being robust and reliable in comparison with the recognition performance by human. The primary challenge to speaker recognition technology has therefore been improving the robustness of systems under mismatched conditions. For speaker recognition, the mismatches are caused mainly by (1) intra-speaker variation of speaking style; and (2) acoustic environment variation.

Our vocal system provides primarily the acoustic cues for phoneme classification and secondarily the individual personality for speaker characterization. The inter-speaker variation could be significant even for the same speech content. A speaker recognition system try to figure out such inter-speaker variation upon which a specific speaker is discriminated from the others. At the same time, the vocal system produce a certain degree of intra-speaker variation when uttering the same speech in different time. Many of the recognition errors are caused by such kinds of intrinsic intra-speaker variation.

The acoustic environment variation, on the other hand, is caused by the various unpredictable distortions during data collection and transmission. For example, in telephony speaker recognition applications (e.g. telephone bank transaction), the speech data could be collected in different background environments, with various phones, and via different channels. The background noise and handset/channel distortions change the spectral structure of the speech data and the derived acoustic features (e.g. the MFCC parameters) can not represent the speaker information correctly.

Many efforts have been devoted to improve the robustness of speaker recognition systems in real-world applications. They can be categorized as follows:

- High-level features extraction [21][87]. Generating new features (e.g. prosodic, speaking style, diction, etc) to supplement the low-level acoustic feature for robust speaker recognition has become a hot topic in this area. However, the effective feature extraction, appropriate data modeling and

efficient information fusion as well have not been thoroughly exploited.

- Feature and model transformation [7][84][90][118]. These approaches try to reduce the effect of spectral distortion either by compensating the parameterized features to equalize the mismatched conditions or by transforming the speaker model to better reflect the distorted features. These methods, although improve the robustness to a certain extern, still can not guarantee an desirable performance for real applications.

- Multi-modality speaker recognition which adopt other bio/non-biometric patterns besides speech for identity authentication. For example, a multi-modality system combines face, fingerprint and voice patterns has been presented [71]. The complexity in data acquisition restricts its application scenarios. Another multi-modality system combines the speaker verification and verbal information verification (VIV) [79]. Only speech data is collected in such system. However, it requires the speaker to speak out some personal information (e.g. birthday, address, etc) and is vulnerable to impostors who have stolen the personal information.

1.4 Motivation and Goal of This Thesis

Motivated by improving the recognition accuracy by fusion of different information sources, this thesis focuses on exploiting the speaker-specific vocal source information for speaker recognition. According to the speech production theory, human speech is produced as the vocal cords phonation followed by vocal tract articulation and lip radiation. Acoustic features representing the vocal tract characteristics (e.g. MFCC and LPCC) have been widely applied for speaker recognition. Although it has been revealed that glottal phonation plays an important role in speaker characterization and human beings rely partially on this kind of vocal source information to recognize familiar speakers [77], the usefulness of the vocal source characteristics for automatic speaker recognition, as well as its effective feature extraction technique, has not been thoroughly studied and fully exploited.

The goal of this research is to improve the speaker recognition performance by using the complementary vocal source and vocal tract speaker information. Towards this goal, we try to answer the following questions:

1. Is it really useful to take into account the vocal source information for speaker recognition?

2. How to effectively represent the speaker-specific information from the vocal source signal?

3. How to take full advantage of the fusion of vocal source and vocal tract information for speaker recognition?

Many efforts have been devoted to study the role of vocal source excitation in speech production. Most of these work are motivated by study the effect of excitation types in natural speech production with applications in speech synthesis and clinical diagnoses of voice disorder. Miller has demonstrated in 1963 that the vocal source signal has more variation among different speakers than among different utterances of the same speaker [72]. From then on, a few work have been conducted to investigate the usefulness of vocal source information for speaker recognition. Most of these work focused on exploiting the role of pitch in speaker recognition [6][34][99]. The pitch, though a very important parameter characterizing the vocal source activity, cannot cover the diverse time-frequency properties of the glottal phonation. A more efficient feature extraction technique is highly desired to retrieve the discriminative source information for speaker recognition.

This thesis first introduces the physiological process and the acoustic theory of speech production. The speaker-specific information associated with the glottal phonation is analyzed. The usefulness of vocal source information for speaker recognition is discussed. In particular, the complementarity of the vocal source and vocal tract speaker information is addressed.

This thesis proposes a time-frequency feature extraction technique to capture the spectro-temporal characteristics from the linear predictive (LP) residual signal. The LP residual signal, though not giving the true glottal pulse,

is regarded as a good representative of the excitation source. The LP residual signal of voiced speech is a quasi-periodic signal with pitch epochs appearing at around the glottis closing instant (GCI). The dominant speaker information is embedded in the pitch epochs. To represent the pitch-related low frequency properties and the high frequency information associated with the pitch epochs, two time-frequency analysis tools, the Haar transform and wavelet transform are applied for the multi-resolution analysis of the residual signal. The derived feature parameters, namely the Haar octave coefficients of residues (HOCOR) and wavelet octave coefficients of residues (WOCOR), are believed to effectively capture the speaker-specific spectro-temporal characteristics of the LP residual signal and are verified with several experiments to be useful for speaker recognition.

The presented vocal source features are used complementarily to the vocal tract features to improve the system performance and robustness. To maximize the benefit obtained from the source-tract information fusion, optimized fusion techniques, in which the source-tract fusion weights are derived with discriminative analysis, are developed to further improve the system performance. A varying weighting scheme, in which the fusion weight is derived based on the source-tract discrimination ratio on each trial, is developed for robust speaker identification. For speaker verification, a text-dependent weighting scheme is also developed. It is found that there is significant difference in speaker discrimination power among different sounds due to the diversity of vocal system configuration in producing different sounds. This thesis analyzes the vocal source and vocal tract discrimination power of the Cantonese digits, upon which a digit-dependent fusion scheme is developed to improve the speaker verification performance. Such a fusion scheme will be useful for real-world verification applications where the spoken text are known to the system. Speaker recognition experiments illustrate that the proposed source-tract information fusion techniques relatively improve the identification and verification performance by 46.8% and 39.6% in matched conditions, and 15.3% and 12.6% in mismatched conditions, in comparison with that use vocal tract information only.

1.5 Thesis Outline

The rest of this thesis is organized as follows:

Chapter 2 provides an overview of automatic speaker recognition including its three fundamental components, i.e. feature extraction, pattern generation and matching, and classification decision. The robustness consideration for speaker recognition over telephone network is also addressed.

Chapter 3 reviews the acoustic theory of speech production and compares the speaker characteristics embedded in the vocal source signal and the vocal tract system. The existing vocal tract and vocal source feature extraction techniques are introduced and their usefulness in speaker recognition is discussed.

In Chapter 4, we conduct an extensive investigation on extracting efficient speaker discriminative source features to enhance the performance of speaker recognition. Two time-frequency transforms, i.e. Haar transform and wavelet transform, are applied to analyze the LP residual signal and their effectiveness in feature extraction are studied.

Chapter 5 evaluates the performances of the proposed vocal source features for speaker recognition. The performances of vocal source feature and the vocal tract feature are compared. The improvement of performance by using these two complementary features is demonstrated.

In Chapter 6, we analyze the speaker discrimination power of the vocal source and vocal tract features, upon which the discrimination-dependent source-tract information fusion schemes are developed to take full advantage of the complementarity of the two feature for speaker recognition.

Chapter 7 discusses the impact of pitch variation on speaker recognition and compares the presented features with existing vocal source features for speaker recognition. Finally Chapter 8 concludes this thesis and includes a brief outline of future work.

Chapter 2

Speaker Recognition: Technical Review and New Thoughts

As any other pattern recognition systems, a speaker recognition system also consists of three components: (1) feature extraction, which transforms the speech waveform into a set of parameters carrying salient speaker information; (2) pattern generation, which generates from the feature parameters a pattern representing the individual speaker; and (3) pattern matching and classification, which compares the similarity between the extracted features and a pre-stored pattern or a number of pre-stored patterns, giving the speaker identity accordingly. Figure 2.1 depicts a speaker recognition system. As illustrated, there are two stages in a speaker recognition system, training and recognition. In the training stage, speaker models (or patterns) are generated from the speech samples with some feature extraction and modelling techniques. In the recognition stage, feature vectors are generated from the input speech sample with the same extraction procedure as in training. Then a classification decision is made with some matching techniques. According to the classification types, recognition task can be divided into either identification or verification. The former is a multi-classes matching. The input features are compared with all the registered patterns and the one giving the highest score is identified as the target speaker, or an unknown impostor in open-set identification. The latter is a binary classification. The input features are compared with the claimed

speaker pattern and a decision is made to accept or reject the claiming.

The following of this chapter will present a technical review of speaker recognition. Section 2.1- 2.3 will specify the technical details of feature extraction, pattern generation/matching and classification. Various most popular techniques will be introduced. Section 2.4 will introduce some prevalent performance evaluation metric. Finally, Section 2.5 will discuss several techniques widely adopted for robust speaker recognition over telephone network.

2.1 Feature Extraction and Selection

The objective of feature extraction is to estimate feature parameters representing the speaker individuality, which is the results of the physical differences in the vocal system and the learned speaking habits and styles as well. Accordingly, speaker specific information embedded in the speech signal can be classified into two categories: (1) the low level information, which is related to anatomical structure of the vocal apparatus and (2) the high level information, which is related to the learned habits and styles. Table 2.1 gives the hierarchical features for human and machine speaker recognition.

Acoustic features representing the low level information have been widely applied in speech and speaker recognition. These low level features focus on revealing the speech- and/or speaker-dependent vocal tract configurations. The most prevalent features are those based on cepstral analysis of speech, such as the linear predictive cepstral coefficients (LPCC) [7][38] and the Mel-frequency cepstral coefficients (MFCC) [28]. There are also some features aiming at capturing the vocal cords vibration characteristics, such as the fundamental frequency F0 [6] [34][98] and the harmonic intensity information [47]. Unlike the speech recognition task, where it is believed that different sounds are mainly related to formant structure of the vocal tract system, in speaker recognition, a number of experiments have shown that the vocal cords vibration style carries rich speaker specific information and is useful for speaker recognition. This thesis focuses on developing effective techniques to extract the vocal source

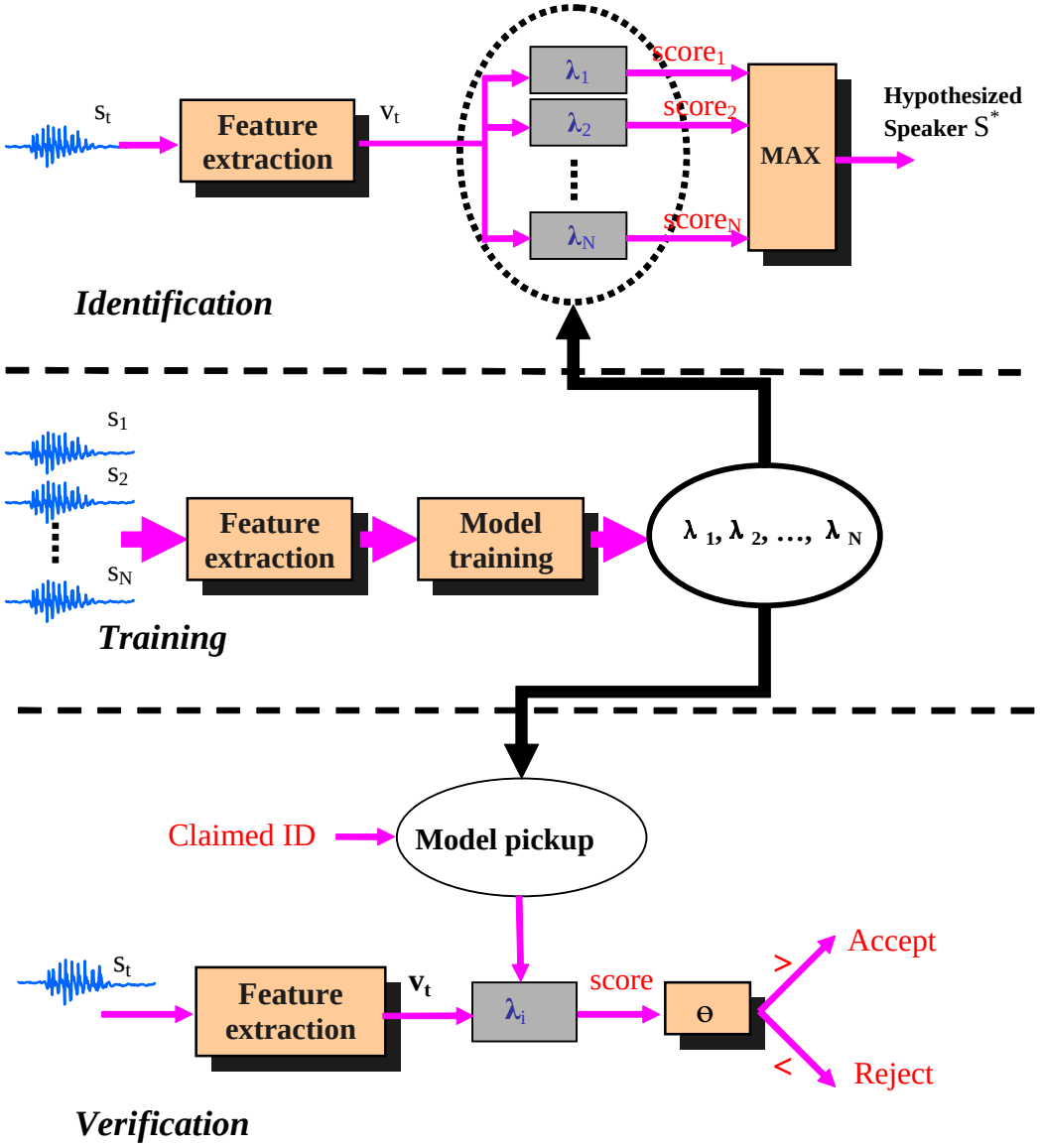


Figure 2.1: Speaker Recognition system structure

Table 2.1: Hierarchical features for human and machine speaker recognition

Physical/social filiation	Perceptual cues for human	Features for ASR	Feasibility in ASR
socio-economic status, education, place of birth, etc.	accent, diction, semantics, idiosyncrasies, etc.	word, phrase and syntax usage, ...	high level features, effective feature representation is pending
personality type, parental influence, etc	speaking style, prosodics, rhythm, intonation, volume modulation, speaking rate, etc.	f0 contour, energy fluctuation, pause, duration, etc.	moderate level features, have been used to supplement the low level features
anatomical structure of vocal apparatus	acoustic aspect of speech, nasal, deep, breathy, rough, harsh, etc.	f0, harmonics, spectrum envelop, energy, etc.	low level features, widely and effectively used in current ASR

excitation related speaker-specific information to improve the speaker recognition performance of the conventional system that uses only vocal tract features. Vocal source and vocal tract related feature extraction techniques will be expatiated in Chapter 3.

To date, high-level information has not been commonly implemented in speaker recognition, mainly due to the difficulty in automatic and quantitative measuring of this kind of high level information. Nevertheless, recently, joint efforts have been put together from more than 10 institutes, such as MIT, IBM, OGI, and so on, to exploit the effectiveness of the high level information for accurate speaker recognition. In this joint project, wide ranging approaches using pronunciation models, prosodic dynamics, pitch and duration features, phone streams, and conversational interactions were explored and developed. An it is showed that these novel features and classifiers indeed provide complementary information and can be fused together to drive down the recognition error [87].

Feature selection is the transformation of feature vectors to a lower dimensional feature vectors while still preserves the relevant information. This is particularly necessary for real applications where available training data are

usually restricted. A useful transform to reduce the feature dimension is the *Principal Component Analysis* (PCA) [49]. In PCA, the original feature vector is transformed into another feature space with orthogonal coordinates. The feature selection is based on the eigenvectors of the covariance matrix of the given data. That is, components in the orthogonal space corresponding to large eigenvalues are remained, while those corresponding to the small eigenvalues are discarded. Thus, the transformed feature vectors remains the most prominent information, giving an optimal representation of the original features. In addition, the orthogonality between the feature components is particularly suitable for data modeling by a diagonal-covariance multivariate Gaussian distribution, which is a necessary assumption in data modeling and parameter estimation.

Another widely applied transformation is the *Linear Discriminant Analysis* (LDA) [70]. The feature selection by LDA is based on a discriminant criteria. That is, only feature components with large inter-classed variation and small intra-class variation are remained. Such discriminant criteria are particularly compatible to speaker recognition, which is a discrimination problem rather than a representation one. There have been a number of papers demonstrate the applications of PCA and LDA for feature selection in speech and speaker recognition [48][107].

2.2 Pattern Generation and Matching

For speaker recognition, pattern generation is the process of generating speaker specific models with collected data in the training stage. Pattern matching is the task of calculating the matching scores between the input feature vectors and the given models in recognition. Generally, speaker models can be classified into two categories: the generative model and the discriminative model.

2.2.1 Generative vs. discriminative models

Generative models attempt to capture all the underlying distribution, i.e., the class centroids and the variation around the centroids, of the training data. The

most popular generative model in speaker recognition is the stochastic model, e.g. Gaussian Mixture Models (GMM) [89], Hidden Markov Model (HMM) [81], etc. The template models, e.g. , Vector Quantization (VQ) codebooks [101], can also be regarded as a generative model, although it does not model the variations. A generative model is trained to best represent the whole distribution space of the training data generated from a specific class. The training of a class model takes into account only the corresponding data, ignoring the distribution of the competing classes.

Discriminative models, on the other hand, does not necessary model the whole distribution, but the most discriminative regions of the distribution. The objective of training a discriminative model is to minimize the classification error on a set of training samples. Therefore, not only samples from the corresponding class, but also those from all the competing classes are considered when training the discriminative model for each class. Discriminative modles include multilayer perceptron (MLP) [61], polyminal classifiers [19] and support vector machine(SVM) [110], etc.

In terms of speaker recognition the discriminative models seem to be desirable since the modeling criteria is coherent with the classification objective. An other advantage of discriminative modeling is that it models only the boundary between classes where the distributions overlap each other, and ignore the regions within each class, where no overlap happens. Thus, it has greater capacity to model the variations in the boundary since it does not need to model the regions where the input features certainly belong to a specific speaker. While the generative models still try to model these certainty regions which make little contribution to the classification performance.

However, the generative models have been verified to be more appropriate for speaker recognition. Training a discriminative model requires the training sample from both the target speaker and all the competing speakers (impostors). Modelling only the boundary may discard some client information which may carry the boundary information between the target speaker and other unseen impostors. Therefore, discriminative models may work poorly for these

unseen impostors. The generative models, on the other hand, are more robust against these impostors since all the target information are retained. Furthermore, the training of discriminative models is more complicated. If the clientele are updated with some new speakers, all the discriminative models should be retrained. The generative models does not require retraining since each target model is trained independently.

In the rest of this section, we shall focus on the generative modelling technique which is adopted in this thesis. Expatiation on the discriminative model is beyond the scope of this thesis and can be referred to [20][12][55][108]. There are also training methods to take advantages of both generative and discriminative models. The incorporation of discrimination into generative models can be achieved by combining the generative and discriminative models, e.g., the Radial Basis Function (RBF) network which combines the GMM with an MLP [12], the GMM/SVM combination [56], and the HMM/MLP hybrids [15]. A generative model can also be made discriminative by choosing a different optimization criterion. For example, the maximum likelihood criterion in HMM/GMM traning can be replaced with a maximum mutual information (MMI) criterion [75] or a minimum classification error (MCE) criterion [50].

2.2.2 Generative modeling

Template models

The simplest template model is the long-term mean or centroid of feature vectors

$$\bar{x} = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.1}$$

The matching score is the distance measurement between a particular observed feature vector and the centroid of the template, i.e.,

$$d(x_i, \bar{x}) = (x_i - \bar{x})^T W (x_i - \bar{x}) \tag{2.2}$$

where W is the weighting matrix. Equation 2.2 gives the Euclidean distance if W is an identity matrix, and the *Mahalanobis distance* if W is the inverse

covariance matrix, which gives less weight to feature components with large variance. Such a long-term mean is a very coarse representative model. Essentially, it averages out the speaker variability for any specific sounds.

More accurate and complicated models are the Dynamic Time Warping (DTW) based text-dependent template models [94] and the Vector Quantization (VQ) based text-independent codebook models [101]. In DTW, the text-dependent templates must be matched to an input feature vector sequence with time warping technique, which does a constrained, piecewise mapping of the time axes to align the two sequences with a minimized distance

$$z = \sum_{i=1}^{L} d(x_i, \bar{x}_{j(i)}), j(i) \in 1, 2, ..., N \tag{2.3}$$

where $(x_1, x_2, ..., x_L)$ and $(\bar{x}_1, \bar{x}_2, ..., \bar{x}_N)$ are the input and template sequences, respectively. The length of the template sequence, N, is always different from that of the input sequence, L, due to speaking rate variability. The template matching index $j(i)$ is determined by DTW algorithm [82].

VQ modeling uses multiple templates (codebook) to represent the training feature vectors. The VQ codebook $C = \{\bar{x}_i | i = 1, 2, ..., K\}$ is designed for each enrolled speaker with the training data using standard clustering algorithms (e.g., k-means, LBG, etc. [60][42]) . Each centroid of the codebook represents a subclass of the pronunciation characteristics. The input feature vectors are quantized using the codebook of an individual speaker, and the pattern matching score is the accumulated minimum distance between the input vectors and the centroids

$$z = \sum_{j=1}^{L} \min_{\bar{x}_i \in C} d(x_j, \bar{x}_i) \tag{2.4}$$

where L is the length of the input vectors. The lack of time warping greatly reduce the computational complexity of the system. However, the clustering procedure used for forming the codebook also averages out the speaker-dependent temporal information. Nonetheless, experimental results showed that when there is not enough training data available, VQ with properly selected codebook size can be a good candidate for data modelling [69].

Stochastic models

Unlike the template models, the stochastic models aim at representing the distribution, i.e., the centroid (mean) and the scattering around the centroid (variance) as well, of the feature vectors in a multi-dimensional space. The pattern matching can be formulated as measuring the probability density (or the likelihood) of an observation given the speaker model. Usually, the underlying distribution of the data is assumed to be Gaussian. And the multivariate Gaussian probability density is adopted for parametrization. As for speaker recognition, the Gaussian Mixture Model (GMM) has been the most popular clustering technique, especially in text-independent tasks [89]. The likelihood of an input feature vector given by a specific GMM is the weighted sum over the likelihoods of the M unimodal Gaussian densities

$$P(x_i|\lambda) = \sum_{j=1}^{M} w_j b(x_i|\lambda_j) \tag{2.5}$$

where $b(x_i|\lambda_j)$ is the likelihood of x_i given the j-th Gaussian mixture

$$b(x_i|\lambda_j) = \frac{1}{(2\pi)^{D/2}|\Sigma_j|} exp\{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1}(x_i - \mu_j)\} \tag{2.6}$$

where D is the vector dimension, μ_j and Σ_j are the mean vectors and covariance matrices of the training vectors. The mixture weights w_j are constrained to be positive and must sum to one. The parameters of a GMM, w_j, μ_j and Σ_j, can be estimated from the training feature vectors using the maximum likelihood criterion, via the iterative Expectation-Maximization (EM) algorithm [10].

A GMM is similar to a vector quantizer in that the mean of each Gaussian density can be regarded as a centroid among the codebook. However, unlike the VQ approach, which makes"hard" decision (only a single class is selected for each feature vector) in pattern matching, the GMM makes a "soft" decision on mixture probability density function (pdf). This kind of soft decision is extremely useful for speech to cover the time variation.

In text-dependent speaker recognition system, or in applications where the spoken text is known, a finite state hidden Markov model (HMM) [81] is preferred to GMM since HMM is capable of modelling the temporal variations in

the utterances. The likelihood of a sequence of speech frames given a HMM model can be calculated as

$$P(x(1:L)|\lambda) = \sum_{\substack{\text{all-state} \\ \text{sequences}}} \prod_{i=1}^{L} b(x_i|s_i)p(s_i|s_{i-1}) \tag{2.7}$$

where s_i is the state of the i-th vector, $p(s_i|s_{i-1})$ is the transition probability from the state of $i-1$-th vector to the state of the i-th vector. Each state of the HMM can be associated with a specific phoneme or a phonetic class. The temporal information along the pronunciation is encoded as the transition of the allowed state sequences. Therefore, time variation in uttering each phoneme or phonetic class can be represented by the total time spent on the corresponding state. In this case, the temporal variation is modelled and speaking rate variability is allowed.

In text-dependent speaker recognition, the *prior* knowledge of speech content makes the HMM capable of modelling not only the speech sound units, but also the temporal sequence among these sounds. However, the advantage of HMM in modelling temporal structure vanishes in text-independent speaker recognition, where the sound sequence in the training utterances is usually different from that in the testing utterances. Unlike HMM, GMM models only the sound units, neglecting the temporal sequence information. A GMM can be regarded as providing an implicit segmentation of the sound units without labelling the sound classes. The sound ensemble is classified into acoustic classes, each of which represents some speaker-dependent vocal system configurations, and is modelled by a couple of Gaussian mixtures. GMM has been the modeling technique state of the art in speaker recognition. Its superiority over other modeling techniques in text-independent speaker recognition has been demonstrated and widely accepted by the research community.

The UBM-GMM training technique

The UBM-GMM method has been adopted as a dominant method for training speaker models in text-independent speaker recognition [68][88]. Instead of training the speaker GMM directly using the training speech, this method

adapts the speaker model from a universal background model (UBM). The UBM is a speaker-independent GMM trained using speech data from a large number of speakers (usually hundreds or even thousands to represent the universal speaker set). Generally, the UBM should have a large number of mixtures so as to model all the possible pronunciation of sound units by many speakers. By adapting the speaker model from the UBM using the corresponding speech data, speaker specific information can be emphasized and represented in some of the mixtures, while the other mixtures remain unchanged or slighted changed. The advantage of the UBM-GMM training technique is that the UBM can be used for score normalization in classification (see next section).

Another advantage of adapting the speaker model from the UBM is that it requires less training data than training the speaker model directly. This is extremely beneficial in real implementations where the amount of training data from the speakers is usually restricted. In this case, one can train a UBM using existing publicly available speech databases. Then the speaker models can be adapted from the UBM with relative less training data. This database for training UBM should be in the same language as that for adapting the speaker models. And it is preferred if the data collection conditions are also the same or at least quite similar, i.e., both are telephone speech or wide-band microphone speech.

2.3 Classification

2.3.1 Multi-classes classification for speaker identification

For speaker identification, the classification is quite simple. The hypothesized speaker is the one whose model best matches the input feature vectors. Or in an open-set identification, the speaker is classified as unknown speaker if the highest matching score is lower than a preset threshold.

2.3.2 Binary classification for speaker verification

For speaker verification, the classification is a binary choice decision. That is, given the matching score of the input feature vectors against the claimed speaker model, the system should make a decision between two hypotheses: the input speech is from the claimed speaker, $H1$, or from an impostor, $H0$. It has been shown that the performance can be greatly improved by normalizing the raw speaker model scores over the background speaker model scores. Taking the stochastic models as an example, the decision can be made upon the log-likelihood ratio

$$\Lambda = \log \frac{P(x_i|H_1)}{P(x_i|H_0)} = \log P(x_i|H_1) - \log P(x_i|H_0) \overset{?}{>} \theta \tag{2.8}$$

where $P(x_i|H_1)$ is the probability of an observation x_i generated by the claimed speaker, and $P(x_i|H_0)$ is the probability of the observation NOT generated by the claimed speaker. The threshold θ can be determined by

1. Setting $\theta = p_1/p_0$, where p_1 and p_0 are the *a priori* probabilities that input speech is from the true speaker and from the impostor, respectively.

2. Choosing θ to satisfy a fixed false accept (FA) rate or false reject (FR) rate according to the Neyman-Pearson criterion.

3. Experimentally determined in developing stage. That is, Varies θ to find different FA and FR and choose θ to give the desired FA/FR ratio. This method has been adopted in most speaker verification systems.

2.3.3 Background model for score normalization in speaker verification

In equation 2.8, the calculation of the impostor probability $P(x_i|H_0)$ requires an impostor (or background) representation. Generally, there are two kinds of background representations, the speaker-dependent background sets and the speaker-independent UBM. For speaker-dependent background selection, the system finds the most alike speakers, who have the least inter-speaker distance

from the corresponding target speaker, representing the most competitive impostors, as well as the most unlike speakers representing the most dissimilar impostors [89][92].

The UBM has been demonstrated to outperform the speaker-dependent backgrounds [85]. We can assume that information associated with speech signal can be classified as either speech specific (common to all speakers) or speaker specific. By adapting speaker model from the UBM, speaker specific elements can be emphasized and represented in some of the mixtures. Thus normalizing $P(x_i|H_1)$ over $P(x_i|H_0)$ will emphasizes the speaker specific frame scores. A frame common to all speakers contributes less to speaker discrimination but may still has a high probability $P(x_i|H_1)$. Normalizing it over $P(x_i|H_0)$, which should be closed to $P(x_i|H_1)$, results in a small Λ. Contrarily, a speaker specific frame will have a large Λ if the claimed speaker is the true speaker, and still small if it is not. What's more, the UBM trained using a large scale database can be regarded as representing all phonetic classes speaker-independently. Thus it is more robust to both the unseen vocabulary of the target speaker and unseen impostor as well.

2.4 Performance Evaluation Metric

The commonly used performance evaluation metric for speaker identification is the identification error rate (IDER)

$$\text{IDER} = \frac{\text{number of misidentified trials}}{\text{total number of identification trials}} \times 100\% \tag{2.9}$$

For speaker verification, the decision threshold should be selected to make trade-off between the false acceptance (FA) and false rejection (FR) errors. Generally, the threshold is selected such that the false acceptance rate equals to the false rejection rate, usually referred as the equal error rate (EER). Another commonly adopted metric is the detection cost function (DCF), which takes into account the cost of FA and FR [67],

$$\text{DCF} = C_{FR} \cdot P_{FR} \cdot P_{Target} + C_{FA} \cdot P_{FA} \cdot P_{NonTarget} \tag{2.10}$$

where C_{FR} and C_{FA} are the detection error cost for false rejecting a target speaker and false accepting an impostor. P_{FR} and P_{FA} are the probabilities of false rejection and false acceptance. P_{Target} and $P_{NonTarget}$ are the probabilities of target and impostor tests among all the tests.

The EER and DCF define two particular trade-offs between false acceptance and false rejection. A pictorial illustration of the performance of the classifier is the Receiver Operating Characteristic (ROC) curve [32]. An ROC curve illustrates the trade-off between the two error types. The false acceptance probability is plotted on the horizontal axis and the correct acceptance probability (equals to the one minus the false rejection probability) is plotted on the vertical axis. Figure 2.2 shows an example of two ROC curves. In an ROC curve, each point corresponds to a particular decision threshold. The closer an ROC curve to the upper-left corner, the better the system performance.

Another commonly used pictorial illustration of the system performance is the Detection Error Trade-off (DET) curves [66], as shown in Figure 2.3. A DET curve plots on its vertical axis the false rejection probability instead of the correct acceptance in ROC curve. It depicts the false rejection vs. false acceptance in different decision thresholds. A large threshold corresponds to the upper-left region, where high security is the prerequisite. On the other hand, a small threshold corresponds to the lower-right region where user convenience are preferred. The curves are plotted in the normal scale. Therefore, they should be straight lines if both target and impostor scores are Gaussian distributed. One can easily compare the system performance with their DET curves. A curve closer to the original point corresponds to a better classifier.

2.5 Speaker Recognition over Telephone Network

Today, most of the speaker recognition systems work very well in laboratory experiments or under some sophisticatedly designed training and testing conditions. However, the performances of these systems degrade dramatically in real

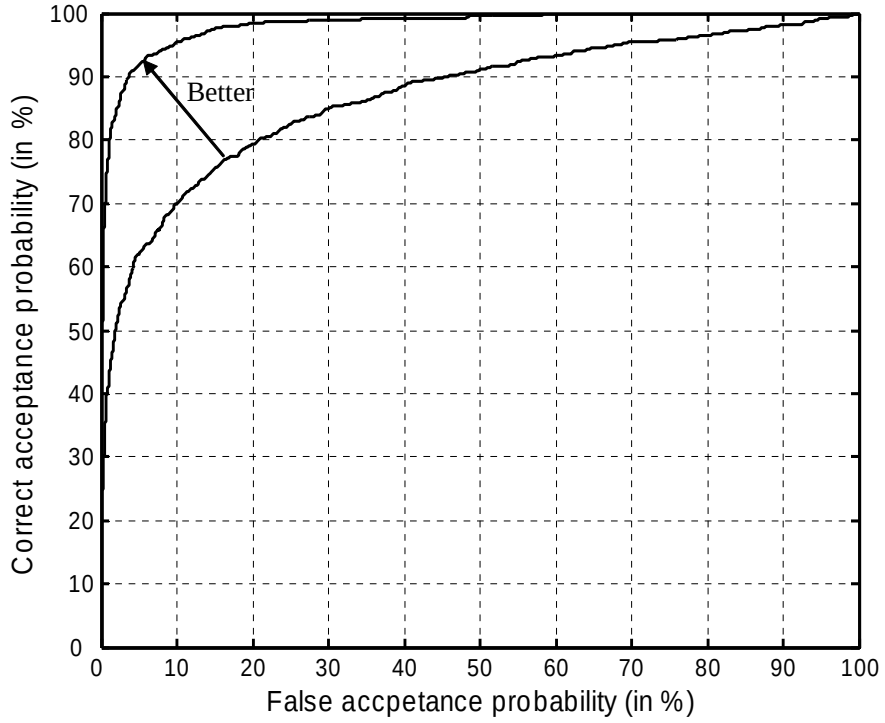


Figure 2.2: ROC curves of two speaker verification classifiers

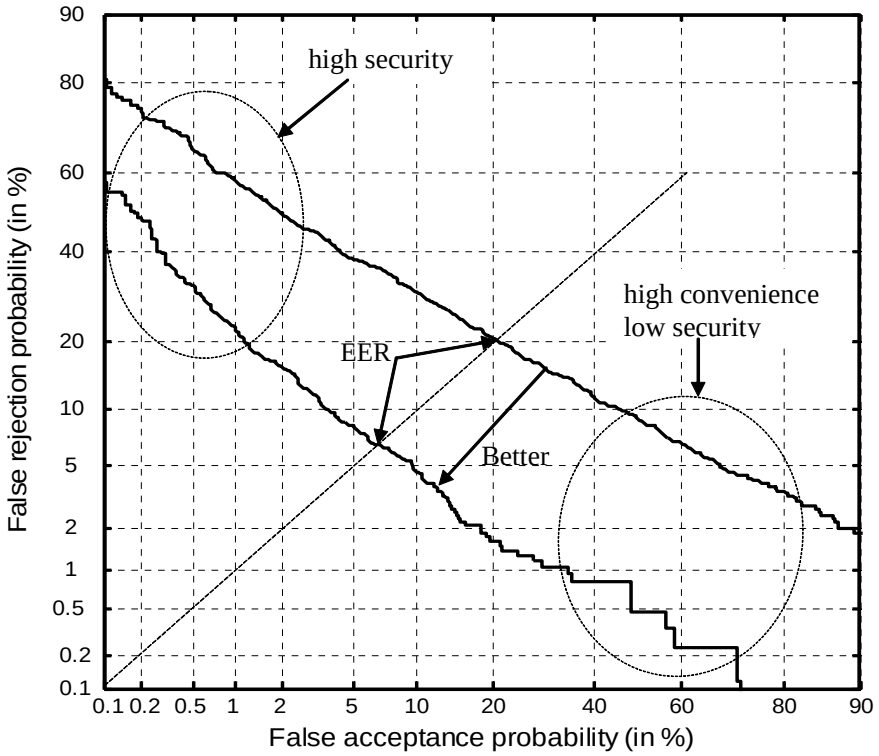


Figure 2.3: DET curves of two speaker verification classifiers

implementations where a certain degree of mismatching conditions always exists. For example, in telephone speech, the mismatches include the additive background noise, the convolutive channel and handset distortion, and others. In the past years, many research efforts have been devoted to tackle the mismatch problems and significant improvements have been reported for speaker recognition over telephone network [43][87][89][119]. The commonly applied methods can be classified into three categories, corresponding to the three components of speaker recognition: (1) robust feature extraction and feature transformation, (2) model transformation, and (3) score normalization.

Robust feature extraction and transformation

The objective of robust feature extraction is to find new features which are robust to the mismatch conditions. These new features can either be derived from the conventional feature parameters, i.e., incorporating auditory models in feature extraction [44][52], calculating the dynamic features [41][100], etc; or new feature parameters such as pitch and other high level features as listed in Table 2.1. These new features are believed and have been verified by some experiments to be more robust to noise and channel distortions than the conventional acoustical features [21][87]. This thesis also focuses on extracting new robust features for speaker recognition. It will be demonstrated in the following chapters that our new feature set, instead of extracting the vocal tract related speaker-specific information, effectively captures the spectro-temporal characteristics of the vocal source excitation. The new feature parameters provide additional speaker discrimination to the conventional vocal tract features and improves the robustness of the speaker recognition system.

Feature transformation approaches attempt to modify the distorted feature to represent the clean speech better. These approaches include cepstral mean normalization (CMN) [7] and signal bias removal [84]. In CMN and signal bias removal, it is assumed (most of time it is true) that for a specific telephone conversation, the channel and handset distortion vary much slowly compared with the variation within speech. Thus, the frequency response of the handset and

channel can be assumed to be time invariant and be approximated by the long-term average of distorted cepstral vectors. These approaches, however, do not consider the effect of background noise. Hermansky *et al* proposed a spectral filtering technique, called Relative Spectral (RASTA), to tackle speech recognition problems with both background noise and channel distortion [44][45]. With the knowledge of signal-to-noise ratio (SNR), it either reduces the additive background noise in low SNR, or reduces the convolutive channel distortion in high SNR. The effectiveness of RASTA relies upon the accuracy of SNR estimation, which limits its practicality. Other approaches include the codeword-dependent cepstral normalization (CDCN) and the SNR-dependent cepstral normalization (SDCN) [1]. In CDCN, the additive noise and the convolutive distortion are modeled as codeword-dependent cepstral biases. In SDCN, SNR-dependent cepstral biases are estimated in a maximum likelihood framework, and it also requires an accurate noise estimation.

Model transformation

The model transformation approaches modify the clean speech models such that the density functions of the resulting models fit the distorted data better. These approaches include: (1) the stochastic transformation [90][95], where the clean models' means and variances are adjusted by stochastic biases; (2) the maximum likelihood linear regression (MLLR) [58], where only the means of clean speech models are linearly transformed; and (3) the constrained re-estimation of Gaussian mixtures [29], where both means and covariance matrices are re-estimated.

Recently, speaker recognition systems combining both feature and model transformation were reported to improve the overall performance. For example, a system that combined handset selector with feature and model transformation claimed to be able to outperform the conventional approaches and significantly reduce recognition errors under several different coders with bit rates ranging from 2.4 kbps to 64 kbps [118][119]. In this system, Coder-dependent GMM-

based handset selectors are trained to identify the most likely handset used by the claimants. Then stochastic feature transformation and model transformation are applied to reduce the acoustic mismatch between different handsets and speech coders. A recently published book by Kung *et al* [55] gives a good review on feature and model transformation for handset and channel distortion in speaker recognition and serves as a useful literature for this topic.

Score normalization

In section 2.3.3, the purpose of score normalization is to reduce the speaker-independent speech mismatch bias so as to find a global speaker-independent threshold for the decision making process. When there are handset and channel distortions, the log-likelihood ratio (LLR) scores Λ, calculated with different speaker models, can still have handset- or channel-dependent biases. This makes the global threshold unacceptable. To avoid or to reduce these biases, several score normalization methods have been developed [9][88]. In these methods, the mean and standard deviations of the LLR score distribution (which is usually Gaussian) are first estimated in training stage and then used for score normalization in recognition. The normalized LLR score is

$$\Lambda^{NORM} = \frac{\Lambda - \mu}{\sigma} \tag{2.11}$$

where μ and σ are the mean and standard deviation estimated. They can be handset- or channel- dependent or independent according to the application conditions. Generally, μ and σ are estimated only from the LLR scores of the impostor such that the distribution of Λ^{NORM} of the impostor will be zero mean, unit variance gaussian distribution. Experiment result shows that after normalization, the LLR score distributions of the target speaker and impostor have less overlap than the LLR score distributions before normalization [88].

Chapter 3

Speech Production and Feature Extraction

Speech is a kind of very special signal among all the signals in the world. It is one of the public, outer form of language. It is also the most efficient information carrier for human communication. Biologically, speech communication is supported by three neurophysiological systems: (1) the central nervous system, particularly Broca's Area, Wernicke's Area, and Planum Temporale, which are involved in the production and understanding of human speech [114]; (2) the motor system, particularly the speech tract including the mechanisms for respiration, phonation and articulation, which produce speech as modulated air pressure with certain temporal and spectral structures [14]; and (3) the sensory system, particularly the ear and the auditory pathways, which decode the speech signal into neural signal such that they can be understood by the central nervous system [5][114].

This chapter only focuses on the motor system of speech production. Section 3.1, reviews the biological process and the acoustic theory of speech production. The acoustic cues for speech and speaker recognition are also discussed. Section 3.2 introduces some techniques for separating the vocal source and vocal tract components from the speech signal. Section 3.3 outlines the prevalent feature extraction techniques for extracting the vocal source and vocal tract related acoustic features for speaker recognition. Finally the effectiveness of these two

kinds of features for speaker recognition is compared and their complementarity for speaker recognition is discussed in section 3.4.

3.1 Speech Production

3.1.1 The three phases of speech production

Speech production is a physiologically complex activity involving coordination of three functionally distinct systems: (1) the subglottal lungs, (2) the larynx, and (3) the supralaryngeal vocal tract. Accordingly, the process of speech production can be classified into three phases: (1) respiration, which provides the air stream and is the primary source of power for phonation; (2) phonation, which changes the steady air stream into quasi-periodic pulsate signal, rich in harmonic structure; and (3) articulation, which modulates the signal from the larynx in its temporal and spectral structure, resulting in different categories of sounds. This section will give a brief review of the three phase of speech production. The details can be find in a number of references, e.g. [14][59][77].

Respiration

The respiratory activity for producing speech requires finely controlled contractions of the intercostal musculatures and the diaphragm. This is markedly different from the relatively passive process in quiet breathing. The air inspiration into the lungs is achieved by contracting the diaphragm to produce a negative pressure in the thoracic cavity. In most languages speech is produced during expiration which suppllies the egressive air stream upon which signal is superimposed.

Phonation

In normal breath, the air stream passes unconstrainedly through the pathway thus produces little sound. The larynx, particularly the vocal cords, provides two kinds of excitation for speech production: *turbulent noise* or quasi-periodic *phonation*. When the vocal cords close sufficiently with only a small portion

open, the constrained airflow passes through the open portion (also named as the *glottis*), creating a turbulent noise from which the aspiration or whisper speech is produced. When the vocal cords behave as a quasi-periodic opening and closing vibration, pulse sequences are generated from which voiced sounds are produced. The rate of the periodic phonation is called the *fundamental frequency*, F0. The phonation (the vibration of the vocal cords) is the result of joint effects of the subglottal and supraglottal air pressure difference, the laryngeal muscle tension, the elasticity of vocal cords, and the *Bernoulli force*. In producing modal voice, these joint effects cause a gradient opening and an abrupt closing of the vocal cords resulting in an asymmetric glottal pulse. The abrupt closing enriches the high frequency components and results in a glottal spectrum with rich harmonics decaying in a -12dB/octave trend.

There are three principal types of phonation in normal speech: modal, pulse and loft registers, as illustrated in Figure 3.1. The modal register has been described above. It occurs most frequently in normal phonation. The pulse register is also called *glottal fry*, which occurs in almost every utterance in some speakers. The degree of pulse register usage is according to mood, level of fatigue, or the misuse of the laryngeal system of individual speakers. It is characterized as very low and irregular $F0$, very long glottal close phase, and 1-3 overlapping pulses. The loft register is also called *falsetto*, which happens rarer than pulse register in normal speech, but often observed in singing, shouting, and gentler activities such as yodelling, giggling, and laughing. It is often identified with high $F0$ and the vocal cords often do not close completely due to the very fast vibration.

In addition to the normal types of voice just described above, another two voice types are: (1) the breathy voice, in which the posterior parts of vocal cords are keeping open, allowing only the anterior parts to vibrate. The breathy voice is often accompanied with falsetto where the vocal cords are not completely closed; and (2) creaky voice, in which the vocal cords are tightly closed, allowing only a small parts to vibrate. Creaky voice is a mixture of modal and pulse registers.

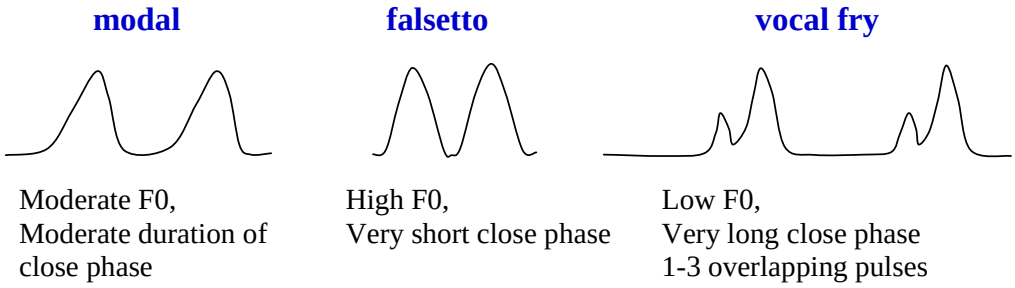


Figure 3.1: The three principal types voice register

Articulation

The vocal tract, including the oral, pharyngeal and nasal cavity, is the the most important component in speech production. The vocal cords classify the speech into voiced or unvoiced sounds by their periodicity. But the vocal tract articulation, involving the systematic movement of the articulators such as tongue, lips, velum and palate, shapes the sounds into a rich variety of phonetic classes. The vocal tract articulation for generating different sounds is achieved by varying the manner and the place of articulation. The manner of articulation is concerned with the airflow path (emitted from lips or nostrils or both) and the degree of constriction (the narrowness of the vocal tract cavity, mainly correlated to the higher or lower position of tongue). In most languages the voicing and the manner of articulation partition phonemes into broad categories such as: vowel, nasal, stop, fricative, glide, rhotic, etc. The place of articulation is concerned with the point of narrowest vocal tract constriction. It enables further discrimination and enrich the total phonemes.

The vocal tract system can be regarded as a filter with various resonance, or formants, and antiresonances determined by the manner and place of articulation. As air flow passing through the vocal tract, the filter amplifies energy around the formant frequencies, while attenuating energy around the antiresonances between the formants. Finally the air flow is emitted at the lips and/or nostrils with various pressure constituting the speech signal.

3.1.2 Acoustic theory and digital model of speech production

As described in the last section, the speech production process can be summarized as an excitation source from the larynx output modulated by the vocal tract system and finally radiated through the lips and/or nostrils. This process can be acoustically simulated by a source-filter model [35]. That is, a filter system, including a vocal tract model $V(z)$ and a radiation model $R(z)$, is excited by a glottal excitation signal $u(n)$, as illustrated in Figure 3.2.

The glottal excitation $u(n)$ has two types: unvoiced or voiced excitation. The unvoiced excitation is generally a random noise and is always modeled by a simple gaussian noise generator. The probability density of the noise samples does not appear to be critical in speech production. Thus, only speech model with voiced excitation will be described in the following. The voiced excitation is modeled as an impulse sequence exciting a low pass filter $G(z)$ to produce a glottal waveform $g(n)$. Figure 3.3 shows an simplified glottal waveform of an modal voice type. As shown, the Fourier spectrum of $g(n)$ has rich harmonics with a decaying slope of -12 dB/octave. Generally, $g(n)$ has only finite length, its z-transform $G(z)$ has only zeros. Nonetheless, $G(z)$ can be approximated by an all-pole model. For example, $G(z)$ of an modal phonation can be well approximated by an two-pole model [63], i.e.,

$$G(z) = \frac{1}{(1 - z_a z^{-1})(1 - z_b z^{-1})} \tag{3.1}$$

And for breathy phonation, an extra pole was required, i.e.,

$$G(z) = \frac{1}{(1 - z_a z^{-1})(1 - z_b z^{-1})(1 - z_c z^{-1})} \tag{3.2}$$

The extra pole results in a steeper spectral slope (-18 dB/octave) [54].

During speech production, the vocal tract can be approximated as a series of concatenated tubes with various cross-section areas [35]. The first tube starts at the glottis and the last tube(s) end(s) at the lips and/or the nostrils. The number of tubes and the diameter and length of each tube, which acoustically determined the resonance and antiresonance of the tubes, are determined by

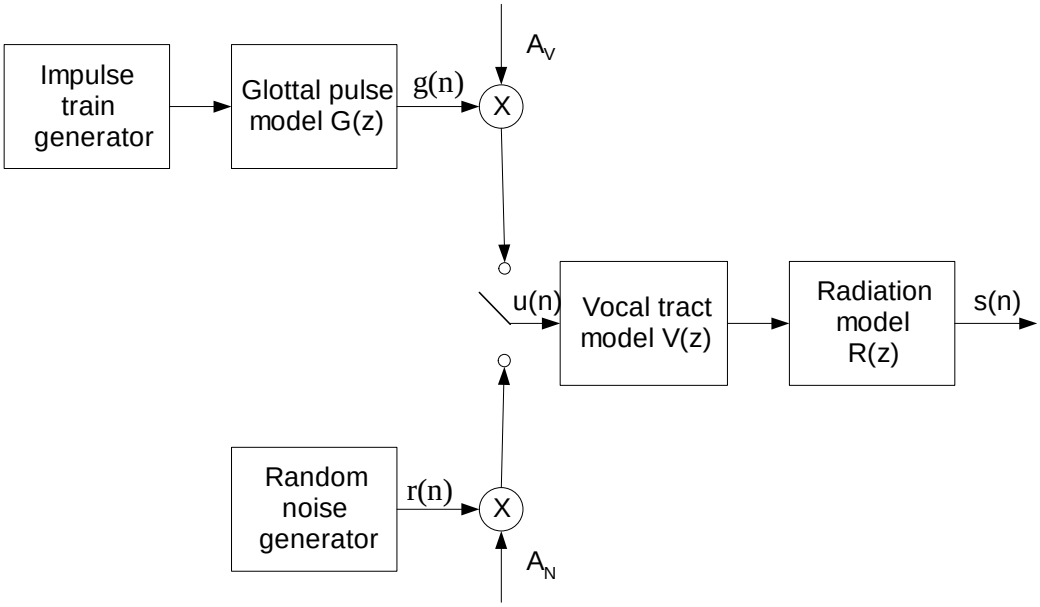


Figure 3.2: Acoustic model for speech production [83]

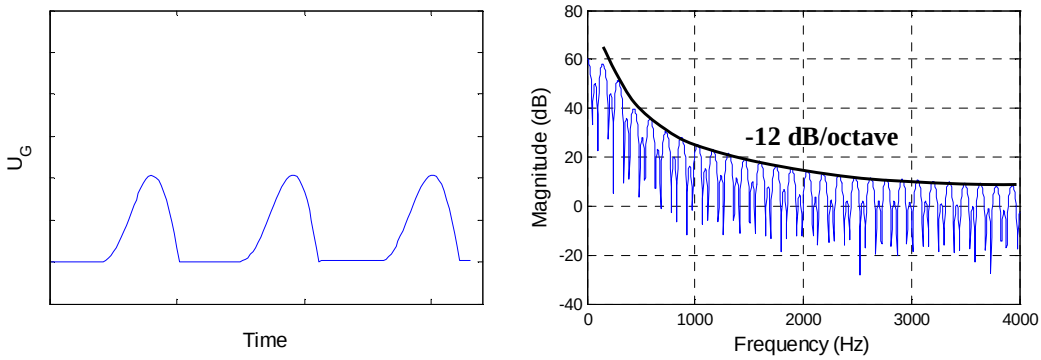


Figure 3.3: Simulation of a typical glottal pulse sequence waveform and its Fourier spectrum. The glottal pulse is simulated by integrating the L-F model waveform [36]

the manners and the places of articulation. In the source-filter model, Such a concatenated tube series is generally assumed to be characterized by an all-pole model with transfer function [83]

$$V(z) = \frac{1}{1 - \sum_{i=1}^{M} a_i z^{-i}} = \frac{1}{\prod_{i=1}^{M} (1 - p_i z^{-1})} \tag{3.3}$$

The poles always appear as conjugate pair, each pair models a formant in the speech spectrum. A typical pair of complex conjugate poles would be $e^{-\sigma_i T \pm j2\pi F_i T}$, where F_i is the formant frequency and σ_i/π is its bandwidth.

The radiation effect is often approximated as a first order difference

$$R(z) = R_0(1 - z^{-1}) \tag{3.4}$$

Thus the z-domain representation of voiced speech production can be formulated as

$$S(z) = A_v G(z) V(z) R(z) \tag{3.5}$$

It should be noted that in this model, the source excitation and the vocal tract modulation are assumed to have separate acoustic effects and can be manipulated independently. In fact, they interact acoustically to a certain degree. For example, the glottal flow during its open phase has an effect on the lower formants (most of time, the $F1$) by widening the formant bandwidth. On the other hand, the first formant modulation on the glottal flow results in the perturbation or ripple in the flow [4]. Nevertheless, the source-tract interaction plays a secondary role in speech production and can be ignored in most speech analysis and application. The independent assumption simplifies the model and permits independent analyses of the glottal and supraglottal systems. However, this kind of interaction could be problematic in some speech processing scenarios. For example, in speech analysis which aims at an accurate glottal flow estimation from speech signal, the interaction should be considered. The glottal excitation estimation can be achieved by inverse filtering. The filter coefficients should be estimated, via linear predictive analysis, within the glottal close phase, to avoid the error caused by the interaction.

3.1.3 Speech and speaker relevant acoustic cues

The supralaryngeal vocal tract configurations provide the primary acoustic cues for phoneme classification. By moving the articulators, mainly the tongue, the vocal tract is altered to be with different configurations, so as to produce different sounds. The vocal tract configurations, so as the phoneme classes, are acoustically associated with the formant structure (central frequency and bandwidth), which is the most important element in constituting the spectrum. The formant structures are relatively stable for a specific phoneme across different speakers. Thus the features derived from the vocal tract spectrum carries the most useful information for speech recognition.

The secondary function of vocal tract articulation is to *color* the speech production with the personality of speakers. Different speakers may have their own style of articulation, resulting in inter-speaker variations in speech spectrum, which can be used for speaker recognition.

The contributions of source excitation in linguistic distinguishing include voice/unvoice classification, sentence melody (intonation) and in tonal languages the tonal forms of words. The phonation types are only related to the speech quality. Different speakers generally hold different phonation types resulting in various voice types, such as modal, breathy, creaky, harsh, etc.

Our auditory system can tolerate the variations of the voice types and the inter-speaker vocal tract configuration as well and make little confusion in distinguishing different sounds. At the same time, the speaker specific larynx configurations, as well as inter-speaker variations of vocal tract structure provide speaker relevant acoustic cues upon which we recognize a familiar speaker without other physical traits.

As to the automatic speaker recognition, feature extraction techniques have been developed to effectively representing the vocal tract configurations by some acoustic features, e.g. MFCC or LPCC. These features, together with some appropriate modelling methods, have been successfully applied for speaker recognition. However, there is still lack of an effective feature extraction technique for the vocal source excitation signal and its usefulness for automatic speaker

recognition has not been thoroughly studied.

3.2 Source-Tract Separation of Speech Signal

In speech analysis and processing, we usually need to characterize the vocal source excitation and the vocal tract modulation separately. However, most of the time we can only observe the speech signal, which is the convolution output of excitation source and vocal tract impulse response (in fact, it is possible to collect the excitation signal directly by the electroglottography (EEG) technique [24][37]) Thus it would be useful to separate the two components from the speech signal. This can be realized in two ways: homomorphic deconvolution and linear predictive (LP) analysis.

3.2.1 Homomorphic deconvolution

As described in section 3.1.2, the speech signal is the convolution output of the vocal source excitation signal and the impulse response of the vocal tract filter system. The formant structure of the vocal tract varies relatively slowly compared to the harmonics of the glottal pulse. Therefore, it is possible to separate the two components from the speech signal by the so called homomorphic deconvolution, or cepstral analysis [83]. The speech signal can be formulated as $s(n) = u(n) * h(n)$, where $u(n)$ is the excitation signal and $h(n)$ is the impulse response of the filter system including the vocal tract modulation and lips radiation. The convolution of $u(n)$ and $h(n)$ corresponds to the product of their Fourier transform $S = UH$. The logarithmic operation transform the product of the two spectra into a sum of the two log-spectra $log(S) = log(U) + log(H)$. Since H changes much faster than U. Contributions of U and H is separable in cepstral domain, which is realized by doing inverse Fourier transform of the log-spectrum. The process of cepstral analysis of $s(n)$ can be formulated as

$$\hat{s}(n) = IDFT(log|DFT(s(n))|) = \hat{u}(n) + \hat{h}(n) \tag{3.6}$$

where $\hat{s}(n)$, $\hat{u}(n)$ and $\hat{h}(n)$ are the cepstrum of $s(n)$, $u(n)$ and $h(n)$, respectively. Thus the convolution of $u(n)$ and $h(n)$ in time domain now becomes additive in

cepstrum domain. It has already been pointed out that $\hat{h}(n)$ usually decreases very fast as n increased, that is, most of the energy of $\hat{h}(n)$ is concentrated near the region of $n = 0$. While for the voiced excitation, $u(n)$ can be approximated as

$$u(n) = \sum_{i=0}^{R} \delta(n - iN_p) \tag{3.7}$$

The cepstrum $\hat{u}(n)$ will also be periodic pulses with period N_p

$$\hat{u}(n) = \begin{cases} nonzero, n = iN_p \\ 0, otherwise \end{cases} \tag{3.8}$$

Figure 3.4 shows the cepstral analysis of both voiced and unvoiced speech segments. As illustrated, the magnitude of $\hat{s}(n)$ decrease quickly as n increases. However, for voiced speech, a large amplitude which corresponds the $\hat{u}(n)$ appears at the position about one pitch period. Thus, by liftering $\hat{s}(n)$ with a low-pass lifter

$$L(n) = \begin{cases} 1, n < N_p \\ 0, n \geq N_p \end{cases} \tag{3.9}$$

$\hat{u}(n)$ and $\hat{h}(n)$ can be separated from $\hat{s}(n)$. As illustrated in Figure 3.4, the cepstrum of voiced speech has an impulse at the around cepstral index of 50, which corresponds to the period of the speech.

The homomorphic deconvolution generates the cepstra of the source excitation signal and the vocal tract system separately. However, we cannot reconstruct the speech signal from the two cepstra since phase information is lost after the $|*|$ operation. Nevertheless, cepstral analysis has many applications in speech analysis and processing such as speech and speaker recognition, voicing/unvoicing detection and pitch detection [74], etc.

3.2.2 Linear predictive analysis

In Linear predictive analysis of speech, the vocal system is assumed to be approximated by an all-pole model,

$$H(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.10}$$

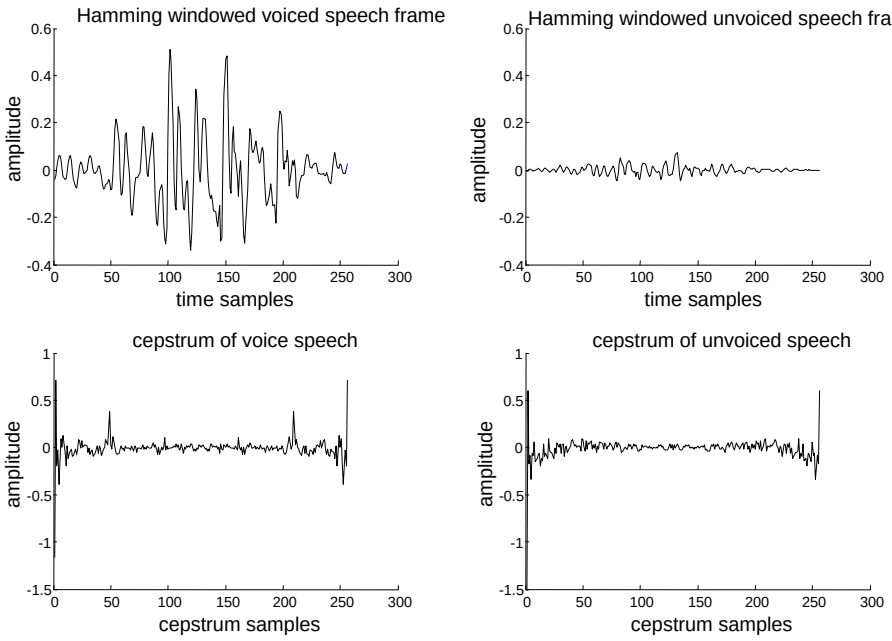


Figure 3.4: Cepstrum analysis of voiced (left column) and unvoiced (right column) speech.

Thus, the z-domain speech production can be written as

$$S(z) = U(z)H(z) \tag{3.11}$$

And the corresponding time domain formula is given by

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + u(n) \tag{3.12}$$

where $u(n)$ and $U(z)$ correspond to the source excitation. The linear predictive analysis assumes that the current sample of the signal is the linear combination of the previous samples. It is therefore predictable if the previous samples are known, i.e.,

$$\hat{s}(n) = \sum_{k=1}^{p} \hat{a}_k s(n-k) \tag{3.13}$$

where $\hat{a}_k$ is the prediction coefficients and can be estimated with various methods, e.g. the autocorrelation and covariance methods [83]. Suppose the pth-order all-pole model accurately represent the underlying speech production

mechanism, and the prediction error is minimized by a proper choice of $\hat{a}_k$, i.e.,

$$\hat{a}_k = a_k, \forall k = 1, 2, \cdots, p \tag{3.14}$$

the prediction error will be equal to the excitation signal

$$e(n) = s(n) - \hat{s}(n) = u(n) \tag{3.15}$$

$e(n)$ is also called the LP residual signal. Generally, the all-pole model for generating speech signal is called the synthesis model, and its inverse version for source and tract separation is called analysis model, as illustrated in Figure 3.5.

Note that the LP analysis assumes an all-pole model for representing the combined effects of the glottal excitation, vocal tract modulation, and the lip radiation. In order to separate the glottal excitation from the vocal tract modulation, it is necessary to perform LP analysis in the glottal closing interval, where the glottal flow is zero so that the estimated LP coefficients best represent the vocal tract system. The the glottal source signal can be achieved by inverse filtering on the whole period of speech signal.

LP analysis is one of the most important techniques in speech analysis and processing. It has been implemented in many of speech related applications such as, speech/speaker recognition, coding, synthesis, etc.

One important properties of LP analysis relies on the minimum error criterion in estimating the LP coefficients. That is, the coefficients are optimized to result a minimum energy (least square) in the residual signal [62]. According to the estimation theory, the only way to achieve the least square criterion is to make the error orthogonal to the estimates, i.e. $e(n) \perp \hat{a}_k$. This will be extremely beneficial in speaker recognition using both vocal source and vocal tract related information. Feature parameters derived from $e(n)$ and $\hat{a}_k$ will be complementary to each other since they are uncorrelated.

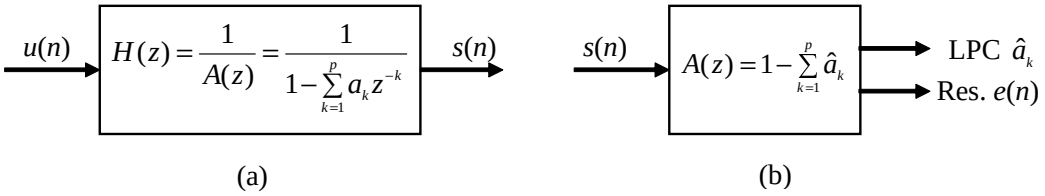


Figure 3.5: Synthesis and analysis model of speech

3.3 Feature Extraction from Speech Waveform

Feature extraction is the first step in automatic speech/speaker recognition. In speaker recognition, the goal of feature extraction is to obtain from the speech waveform the salient features which are critical to the recognition of the speaker identity. As mentioned in Section 3.1, two distinct parts, the vocal source excitation and the vocal tract system, are involved in the speech production process. Speaker-dependent features can also be classified into vocal tract related and vocal source related features. The former is derived from the spectral analysis of the speech signal, and is aimed at representing the shape of the spectral envelop, particularly the formant structure. The latter includes the time and frequency properties of the glottal pulse waveform, e.g. the fundamental frequency/pitch period, the harmonic structure, the pulse open and speed quotients, and so on.

3.3.1 Vocal tract features

The most important vocal tract related features are those derived from the short-time spectral analysis aiming at capturing the spectral envelop and thus the formant structure. There are two major branches in the short-time spectral analysis: linear predictive analysis and short-time Fourier transform (STFT). Correspondingly, there are two prevalent vocal tract features: linear predictive cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC), as described below.

LPCC

The first spectral analysis technique is the linear predictive analysis described in Section 3.2.2. The all-pole model power spectrum $1/|A(\omega)|^2$ provided an estimate of the short-time spectral envelop when p is relatively small (10-16 for speech signal). The LP coefficients can be further transformed into cepstral coefficients, i.e. the LPCC parameters. The cepstral coefficients are essentially the inverse Fourier transform of the log power spectrum. They can be derived from the LP coefficients using the following recursive relationships [7]:

$$\begin{cases} c_1 = a_1, \\ c_n = \sum_{i=1}^{n-1} (1 - i/n) a_i c_{n-i} + a_i, 1 < \mathrm{n} \leq \mathrm{p}, \\ c_n = \sum_{i=1}^{n-1} (1 - i/n) a_i c_{n-i}, \mathrm{n} > \mathrm{p} \end{cases} \tag{3.16}$$

One of the advantage of cepstral coefficients is that it incorporates the nonlinear compression of energy $log|S(\omega)|^2$, which is corresponding to the human auditory perception. More importantly, the cepstral coefficients have been demonstrated to be more robust than LP coefficients. For example, it is easy to reduce the fixed frequency response distortion introduced by the recording apparatus and transmission channels with the cepstral mean normalization techniques [7][38].

MFCC

The second major branch of spectral analysis is based on the short-time Fourier transform followed by various kinds of filter bank smoothing. Currently, the "perceptually motivated" frequency bank structure, e.g. the Mel-scale or Bark-scale frequency banks, are commonly adopted to approximate the human auditory system. As with the LP analysis, the smoothed power spectrum is further transformed into cepstral coefficients. One of the most prevalent feature parameters for speech/speaker recognition is the MFCC parameters [28].

An outline of MFCC feature extraction process is shown in Figure 3.6. The main steps are as follows

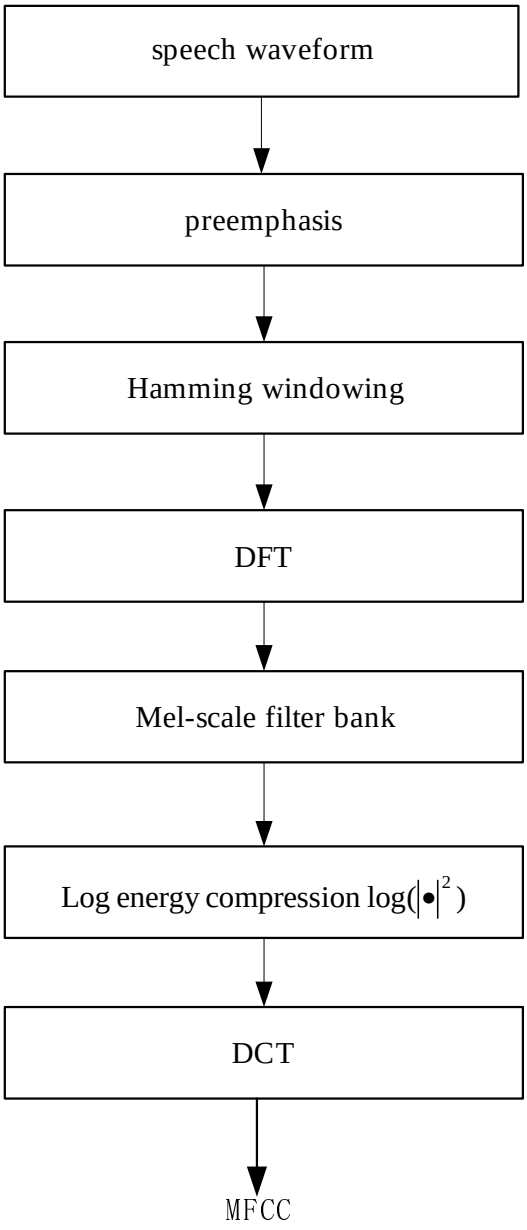


Figure 3.6: MFCC feature extraction process

- Preemphasis: to flatten the speech spectrum so as to reduce the dynamic range using a first order filter $P(z) = 1 - 0.95z^{-1}$.

- Hamming windowing: to block the speech signal into overlapped frames with minimized edge discontinuity effect. Typically a 20-30ms window length is used to tradeoff the spectral and temporal resolution. The time shift of each frame is about 10ms during which human articulatory configuration does not change dramatically.

- DFT: to transform the speech signal into the frequency domain, where the most important speech/speaker information resides.

- Mel-scale bandpass filtering: to approximate the frequency resolution of our auditory system.

- Subbank energy compression: to approximate the nonlinear compression of energy of our auditory perception. More importantly, the log operation makes the subbank energy approximately Gaussian distributed - a requirement for subsequent acoustic modelling.

- DCT: to transform the spectral information to the cepstral domain in which the energy (information) is dominated by less coefficients.

It should be pointed out that the short-time spectral analysis based features, either LPCC or MFCC, are obtained using a short time window, ignoring the dependency on the adjacent frames. However, the sequential (or dynamic) information, i.e. the change in the speech sequence, is believed to have major contribution to the perception and cognition of human speech. Thus, dynamic parameters derived from the static cepstral coefficients have been suggested and shown to improve the performance in both speech and speaker recognition [39][100]. These dynamic features include the delta-cepstrum (the first-order difference of the short-time static cepstrum), the delta-delta-cepstrum (the second-order difference of the static cepstrum), delta- and delta-delta-energy, etc. Especially, the dynamic features is verified to be more robust than the static features in noisy conditions [116].

3.3.2 Vocal source features

As mentioned in Section 3.1.3, vocal source excitation contributes little to linguistic distinguishing of phonemes except the voice and unvoice classification and tone classification in tonal languages. However, it is believed that the source excitation signal carries rich speaker specific information. Unfortunately, it is difficult to fully exploit this kind of source related information for speaker recognition, mainly due to the difficulty in its accurate estimation and the lack of effective parametrization of the true glottal pulse waveform.

The prevalent method for glottal pulse estimation is the linear predictive inverse filtering of the speech signal. While the vocal source and vocal tract are assumed to be independent to each other in LP analysis, the source and tract are actually interactive, especially during the glottal open phase. Thus only the LP coefficients estimated in the glottal closed interval, where the source-tract interaction is minimized, are appropriate for estimating the true glottal pulse. Unfortunately, automatic localization of the glottal closed phase is most of time unreliable and an EGG signal is required in assistance of close phase localization [16][109], which restricts its real time applications. There have been a number of methods presented to automatically estimate the glottal wave by inverse filtering [2][115]. These method works well in modal voice, however, the results in real speech is not convinced especially in strong breathy voice or irregular phonation.

One the other hand, unlike the vocal tract signal within which the salient speech/speaker information resides in the frequency domain and spectral analysis is appropriate for feature extraction, the glottal source signal bears rich speaker-specific information in both time and frequency domain. For example, the fundamental frequency and harmonic structure is very useful for speaker discrimination. Also, the ratio of the open phase to the pitch period, the returning time, and the pulse peak value as well may contribute further speaker information. Before going to the next chapter to introduce our source feature extraction technique, we first review some most commonly adopted source features in speaker recognition.

Fundamental frequency and pitch contour

Pitch period, the period of the vibration of the vocal cords, has been considered as one of the most important factor characterizing the glottal excitation. Pitch, or pitch contour has been verified to be useful for speaker recognition by various tests [6][34][98]. However, it has also been found that for high pitched female speakers the finer spectral structure introduced by the fundamental frequency would degrade the estimation of the vocal tract spectral envelop (particularly, the central frequency and band width of the first formant). When the average pitch varies significantly between enrollment and testing stages, the effect of pitch will lessen the speaker recognition performance [126].

Harmonics related features

The intensity of the harmonics (particularly the first some harmonics with high energy) are often considered important for the perception of vocal quality [33]. Childers et al [25][26] defined two parameters to measure the harmonic structures. The first is the "harmonic richness factor"(HRF), which measures the ratio of the intensity of the harmonics to the intensity of the fundamental frequency, i.e.

$$HRF = \frac{\sum_{i \geq 2} H_i}{H_1} \tag{3.17}$$

where H_1 and H_i are the intensity of the fundamental frequency and the harmonics, respectively. The second is the "noise-to-harmonics ratio"(NHR), which measures the ratio of the energy of the inter-harmonics noise to the energy of the corresponding harmonics, i.e.

$$NHR_i = \frac{N_i}{H_i} \tag{3.18}$$

where N_i and H_i denote the energy of the ith inter-harmonics region and the ith harmonics, respectively.

Imperl et al proposed a method to extract the harmonic structure using Hildebrand-Prony transform. He demonstrated in his experiment that the harmonic feature can improve the recognition performance in a certain degree

[47].

Besides the pitch and harmonics, features generated from the LP residual signal are believed to be very useful in speaker recognition. The LP residual signal, according to the theoretic frame of LP analysis, should be orthogonal to the synthesis filter. Thus features derived from the LP residual signal should be complementary to that derived from the synthesis filter in recognition. The generation of speaker-specific features from the LP residual signal can be classified into two categories, transform based and model base parameter estimation, as described below, respectively.

Cepstral analysis of the LP residual signal

Thevenaz et al [105] investigated the complementary properties of the LP residual signal and synthesis filter in speaker recognition. The cepstral analysis, as described in Section 3.2.1, is done on the LP residual signal. Then the peak value of the cepstrum, are used as the source feature. It is demonstrated in their experimental results that although the residual based features are not as efficient as that derived from the synthesis filter, it did further improve the overall performance of speaker recognition systems as a complementary feature to the latter.

Modeling the glottal flow derivative

Instead of doing cepstral analysis, this method intended to estimate the true glottal flow or its derivative waveform and parameterize them according some existing glottal pulse models [16][78]. In both articles, the glottal flow derivative is estimated using an inverse filter, with the filter coefficients determined by LP analysis within glottal closing interval. Then the glottal flow derivative is modeled using the Liljencrants-Fant model [36]. The model parameters estimated are used as feature parameters for speaker recognition. These methods have at least the following two shortcomings restricting their real-time applications. Firstly, automatic and accurate estimation of the glottal pulse is most of time

unpractical in real-time speech and in [16] it relied on an electroglottalgraphy (EGG) simultaneously recorded with speech for glottal close instance (GCI) localization. Secondly, the existing glottal models are only suitable for normal voice types. Some abnormal voice type, for example, that with secondary or tertiary pulses, can not be modeled. However, these phenomena are common in natural speech for some speakers and are important for speaker characterization. It seems that casting these secondary and tertiary pulses into noise, as did in [78], is not a good representation.

3.4 Comparison of Vocal Source and Vocal Tract Features for Speaker Recognition

The previous section reviews two prevalent vocal tract features and some vocal source features that have been explored for speaker recognition. Generally, the vocal tract features performs much better than the existing vocal source features for the reasons that:

- The vocal tract configurations not only determines the phonetic class, but also terms to *color* the speech waveform with speaker characteristics. The vocal tract related speaker-specific information mainly resides on the frequency domain. The vocal tract features (e.g. LPCC or MFCC) represent the phonetic information and the speaker specific information as well. The appropriate data modeling techniques (e.g., GMM, HMM and SVM, etc.) make the speaker-specific information salient for speaker recognition.

- The vocal tract configurations are comparatively more time-invariant than vocal cords vibrations. As known, the major function of the vocal tract is for phonetic classification. Therefore, a speaker cannot varies the vocal tract configuration significantly in pronouncing a specific phoneme in different time. Contrarily, the vocal cords has a larger degree of freedom in phonation for pronouncing the same phoneme. Actually, the vocal cord phonation changes greatly with other factors such as the mood and the

level of fatigue of the speaker. The larger degree of intra-speaker variation makes the vocal source characteristic a secondary candidate for speaker recognition.

- Although rich speaker information is embedded in the vocal cords phonation, they have not been efficiently used in speaker recognition due to the difficulties in glottal source signal estimation and effective feature representation techniques.

Nevertheless, the vocal source features can serves as a secondary but complementary feature to the vocal tract features. For example, in speaker verification, the performance is determined by the distribution of claimant and impostor scores. A large impostor test score will cause a false acceptance and a false rejection error happens when the claimant test score is very small, as illustrated in Figure 3.7. As mentioned above, the vocal tract configuration cannot vary significantly in pronouncing the same phoneme. Therefore, it is possible that two speakers have very similar vocal tract configurations. As a consequence, the impostor test may have a large score, which results in a false acceptance. In this case, if the vocal source signal happens to be (and it is quite possible) different between the impostor and the target speaker, the false acceptance can be avoided by incorporating the vocal source features.

On the other hand, in real applications, the background noise and channel modulations can greatly distort the speech signal and the resulting vocal tract features may not represent individual speaker correctly. Therefore, a claimant test with a distorted testing speech will result in a low score and the false rejection error arises. Some of the vocal source features, e.g. the fundamental frequency, may be comparatively more robust to these kinds of distortions. Figure 3.8 compares a segment of clean speech and its distorted version. The left column shows the clean speech, speech spectrum, LP residual signal and spectrum of the residual signal (from top to bottom), respectively. The right column gives the distorted counterparts. It is clear that the formant structure changes. Since the vocal tract features (MFCC or LPCC) mainly represent the formant structure, they are also changed and a false rejection error may happens.

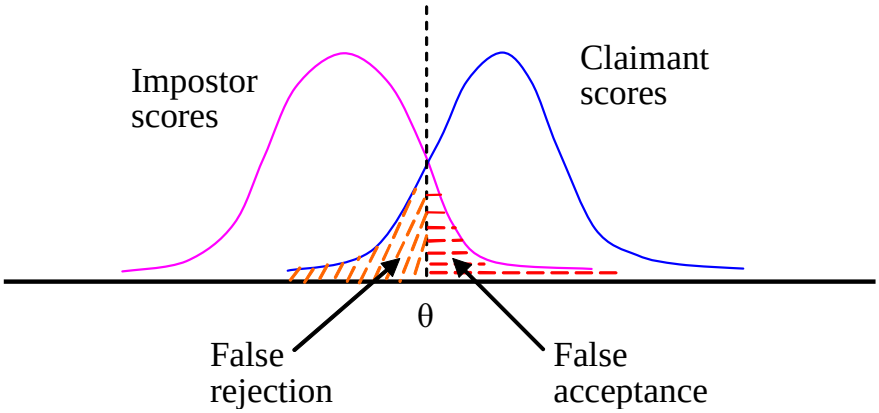
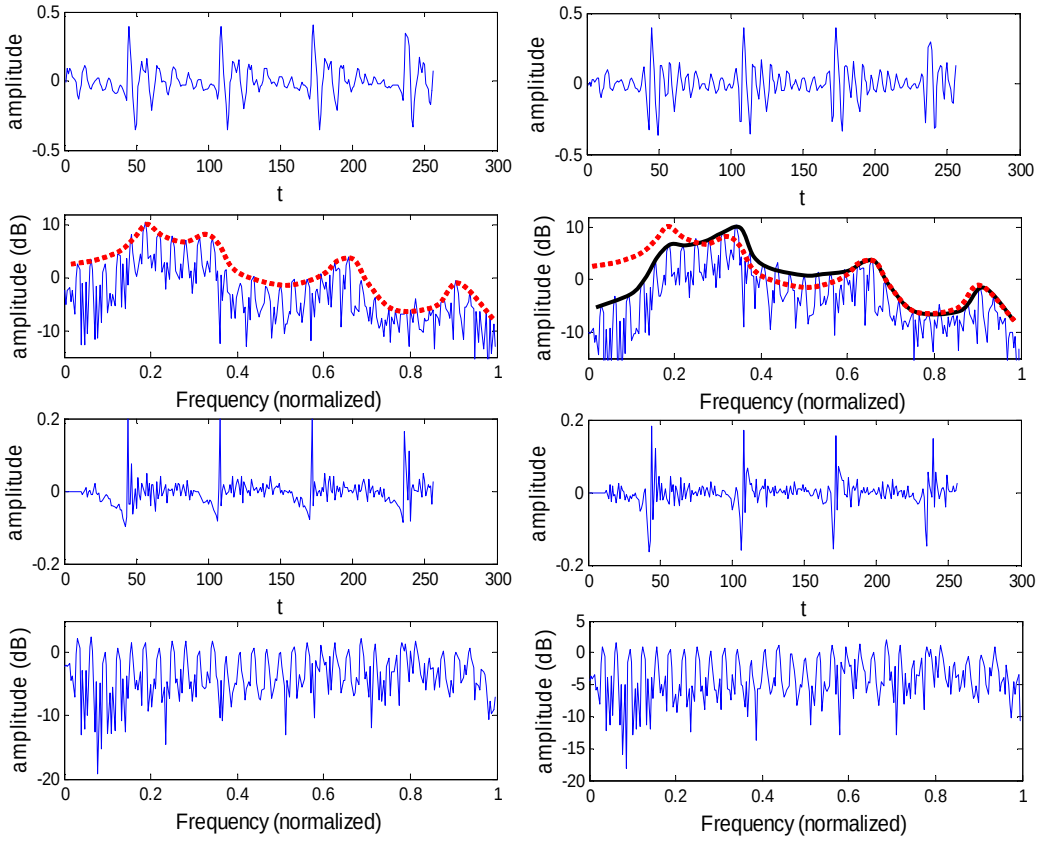


Figure 3.7: The score distributions of a speaker verification system

However, comparing the two LP residual signals, most of the source information, e.g., the positions and the amplitudes of the positive pitch epochs, the degree of noise, has been retained. The inverse filtering essentially is a whitening process, which results in a nearly flat spectrum of the residual signal, as illustrated in the figure. That is, the channel distortions are mainly reflected in the LP coefficients, not in the residual signal. Therefore, incorporating vocal source information could improve the robustness of the speaker recognition system.



a. clean speech b. distorted speech

Figure 3.8: Comparison of the vocal tract and vocal source characteristics of clean and distorted speech. From top to bottom are speech, speech spectrum, LP residual signal, and residual signal spectrum, respectively.

Chapter 4

Time-Frequency Feature Extraction from the LP Residual Signal

Basically, a glottal pulse waveform representing the true vocal cords vibration process is preferred for analyzing the speaker-specific source excitation characteristics. However, estimation of the glottal pulse waveform requires an accurate glottal closing interval localization, which is almost impracticable in real time applications such as telephony speaker recognition. In this thesis, rather than the glottal pulse waveform, the LP residual signal is adopted as a good, though not exact, representative of the vocal source excitation. Figure 4.1 compares the glottal flow derivative and the LP residual signal of two segments of speech waveform from two male speakers. It is clear that two signals are very similar in that: (1) the positions of pitch epochs and (2) the existing of secondary pulses in speaker B.

The LP residual signal of voiced speech is a quasi-periodic signal with pitch epoch at around each glottis closing instant (sometime, secondary and even tertiary epochs within one pitch period). These pitch epochs are the dominant elements in characterizing the underlying vocal cords vibration mechanism, requiring a very short analysis window to capture the temporal properties. On the other hand, a relative longer window is also necessary for capturing the pitch

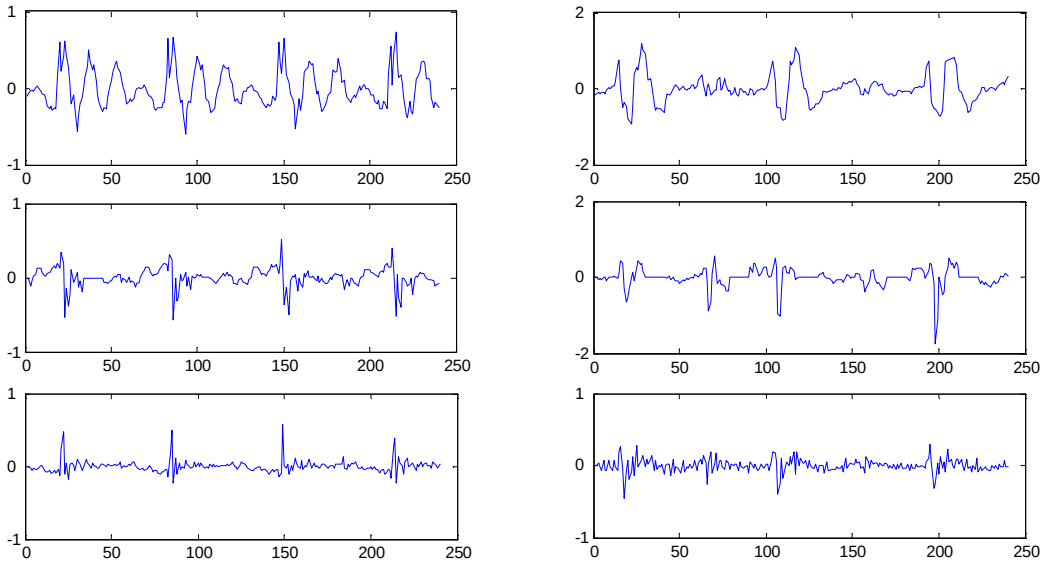


Figure 4.1: Comparison of the glottal flow derivative and the residual signal Left: speaker A, Right: speaker B. Top to bottom: 30ms segment of speech waveform, glottal flow derivative for which the LP coefficients are estimated during the glottal closing interval by covariance method, and the LP residual signal for which the coefficients are estimated with the 30 ms speech segment by autocorrelation method.

related low frequency properties. Therefore, the speaker-specific information of the residual signal resides in both time and frequency domain, requiring a time-frequency analysis rather than Fourier analysis for feature extraction.

Figure 4.2 shows the short-time Fourier spectrum of speech and the corresponding LP residual signals. As illustrated, speech signal is quasi-stationary in a short-time segment of 30 ms. The Fourier spectra can well capture the formant structures. However, the short-time Fourier spectra of the LP residual signals are nearly flat, from which temporal details of the pitch epochs is lost. Therefore, short-time Fourier transform is not a good candidate for feature extraction on the LP residual signal.

In this thesis, we have conducted an extensive investigation on extracting efficient speaker discriminative source features to enhance the performance of speaker recognition. Two time-frequency transforms, i.e. Haar transform and wavelet transform, have been applied for analyzing the LP residual signal and their effectiveness in feature extraction have been studied. The feature extraction techniques described in this chapter have been published in [121][122]. In

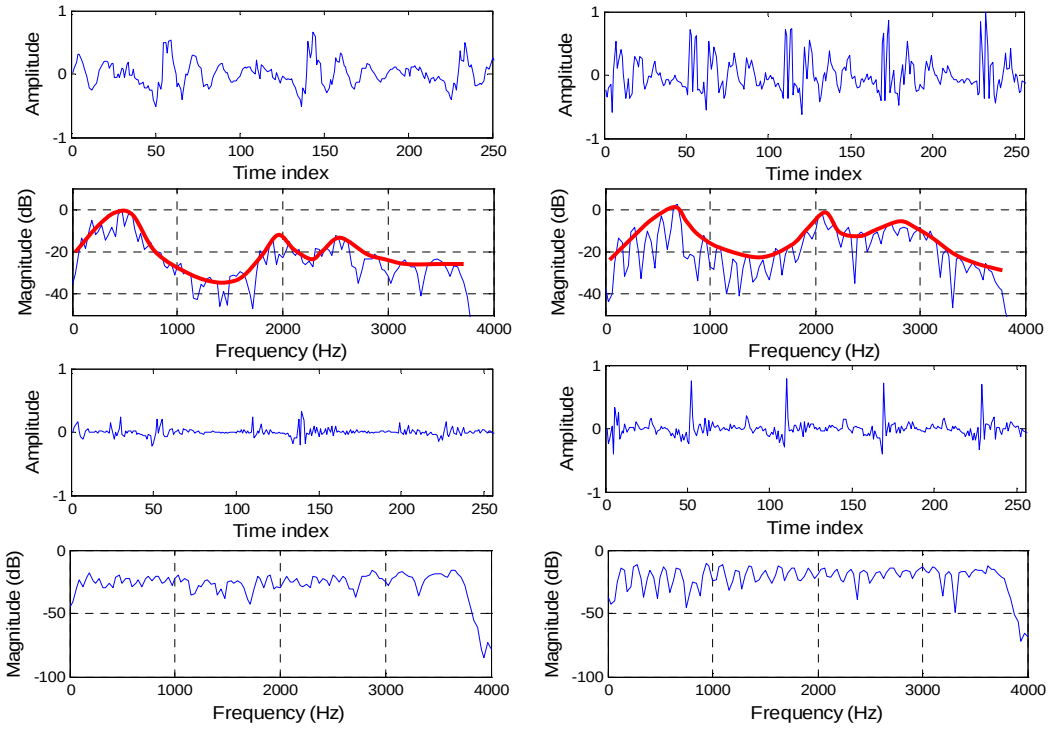


Figure 4.2: Short-time Fourier spectrum of speech signal and LP residue.

Left column: speaker A, Right column: speaker B. Top to bottom: 30ms segment of speech waveform, Fourier spectrum, LP residual signal and its spectrum, respectively.

the rest of this chapter, we first analyze the speaker specific information embedded in the LP residual signal in Section 4.1. Then the time-frequency feature extraction with Haar transform and wavelet transform, and the two new features, namely, Haar Octave Coefficients of Residues (HOCOR) and Wavelet Octave Coefficients of Residues (WOCOR), will be described in Section 4.2 and 4.3, respectively.

4.1 Speaker Specific Information in the LP Residual Signal

In the theoretical speech production model as described in Figure 3.2, if the combined effect of $G(z)$, $V(z)$ and $R(z)$ can really be represented by an all-pole model $H(z)$, the residual signal $e(n)$ is a series of impulse (similar to that of

the Speaker A in Figure 4.1). However, the real configuration of vocal system frequently violates the all-pole assumption. For example, the glottal excitation in real speech generally consists of both pulses and inter-pulses noise, sometime more than one pulses within one pitch cycle (as that of the Speaker B in Figure 4.1). Moreover, the glottal model $G(z)$ and nasal coupling generally introduce zeros and can not be exactly represented by the all-pole model $H(z)$. Therefore, the LP residual signal essentially carries not only the glottal excitation related information, but also all the other information that has not been captured by the LP coefficients.

4.1.1 Glottal excitation

Voicing/Unvoicing

Generally, when tackling the glottal excitation, we only focus on the voiced sounds since theoretically the excitation source of unvoiced sounds is random noise. Our recognition results show that, without voicing detection, the incorporation of unvoiced sounds degrades the performance of source related features for speaker recognition.

Pitch and harmonics information

Pitch information has been considered as one of the most important factor characterizing the glottal excitation. In the LP residual signal, pitch information is manifested as pitch epoches appearing in every glottal closing instant. The harmonics information is closely related to the regularity of the pitch epoches and the degree of aspiration noise.

glottal, subglottal pulses and aspiration noise

As described in Chapter 3, the glottal pulse waveform, the existence of multi-pulses, and the intensity of aspiration noise carries speaker specific information. These features are approximately, if not exactly, retained in the LP residual signal. Appropriate feature extraction to represent these characteristic could

be useful for speaker recognition.

Phase information

Even though human eras are not sensitive to phase mismatch between speeches, it has been found that phase mismatch did cause spectral/cepstral distortion to a certain extern[53]. As to the LP residual signal, phase information can be implicitly identified by epoch localization. In this thesis, it will be demonstrated that the accurate pitch estimation and epoch localization will improve the speaker recognition performance of the feature extracted from the LP residual signal.

4.1.2 Lip radiation effect

The air pressure relation at the lips can be modeled by high-pass filtering operation, as formulated in Eqt. 3.4. At low frequencies, this can be approximated as a differentiator. The speech pressure measured in front of the lips can be expressed as [83]

$$s(t) \approx d[u(t) * h(t)]/dt = [du(t)/dt] * h(t) \tag{4.1}$$

The effect of radiation is typically retained in the source signal, since it corresponds to a one-zero model. Therefore, the LP residual signal is not the real glottal waveform, but its derivative.

4.1.3 Zeros due to the nasal sounds

The all-pole model represents only the the poles related formant structure. Its inverse model can not fully compensate the zeros introduced by glottal excitation and nasal coupling. Thus, the zeros will be reflected in the LP residual signal, resulting in delayed and scaled versions of the glottal epochs. In addition, the existing zeros may also affect the accuracy of vocal tract frequency response estimation, which also introduces additional delayed and scaled versions of the original glottal epoch [3].

4.1.4 Effects of source-tract interaction

In the classical acoustic theory of speech production, it is assumed that the glottal excitation and the vocal tract modulation are linear separable. This assumption does not exactly reflect the real situation. In fact, there is always source-tract interaction during articulation. The most significant interaction effects on the glottal waveform are the ripples, the sub-glottal pulses, the pulse skew and the resultant variance in the glottal spectral slope [36] [4].

4.2 Generating the Haar Transformed Vocal Source Feature HOCOR

4.2.1 Haar transform

Haar transform of signal $x(n)$ and its inverse transform can be formulated as [11]

$$\begin{cases} X(k) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)H(k,n), k = 0, 1, \cdots, N-1 \\ x(n) = \sum_{k=0}^{N-1} X(k)H(k,n), n = 0, 1, \cdots N-1 \end{cases} \tag{4.2}$$

The Haar function $H(k,n)$ is a completely orthogonal function set of rectangular waveforms

$$\begin{aligned} H(0,n) &= 1, 0 \le n \le N-1 \\ H(k,n) &= H(2^{i-1}+j-1, n) \\ &= \begin{cases} \sqrt{2^{i-1}}, \frac{j-1}{2^{i-1}}N \le n < \frac{j-\frac{1}{2}}{2^{i-1}}N \\ -\sqrt{2^{i-1}}, \frac{j-1}{2^{i-1}}N \le n < \frac{j-\frac{1}{2}}{2^{i-1}}N \\ 0, elsewhere \end{cases} \\ & \quad i = 1, 2, \cdots .j = 1, 2, \cdots, 2^{i-1} \end{aligned} \tag{4.3}$$

where i denotes an *octave* subset having a zero-crossing in a given width $N/2^{i-1}$, and j gives the *position* of the function within this subset. Figure 4.3 shows the first 4 octave groups of Haar Function.

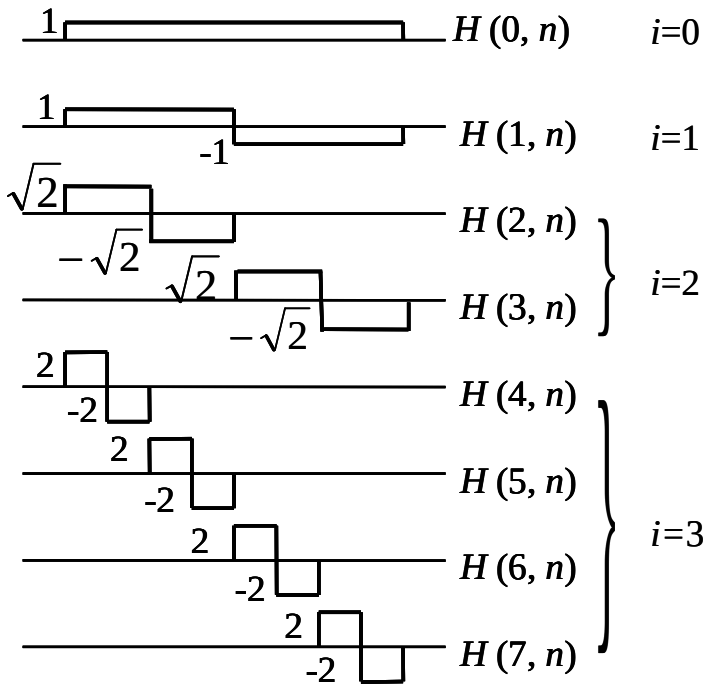


Figure 4.3: The first 4 octave groups of Haar Function

If we define the *frequency* to be the number of zero-crossing within a specific time interval, the Haar transform provides a kind of time-frequency analysis of the signal. We can define the Haar spectrum as

$$G(k) = |X(k)|\,, \qquad k = 0, 1, \cdots, N-1 \tag{4.4}$$

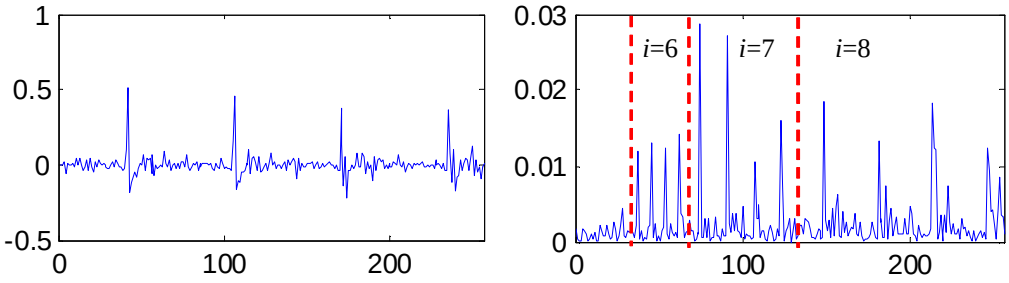
The Haar spectrum can also be partitioned into different octave groups, i.e.

$$\begin{aligned} &\tilde{\mathrm{H}}_0 = \{G(0)\} \\ &\tilde{\mathrm{H}}_\mathrm{i} = \{G(k) \mid k = 2^{i-1}, \cdots, 2^i - 1\} \\ &i = 1, 2, \cdots, \log_2 N \end{aligned} \tag{4.5}$$

All the $G(k)$'s belonging to an octave group can be considered as the result of scanning the signal with a specific Haar function and therefore contain the corresponding frequency information and its time changing property as well. Thus, Haar spectrum retains the spectro-temporal characteristics of the signal.

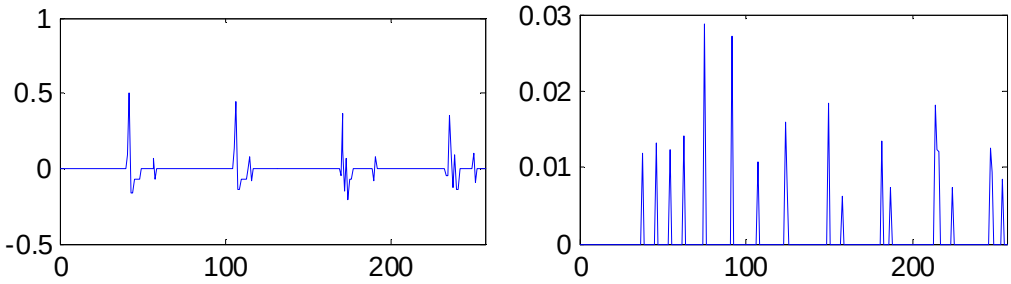
A very important property of Haar transform comes from its efficiency in detecting bursts among the noisy signal [11][106]. This is particular beneficial for analyzing the LP residual signal, which is a noise-like signal with quasi-periodic

pitch epochs. Figure 4.4 shows a segment of LP residual signal and its Haar spectrum. The peaks of $G(k)$'s within an octave group are position sensitive to the pitch epochs in the residual signal. As known, the most useful information of the LP residual signal is embedded in the pitch epochs. The efficiency in burst localization of Haar spectrum makes it possible to approximately recover the signal with relatively small number of Haar coefficients. For example, one can clip the Haar coefficients and retain only the most prominent coefficients representing the pitch epochs. The reconstructed signal from the clipped Haar coefficients can well represent the original signal in the pitch-related information (i.e. the pitch period and pitch epochs), as illustrated in Figure 4.5. As mentioned in the last section, the degree of inter-epochs aspiration noise is kind of speaker-specific information, discarding the small valued coefficients might lose this information. However, in noisy conditions, the pitch epochs are much more resistant to the noise distortion. Therefore, robust feature generation from the Haar spectrum should focus more on these prominent coefficients.



a. LP residual signal $u(n)$ b. Haar spectrum $G(k)$

Figure 4.4: Haar spectrum of a length-256 LP residual signal



a. reconstruct residual signal b. clipped Haar spectrum

Figure 4.5: Clipped Haar spectrum and the reconstructed signal

Table 4.1: Octave copying of Haar coefficients

$X(k)$	$X_1(k)$	$X_2(k)$
$X(0)$	$X(0)$	$X(0)$
$X(1)$	$X(1)$	$X(1)$
$X(2)$	$X(2)$	$X(2)$
$X(3)$	$X(3)$	$X(3)$
$X(4)$	$X(4)$	$\sqrt{2}X(2)$
$X(5)$	$X(5)$	0
$X(6)$	$X(6)$	$\sqrt{2}X(3)$
$X(7)$	$X(7)$	0
$X(8)$	$\sqrt{2}X(4)$	$2X(2)$
$X(9)$	0	0
$X(10)$	$\sqrt{2}X(5)$	0
$X(11)$	0	0
$X(12)$	$\sqrt{2}X(6)$	$2X(3)$
$X(13)$	0	0
$X(14)$	$\sqrt{2}X(7)$	0
$X(15)$	0	0

On the other hand, to the noise-like residual signal with flat Fourier spectrum, there are lot of redundance between the different octave coefficients, especially for the higher octaves which contain mainly the noise components. Therefore, one can approximate the higher octave coefficients from the lower octave coefficients. And the reconstructed signal should approximate the original residual signal, given the new coefficients have the similar octave power spectrum as the original one. One way to reduced the redundance is to replace the higher octave coefficients with the lower octave ones by *octave copying*. To a 16 point Haar transform, for example, Haar coefficients generated by the 1st- and 2nd- order octave copying, $X_1(k)$ and $X_2(k)$, are illustrated as in Table 4.1. Figure 4.6 shows the LP residual signal and its Haar spectrum, and the

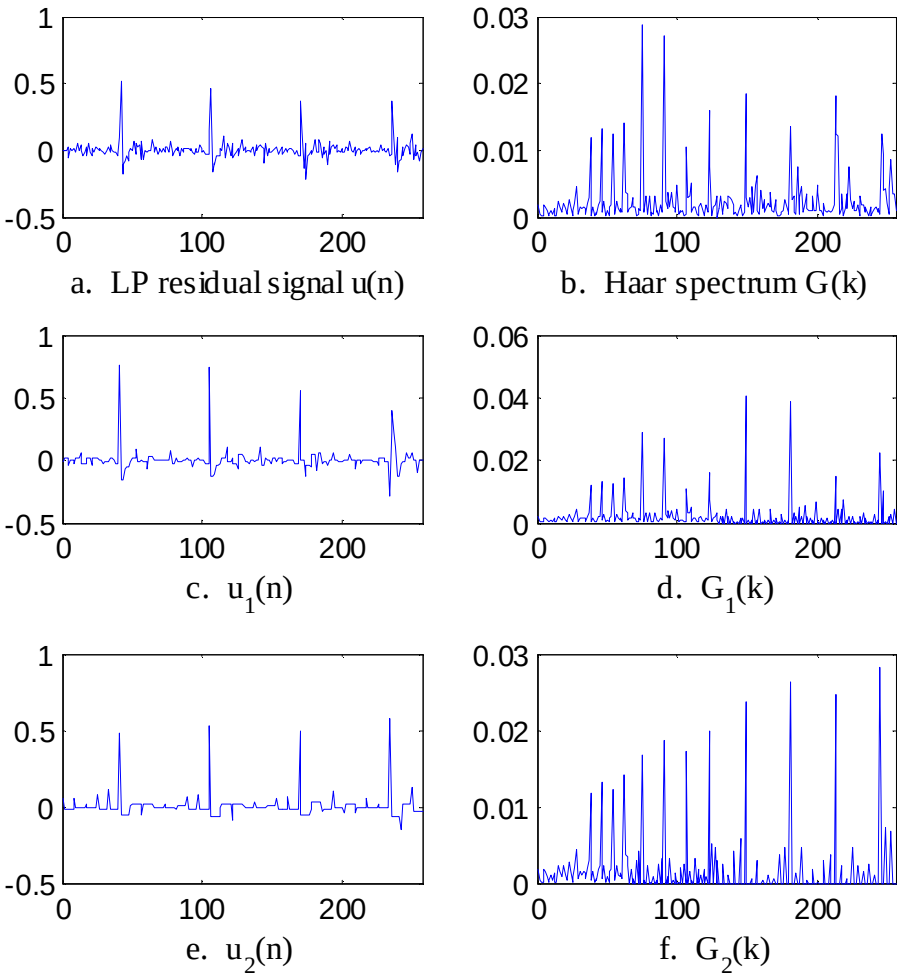


Figure 4.6: Reconstructing signal from Haar coefficients with 1st- and 2nd-order octave copying

1st- and 2nd- order octave copied Haar spectrum $G_1(k)$ and $G_2(k)$, and their reconstructed signals. As illustrated, after octave copying, the reconstructed signal approximately retains the pitch epochs related properties, although some noise details are lost. Therefore, the higher octave coefficients contains lost of noise information and they seems to be less important in characterizing the pitch epochs. This property can be utilized in feature selection, in particular, in the noisy speech.

4.2.2 Feature generation

In this thesis, we generate the HOCOR parameter from every 32 ms segment of LP residual signal (for 8000 Hz sampling data, the window length is 256). The Haar spectrum can be divided into 9 octave groups ($i = 0, 1, \cdots, 8$). Each octave contains the spectral information as follows,

$$\begin{aligned} \tilde{\mathrm{H}}_0 &\rightarrow D.C. \\ \tilde{\mathrm{H}}_1 &\rightarrow f \in [0,\ 32]\mathrm{Hz} \\ \tilde{\mathrm{H}}_2 &\rightarrow f \in [32,\ 64]\mathrm{Hz} \\ &\vdots \\ \tilde{\mathrm{H}}_8 &\rightarrow f \in [2000,\ 4000]\mathrm{Hz} \end{aligned} \tag{4.6}$$

The lowest frequency component we are interested in LP residual signal is the fundamental frequency, which is seldom less than 64 Hz. Therefore the first three octaves can be ignored. To capture the spectral information of each octave sub-bank, the simplest HOCOR feature set can be derived as,

$$\mathrm{HOCOR}_0 = \left\{ \underset{G(k)\in\tilde{H}_i}{\|G(k)\|} \mid i = 3, 4 \cdots, \log_2 N \right\} \tag{4.7}$$

where $\|\cdot\|$ denotes the 2-norm operator. In this case, the feature vector has just 6 elements containing only pitch and harmonics related spectral information, but not the temporal information since all $G(k)'s$ within an octave are summed up together. To retain the temporal information, each octave can be equally divided into 2 sub-groups and then the energy of each sub-group is computed to generate a double sized HOCOR. For convenience, we call it the first-ordered HOCOR, noted as HOCOR_1. There are now 12 elements in the HOCOR_1 feature vector and contains approximate temporal information of the constituent frequency components. To extend further so as to obtain more detailed temporal information, each octave can be divided into 4, 8 and up to 2^{i-1} (where i is the octave number as in Eqt. 4.4) sub-groups, noted as

$$\mathrm{H}_i^\alpha = \{\mathrm{H}_i^\alpha(j) \mid j = 0, 1, \cdots, 2^\alpha - 1\} \quad , \quad (\alpha \leq i - 1) \tag{4.8}$$

where

$$\mathrm{H}_i^{\alpha}(j) = \left\{ G(k) \;\middle|\; \begin{array}{l} k = 2^{i-1} + j \cdot 2^{i-1-\alpha}, \cdots, \\ 2^{i-1} + (j+1) \cdot 2^{i-1-\alpha} - 1 \end{array} \right\} \tag{4.9}$$

And the αth-ordered HOCOR is given by

$$\mathrm{HOCOR}_{\alpha} = \left\{ \underset{G(k) \in H_i^{\hat{\alpha}}(j)}{\|G(k)\|} \;\middle|\; \begin{array}{l} i = 3, \cdots, \log_2 N \\ \hat{\alpha} = \min(i-1, \alpha) \\ j = 0, \cdots, 2^{\hat{\alpha}} - 1 \end{array} \right\} \tag{4.10}$$

The process for generating the HOCOR_{α} feature parameters is illustrated in Figure 4.7, where the LP coefficients a_k are estimated by 12th-ordered LP analysis on the Hamming windowed speech frame using the autocorrelation methods [83]. The LP residual signal $u(n)$ is generated by inverse filtering, i.e.

$$u(n) = s(n) - \sum_{k=1}^{12} a_k s(n-k) \tag{4.11}$$

Before Haar transform, $u(n)$ is normalized to the range of [-1, 1] to reduce the intra-speaker variation.

In summary, HOCOR bears the following properties:

- HOCOR is uncorrelated to the LP coefficients derived vocal tract feature parameters in a large extent, since the residual signal is theoretically orthogonal to LP coefficients;

- The rectangular base function and the time-frequency properties of Haar transform result in better spectral decomposition of the noise like, burst mode changing residual signal. The Haar coefficients are very efficient in capturing the time-frequency properties of the pitch epochs within the residual signal;

- HOCOR_{α} with $\alpha > 0$ represents pitch and harmonics related spectral information as well as temporal characteristics within the segment of LP residual signal, which will be verified to be useful for speaker recognition;

- Computational simplicity, which will be extremely useful in real-time applications where computational simplicity is a critical requirement.

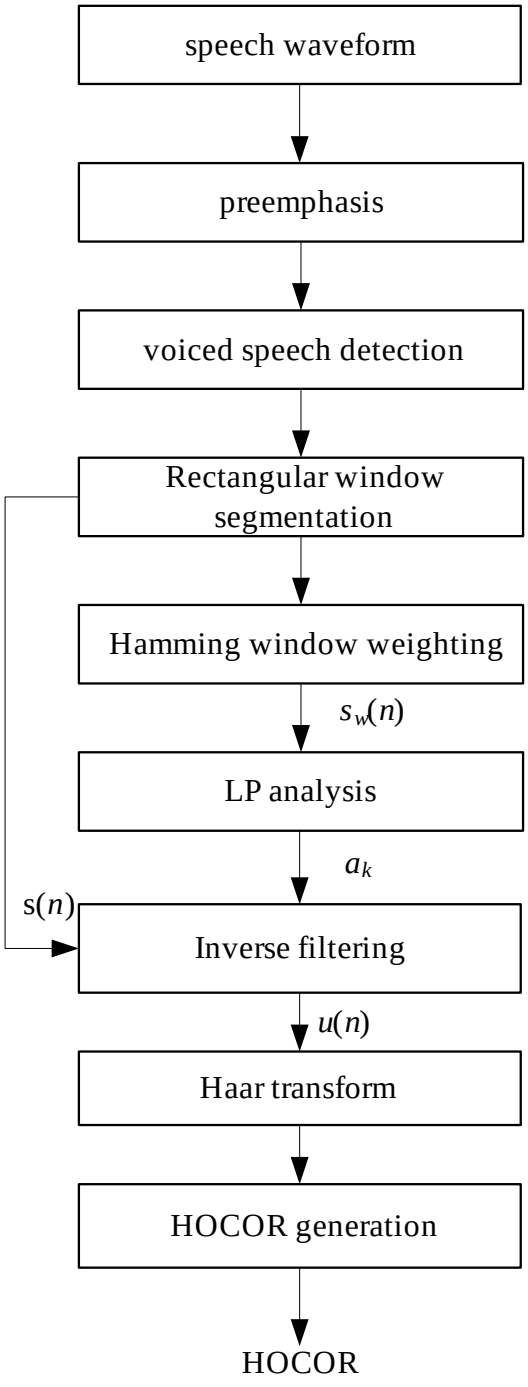


Figure 4.7: The process for generating time-frequency feature HOCOR

4.3 Generating the Wavelet Transformed Vocal Source Feature WOCOR

In the last section, we derived the HOCOR feature parameters by applying Haar transform on every 32 ms of LP residual signal which may cover several pitch periods. Study of the glottal waveform showed that there exist a certain degree of dynamic evolution (e.g. the variation of the periods and the amplitudes of the glottal pulses) from one pitch period to the next [80]. The dynamic evolution of glottal waveform is highly related to the speaker-specific phonation styles, and thus should be useful for speaker recognition. The dynamic evolution of glottal pulse has a close relation with the variation of the period (jitter) and the amplitude (shimmer) of the epoches in the residual signal. What's more, a great pitch variation could result in a strong degree of inter-pulses noise in the LP residual signal. To study such inter-period variations, this section adopts a pitch-synchronous analysis for the LP residual signal. That is, we which restricts the length analysis window to be exactly two pitch cycles and be synchronized with the pitch epochs.

On the other hand, Haar transform has been shown to be very effective in bursts detection from signals [106]. The rectangular basis function is very good for time localization, but not for frequency localization. Figure 4.8 illustrates the spectrum of Haar basis function and that of db4 wavelet function. it is clear that the Haar function is not good at frequency localization in that (1) large bandwidth so as worse frequency resolution; and (2) less side lobe attenuation so as strong interference from neighboring bands. Therefore, it may not be the best candidate for time-frequency analysis of the LP residual signal.

In this section, we proposed a time-frequency vocal source feature extraction by pitch-synchronous wavelet transform, with which the pitch-related low frequency properties and the high frequency information associated with pitch epochs, as well as their temporal variations within a pitch period and over consecutive periods can be effectively characterized.

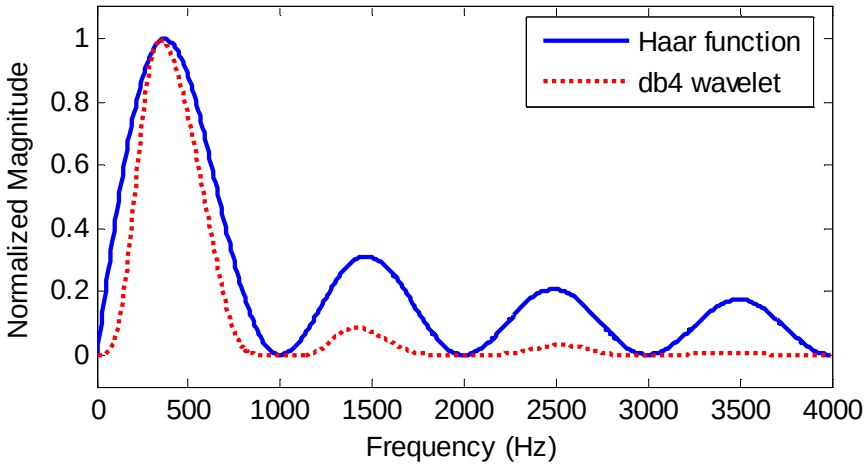


Figure 4.8: Comparison of spectra of Haar function and db4 wavelet

4.3.1 Wavelet transform

The wavelet transform of time signal $x(t)$ is

$$w(a,\tau) = \frac{1}{\sqrt{|a|}} \int_t x(t)\Psi^*(\frac{t-\tau}{a}) \tag{4.12}$$

where $\Psi(t)$, a and τ are the mother wavelet function, scaling (or dilation) parameter and translation parameter, respectively. $\Psi(\frac{t-\tau}{a})/\sqrt{|a|}$ is named the baby wavelets. It is constructed from the mother wavelet by first, scaling $\Psi(t)$ which means to compress or dilate $\Psi(t)$ by a, then moving the scaled wavelet to the time position of τ. The compression or dilation of $\Psi(t)$ will change the window length of wavelet function, thus changing the frequency resolution. Therefore, the ensemble of $\Psi(\frac{t-\tau}{a})/\sqrt{|a|}$ constitutes the time-frequency building blocks of the wavelet transform [22] [103] .

Eqt. 4.12 can be discretized to be

$$w(a,b) = \frac{1}{\sqrt{|a|}} \sum_n x(n)\Psi^*(\frac{n-b}{a}) \tag{4.13}$$

where a and b are now discrete parameters. The Haar transform can be regarded as a wavelet transform with haar function $H(1,n)$ as the mother wavelet, $H(k,n)$ the baby wavelets, the scaling parameter $a = 2^{-k}$, and the translation parameter $b = 2^{-k}$, $k = 1, 2, \cdots$.

The advantage of wavelet transform for LP residual signal first comes from that it performs a const Q analysis of the signal. That is, the changing of frequency resolution necessarily accompanies a central frequency shift. This property could be very useful in speech analysis since our hearing system also performs a const Q analysis of the speech signal.

Another advantage of wavelet transform is that the basis functions in Eqt. 4.12 have not been specified. Therefore, besides the Haar functions, one can select various other basis functions, or even construct a new one with desired time-frequency resolution, according to the signal to be analyzed.

The wavelet transform essentially performs a multi-resolution analysis of the signal since its basis functions have various time-frequency resolution. To achieve the multi-resolution objective, the signal is first analyzed by wavelet function with small a which corresponds to high time, low frequency resolution and the temporal detail are measured. Then, a is increased and the frequency details are obtained while losing some time information. Such a kind of multi-resolution analysis is reasonable since generally the low frequency components are time insensitive while the high frequency components are frequency insensitive. Figure 4.9 shows a segment of LP residual signal (top panel) and its 4 wavelet transforms in different scales (the bottom 4 panels). As shown, as a increases, time resolution decreases, whilst the frequency resolution improves. Also, the time varying characteristics of $x(n)$ can be measured from $w(a, b)$ at different translation parameters.

In order to generate feature parameters from the wavelet coefficients, the scaling parameters for multi-level wavelet transform should be first determined. The selection of a should take into account the time-frequency properties of both the wavelet function and the analyzed LP residual signal. Figure 4.10 shows the waveforms of $\Psi(n/a)$ with $a = 2^k, k = 1, 2, 7$ and their Fourier spectra. As illustrated, the central frequency (f_c) and bandwidth (B) decrease as a increase (the frequency resolution increase). The wavelet function with $a = 128$ ($f_c \simeq 45, B \simeq 40$) seems to be of no use in analyzing the LP residual signal, since the lowest frequency component we are interested is the fundamental frequency,

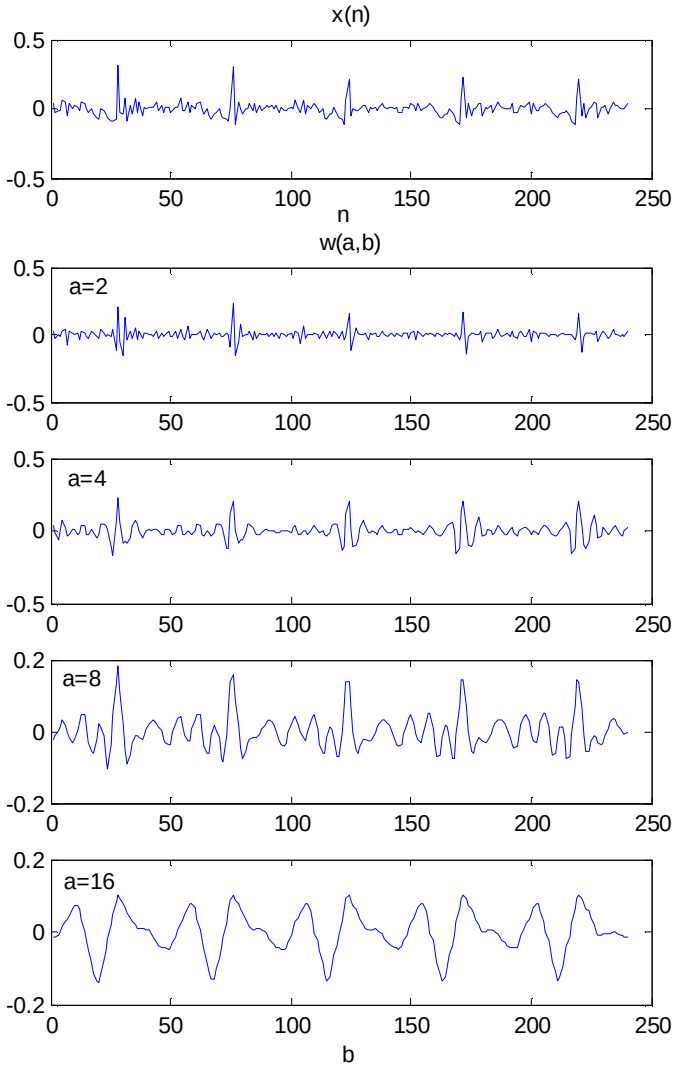


Figure 4.9: Wavelet transform of a segment of LP residual signal

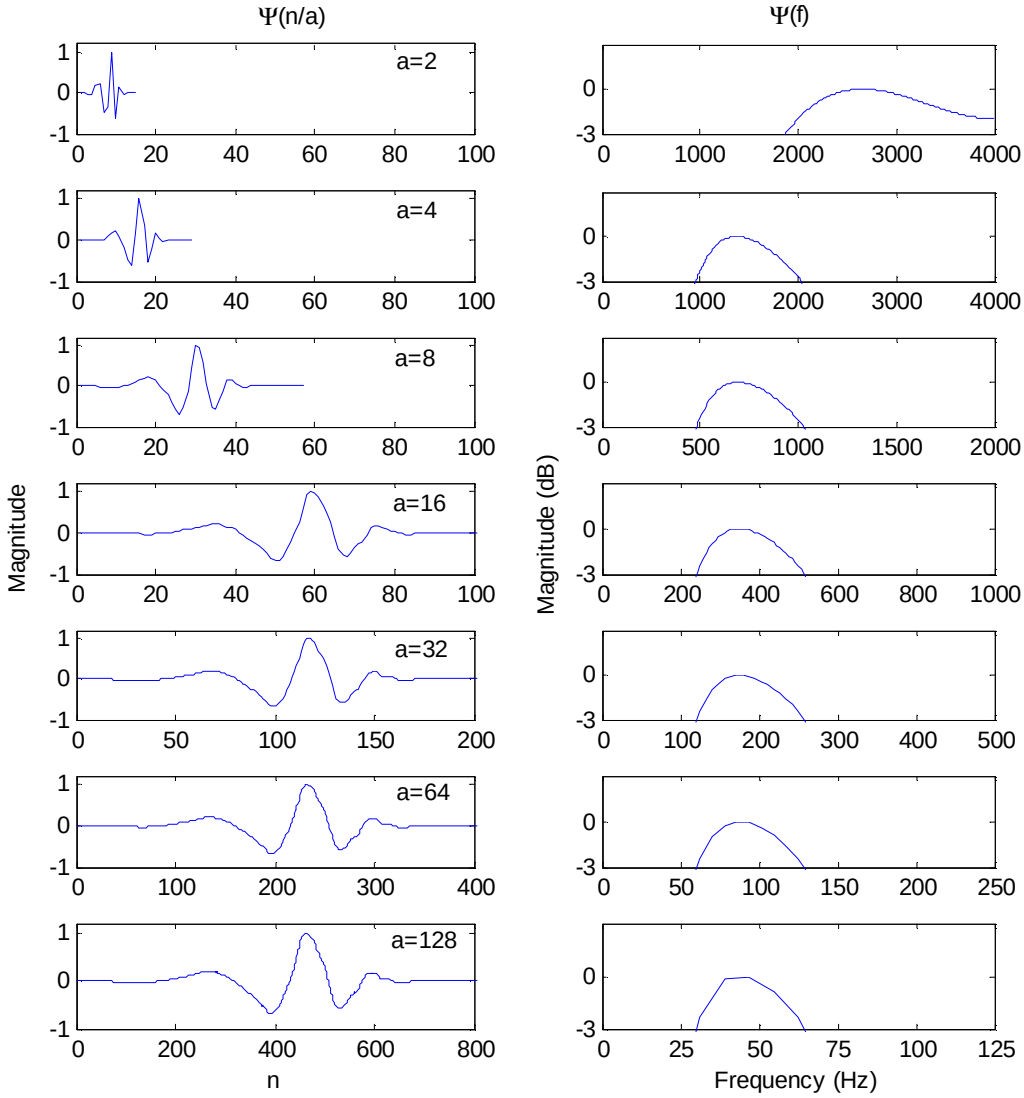


Figure 4.10: Wavelet functions and their spectra

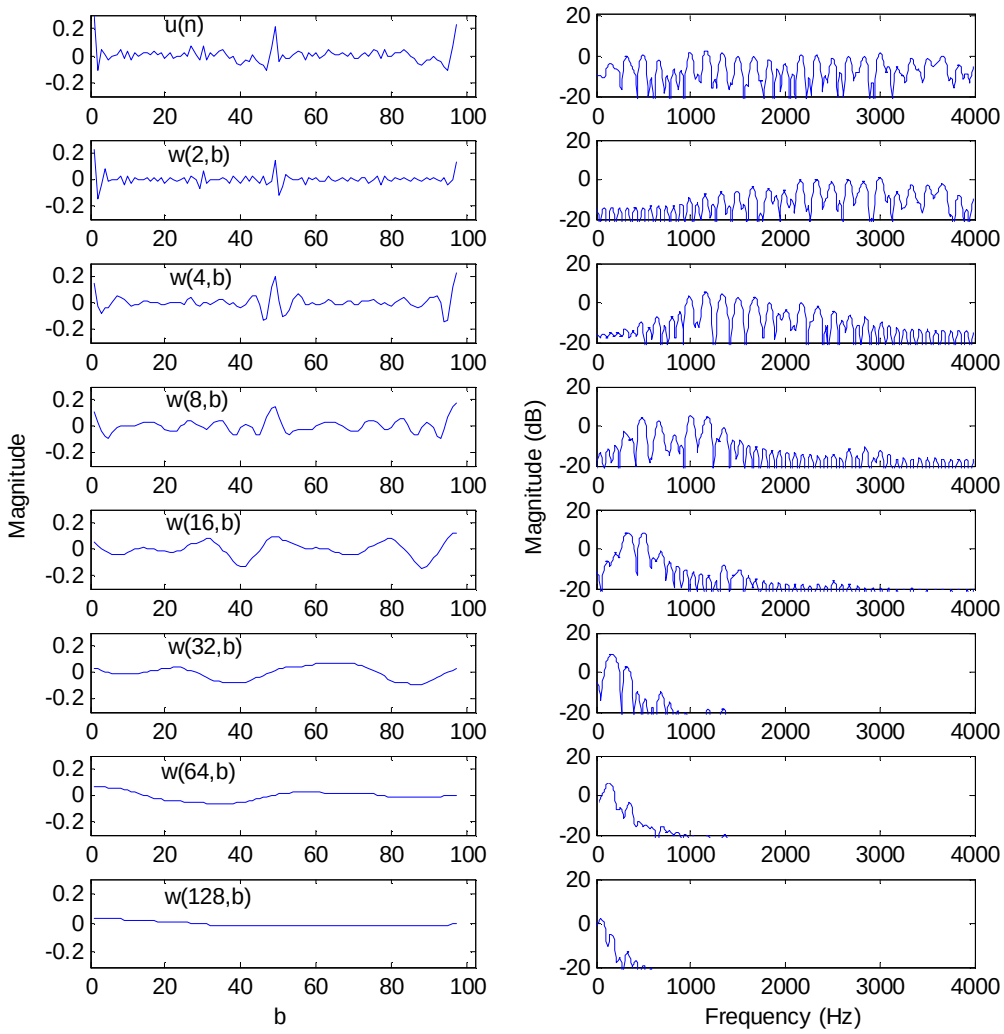


Figure 4.11: Wavelet transform of LP residual signal and their spectra

which is generally grater that 60 Hz. Figure 4.11 shows a segment of LP residual signal (two pitch periods) and its wavelet transform coefficients with the first 7 octave scaling parameters and their Fourier spectra. It is clear that the last one contains little information. Therefore, in this thesis, 6 octave scaling parameters, i.e.

$$a = \{2^k | k = 1, 2 \cdots, 6\} \tag{4.14}$$

are selected for multi-resolution analysis.

4.3.2 Feature generation

The feature extraction process is illustrated in Figure 4.12 and is formulated in the following steps:

1) Pre-emphasis and Hamming windowing

The speech signal is first pre-emphasized and segmented into frames of 30 ms for subsequent processing.

2) Voicing decision and pitch detection

We use the Entropic's Robust Algorithm for Pitch Tracking (RAPT) [104] for pitch detection and voicing decision. In generating the time-frequency feature from the LP residual signal, only voiced segments will be processed since it is believed that unvoiced speech carries little speaker-specific source excitation information.

3) LP analysis and inverse filtering

The LP analysis and inverse filtering for generating the LP residual signal is the same as that in HOCOR generation. The magnitude of $u(n)$ is also normalized to the range of [-1, 1] to reduce the intra-speaker variation.

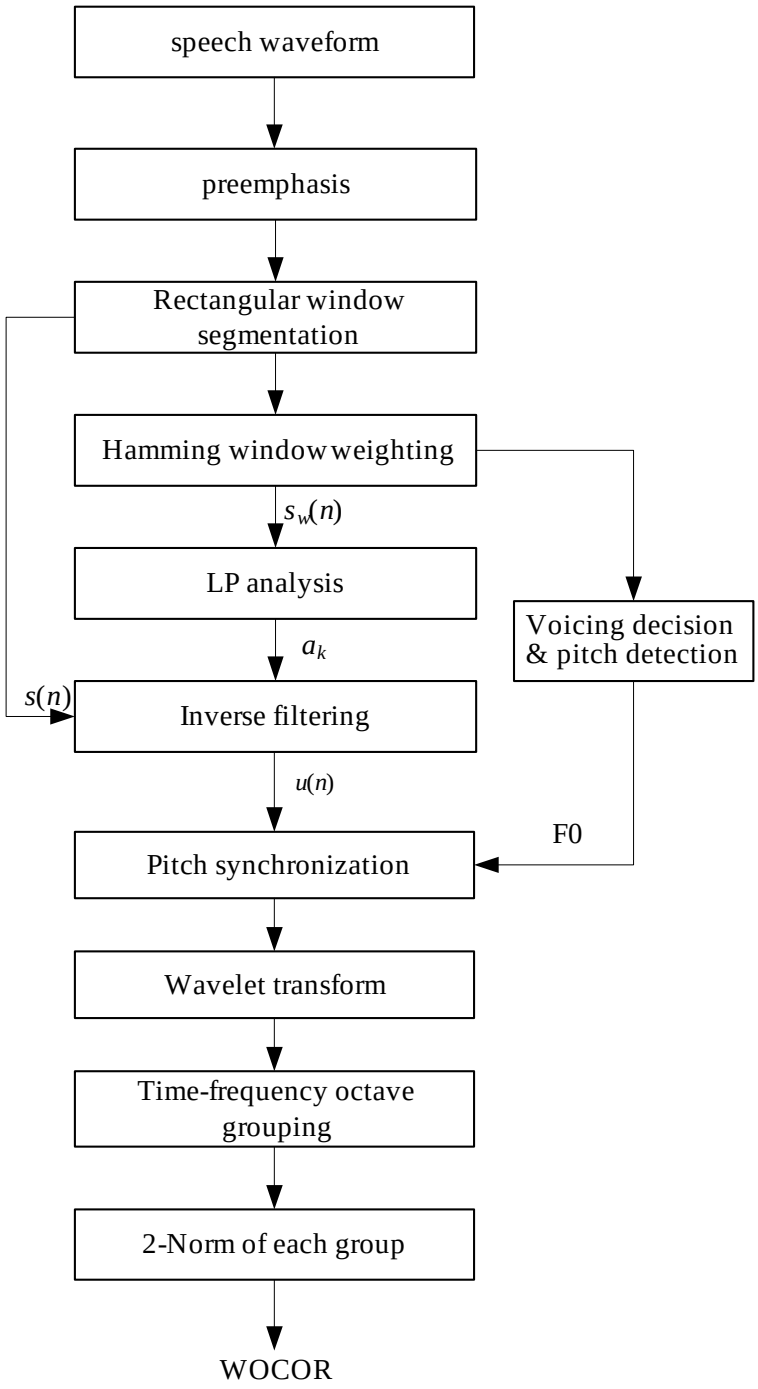


Figure 4.12: The process for generating time-frequency feature WOCOR

4) Pitch-synchronous wavelet transform of the LP residual signal

With the pitch periods estimated in step 2, pitch epochs synchronized analysis windows are located for subsequent processing by wavelet transform. Each window covers two pitch periods and overlaps with the previous window by one pitch period. Wavelet transform of the pitch-synchronized LP residual signal can be expressed as

$$w(a,b) = \frac{1}{\sqrt{|a|}} \sum_n u(n) \Psi^*(\frac{n-b}{a}) \tag{4.15}$$

where $\Psi(n)$ is the 4th-order Daubechies wavelet.

5) Generation of time-frequency features

To generate the feature parameters for pattern recognition, the wavelet coefficients with a specific scaling parameter are grouped as

$$W_k = \left\{ w(2^k, b) \,|b = 1, 2, \cdots, N \right\} \tag{4.16}$$

where N is the window length. Each W_k is called an octave group. Then the WOCOR parameters can be derived as

$$\text{WOCOR}_1 = \{ \|W_k\| \,|k = 1, 2, \cdots, 6 \} \tag{4.17}$$

WOCOR_1 has 6 elements and contains only spectral information. It includes no temporal characteristics within the analysis window. To retain the temporal details, each octave group can be equally divided into M sub-groups

$$\begin{aligned} &W_k = \left\{ w(2^k, b) \,|b \in (m-1:m] \times Round(N/M) \right\} \\ &m = 1, 2, \cdots, M \end{aligned} \tag{4.18}$$

Finally, a feature vector with $6M$ parameters can be generated as

$$\text{WOCOR}_M = \left\{ \|W_k(m)\| \,\middle|\, \begin{array}{l} m = 1, 2, \cdots, M \\ k = 1, 2, \cdots, 6 \end{array} \right\} \tag{4.19}$$

With multi-level wavelet transform, the pitch-related low frequency properties and the high frequency information associated with pitch epochs can be captured with different time-frequency resolutions. Pitch synchronization

and dividing each octave group into several sub-groups enable the measuring of temporal variations of spectral components within a pitch period and that over consecutive periods. Therefore, WOCOR_M is capable of capturing the spectro-temporal characteristics of the LP residual signal.

4.4 Summary

We adopted the LP residual signal as a representative of the vocal source excitation. The LP residual signal contains not only rich speaker-specific time-frequency information of the glottal phonation, but also all the speech characteristics that cannot be represented by the LP coefficients. The two source features derived, i.e. HOCOR_α and WOCOR_M effectively capture the spectro-temporal characteristics of the LP residual signal. Particularly, WOCOR_α captures the temporal variations within each pitch period and that over consecutive periods. The HOCOR_α feature set, though does not explicitly represent such temporal variations as with WOCOR_M, it still contains the temporal information among the segment (32 ms) of analyzed LP residual signal. The efficiency in burst detection of Haar transform makes the HOCOR_α parameters more robust to noise distortions. Both feature sets are complementary to the vocal tract features derived from the LP coefficients, which benefits the fusion of source and tract features for speaker recognition.

Chapter 5

Speaker Recognition Using Time-Frequency Vocal Source Features

We have described in Chapter 4 the details how we generate the time-frequency feature parameters HOCOR_α and WOCOR_M from the LP residual signal. We also demonstrated that the new features can effectively capture the spectro-temporal characteristics of the residual signal, which is expected to be useful for speaker characterization and recognition. In this chapter, both speaker identification and verification systems are developed to evaluate the effectiveness of the new features. Section 5.1 describes the experimental procedure and the process for vocal tract feature extraction. Section 5.2 demonstrates how the increased temporal information in HOCOR_α and WOCOR_M can improve the speaker recognition performance, and the feature dimension is determined accordingly. The performances of the proposed vocal source and vocal tract features are compared in Section 5.3. A preliminary source-tract information fusion system will also be evaluated.

5.1 Experimental Procedure

5.1.1 Speech corpus

Speaker recognition performance in matched conditions is evaluated over a Cantonese database, which is a male subset of the database collected in the Chinese university of Hong Kong for speaker recognition purpose [125]. There are 50 registered male speakers in this corpus, each having 6 training sessions and 12 testing sessions. In this experiments, the first 3 training sessions will be used for training speaker models. The other 3 training sessions will be used for training the weighting parameters for the fusion of vocal source and vocal tract features. Each training and testing session contains respectively 30 and 6 utterance. For each speaker, the time span of data collection varies from 4 to 9 months with at least one week interval between two consecutive sessions. There are also 9 unregistered speakers totally providing 1230 testing utterances. These utterances will be used for unseen impostor testing.

Each utterance in this corpus contains a digit string consisting of 14 randomly generated digits. The speakers were prompted with the digit string and asked to read out continuously with short pause in preset positions, i.e., 8-21-447-7890-5536. The speech data were collected by the telephone handset and transmitted over a public fixed-line telephone network. Finally it was recollected via a dialogic card and saved in the computer disk with 8-bit mulaw binary format. There is no handset and channel mismatches between training and testing sessions since all the speech data were collected with the same handset and transmitted over the same fixed line telephone channel.

The collected data have been carefully validated and accurate annotation is provided. Therefore, it can be used for text-independent (but text-constrained) speaker recognition test since all the digit string are randomly generated, and text-dependent recognition test with the annotation provided.

5.1.2 The baseline system

To compare the performances of vocal source and vocal tract features, a baseline system is adopted, which employs the prevalent vocal tract feature parameters including the 12 dimensional Mel-frequency cepstral coefficients (MFCC), the log energy, and their first and second ordered dynamic coefficients. These feature parameters are concatenated together forming a 39 dimensional feature vector, noted as MFCC_E_D_A. The procedure for MFCC generation is illustrated in Figure 5.1. It is different from the conventional process as illustrated in Figure 3.6, where the signal spectrum is calculated by applying FFT on the Hamming windowed speech frame. The conventional MFCC parameters contain both vocal tract information characterized by the spectral envelop, as well as the vocal source information embedded in the fine structure of the fundamental and harmonic frequencies. It has been pointed out that when the pitch varies significantly between training and testing stages, the fine structure also changes significantly and degrades the performance of conventional MFCC parameters [126]. In this thesis, instead of applying the Fourier transform, the signal spectrum is calculated from the LP coefficients a_k, i.e.

$$H(\omega) = \frac{1}{\left|1 + \sum_{k=1}^{12} a_k e^{-j\omega k}\right|} \tag{5.1}$$

The advantage of calculating Fourier spectrum via Eqt. 5.1 is that the spectrum is smoothed and the effect of fundamental frequency and harmonics intensity is eliminated. Furthermore, this process makes the vocal tract parameters orthogonal to the vocal source parameters, since they are derived from two orthogonal components, the LP coefficients and the LP residual signal. Thus, the complementarity of the source and tract features is maximized.

5.1.3 Vocal source feature selection

The two vocal source features HOCOR_α and WOCOR_M as described in Chapter 4 will be evaluated. The feature dimension is determined by the number of octave groups selected (the interested frequency component) and the number of

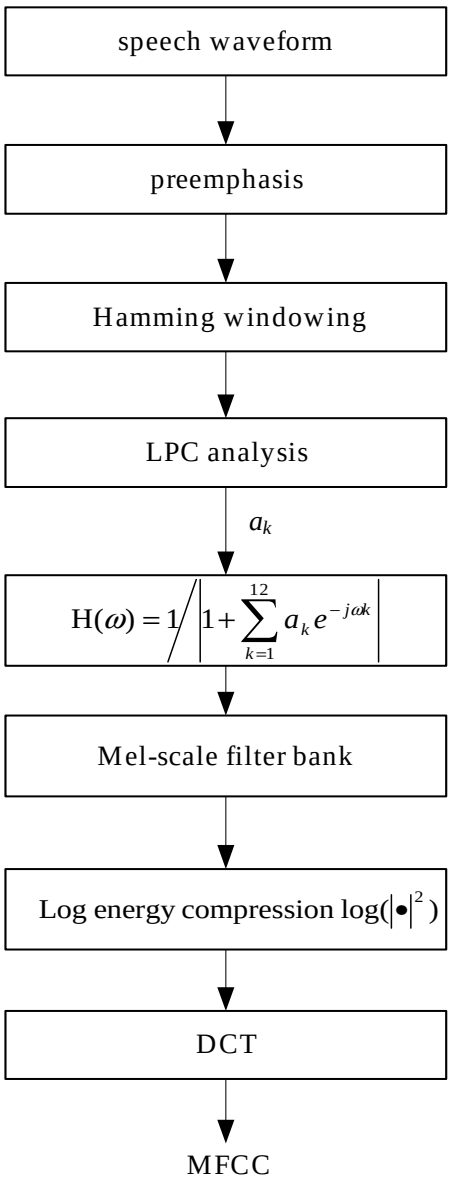


Figure 5.1: Baseline system MFCC feature extraction process

sub-groups within each octave group (the degree of temporal detail required). In Chapter 4 we have described that 6 octave gropus were selected according the properties of the analyzed LP residual signal and the Haar and wavelet functions. The number of sub-groups within an octave will be determined according to the experimental results as described in Section 5.2.

5.1.4 Model training

The UBM-GMM, which has been the dominant technique for model training in speaker recognition, is adopted in our experiments. Firstly, a universal background model (UBM) is trained using the training data from all the registered speakers. Then each speaker's model is adapted from the UBM with the corresponding training data using the *maximum a posterior* (MAP) adaptation approach [88].

5.1.5 Verification and identification tests

The speaker verification performance is evaluated using the testing data from both the registered speakers and the unregistered speakers. For each of the 50 registered speakers, there are 72 claimant tests using his own testing data, 3528 impostor tests from the other 49 speakers, and 1230 unseen impostor tests from the 9 unregistered speakers.

Only the close-set identification performance is evaluated in the experiments. That is, only the testing utterances from the registered speaker (totally 3600 utterances) are tested.

Three feature sets including the two vocal source features HOCOR_α and WOCOR_M and the vocal tract feature set MFCC_E_D_A (for simplicity, noted as MFCCs in the follows) will be evaluated and their performances will be compared. Preliminary fusion of the vocal source and vocal tract feature, where a fixed fusion weight trained using the last 3 sessions of training data, for speaker recognition will also be evaluated.

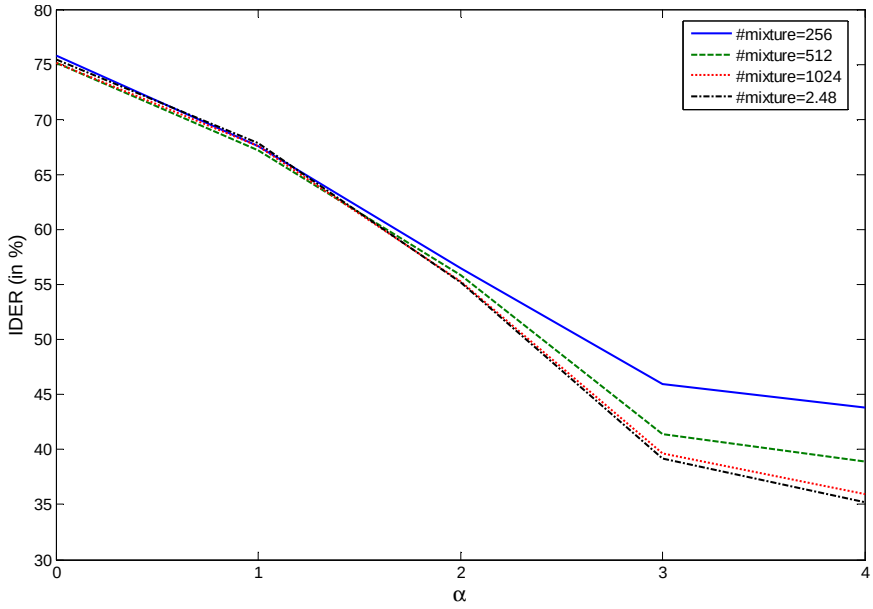
5.2 Speaker Recognition with Vocal Source Features

5.2.1 Feature selection of HOCOR$_\alpha$

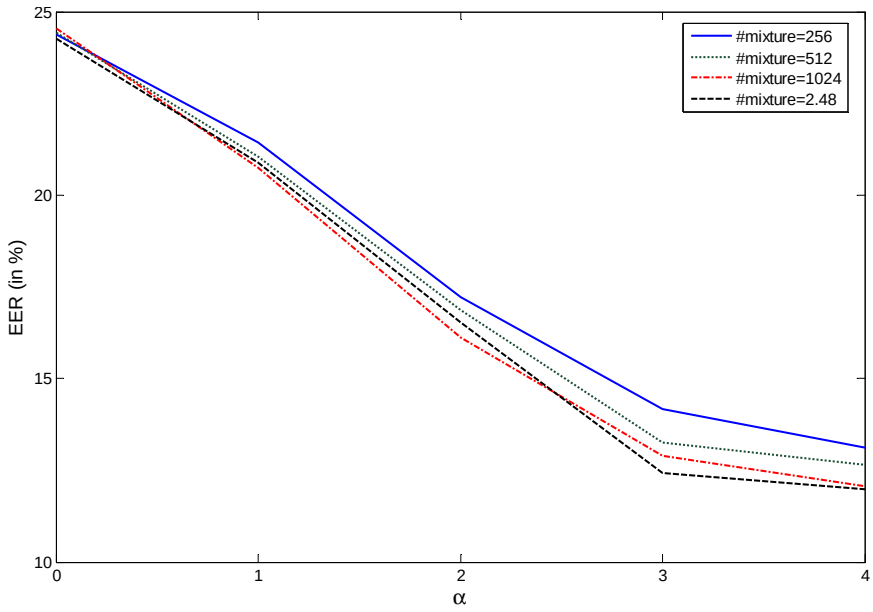
Equation 4.10 defined the time-frequency feature HOCOR$_\alpha$ which captures the spectro-temporal information of the LP residual signal. As mentioned, for a length-256 LP residual signal, the HOCOR$_\alpha$ parameters contains spectral information of 6 octave sub-banks . The temporal details incorporated in the feature set is determined by the parameter α, which will be experimentally determined in this section. To investigate the impact of temporal details on the performance of speaker recognition. Speaker identification and verification performances are evaluated with HOCOR$_\alpha$ in various α. Figure 5.2 elaborates the impact of the temporal information for speaker identification and verification, where the IDER refers to the identification error rate and EER is the verification equal error rate. Since increasing α will increase the feature dimension, which requires more Gaussian mixtures to train the speaker model, recognition results with 256, 512, 1024 and 2048 mixtures GMM are also given in the figure. As illustrated, increasing α from 0 to 3 (feature dimension from 6 to 44) significantly improves the identification and verification performances. HOCOR$_4$ does not outperforms HOCOR$_3$ as significantly. For larger α, more Gaussian mixtures, i.e. 1024 or 2048 as illustrated, are required to model the feature distribution.

5.2.2 Feature selection of WOCOR$_M$

Figure 5.3 shows the identification and verification results of WOCOR$_M$ with M from 1 to 6 and that with various Gaussian mixture number. Similar to that with HOCOR$_\alpha$, as M increases from 1 to 4 (feature dimension from 6 to 24), both identification and verification performances are significantly improved. For $M > 4$, the performances are not improved significantly ($M = 5$), or even degraded ($M = 6$). Increasing Gaussian mixture number from 1024 to 2048 does not improve the performances significantly.

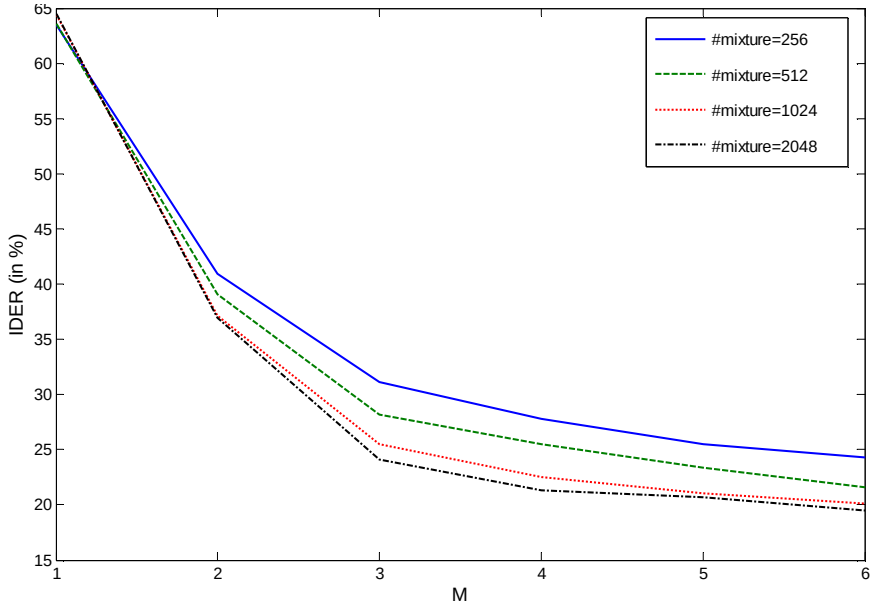


a. Speaker identification performance of HOCOR_α with various α

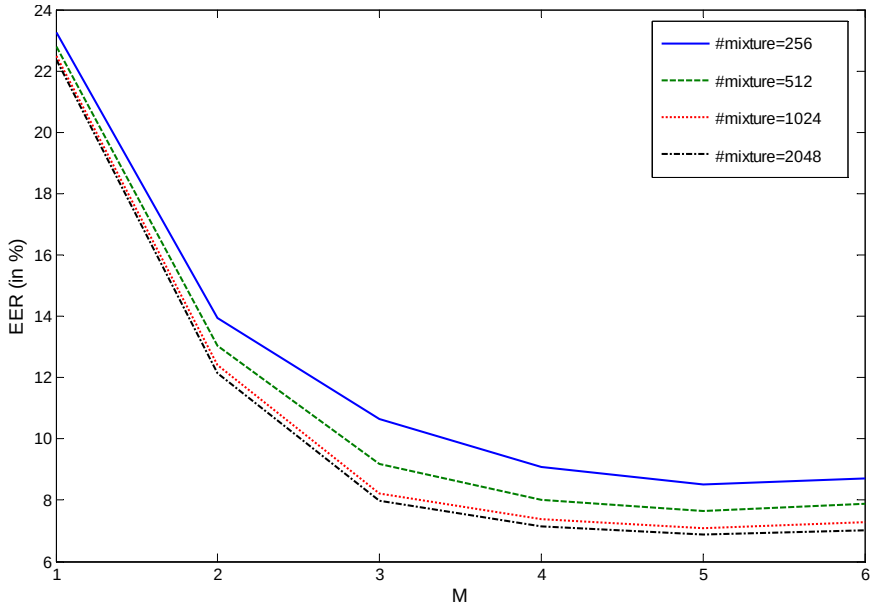


b. Speaker verification performance of HOCOR_α with various α

Figure 5.2: Impact of temporal information in HOCOR_α for speaker recognition



a. Speaker identification performance of WOCOR_M with various M



b. Speaker Verification performance of WOCOR_M with various M

Figure 5.3: Impact of temporal information in WOCOR_M for speaker recognition

5.2.3 Comparison and analyses

As mentioned, the HOCOR_α parameters are extracted form a length-256 segments of LP residual signal. For male speaker whose F0 is generally within the range of 60-250 Hz, a segment of signal could have as more as 8 pitch periods (generally there are 3 or 4 pitch periods in each segment since the most common F0 values are within 80-150 Hz). Therefore it requires larger α (e.g. $\alpha = 3$ and there are 8 sub-groups in each octave) to capture more temporal details for improved recognition performances. Similarly, incorporating more temporal information also improves the performance of WOCOR parameters. However, the WOCOR parameters, are generated with LP residual of exact two pitch periods. Results in Figure 5.3 tell that temporal details with $M = 4$, with which temporal information of every half pitch cycle is retained, is appropriate. The recognition system does not benefit from incorporating more temporal details.

According to recognition performances illustrated in Figure 5.2 and 5.3, the 44 dimensional HOCOR_3 and 24 dimensional WOCOR_4 will be used throughout the experiments in the rest of this thesis. The identification and verification performance of HOCOR_3 and WOCOR_4, both with 1024 mixture GMM training, are given in Table 5.1. It is clear that WOCOR_4 parameters performs better than HOCOR_3. The superiority of WOCOR_4 in speaker discrimination is resulted from:

- Temporal information is better represented in WOCOR_4 due to the pitch-synchronous analysis.
- Better basis function for time-frequency transform in generation of WOCOR_4 as addressed in Section 4.3.

Nevertheless, HOCOR_3 shows a certain degree of speaker discrimination capability. Particularly, the extremely computational simplicity of HOCOR_3 generation makes it a convictive candidate for vocal source speaker discriminative feature in speaker recognition applications where the computational complexity is a critical requirement.

Table 5.1: Speaker recognition performance with $HOCOR_3$ and $WOCOR_4$

Feature	$HOCOR_3$	$WOCOR_4$
IDER (%)	39.64	23.50
EER (%)	12.42	7.39

5.3 Speaker Recognition Using Complementary Vocal Source and Vocal Tract Features

We have evaluated the speaker recognition performance with the proposed vocal source features. The effect of vocal source and vocal tract speaker-specific information are expected to be complementary to each other, since they are derived from two orthogonal components as described in Section 5.1.2. This section evaluates the effectiveness of fusion of source and tract features for speaker recognition.

5.3.1 Training the fusion weights

Generally, the information fusion can be done either on the feature level or on the score level. For feature level fusion, two features are concatenated together to form a single feature vector. Therefore, only one GMM is trained for a speaker. The speaker GMM models not only the distributions of the vocal source and vocal tract feature parameters, but also their joint probability distribution. However feature level fusion requires the two features to be associated with the same time resolution. That is, they should be extracted from the same speech segments. In our feature extraction process, the vocal tract feature MFCCs is generated from speech segments including both voice and unvoice speech, while the source features are generated from only the voiced speech. Especially, the WOCOR parameters are generated from every two pitch periods of voiced speech. Thus, they cannot be concatenated together directly. In this thesis, the vocal source and vocal tract information are fused in the score level. To do so, two GMMs are trained for each speaker, one modeling the MFCCs and

the other modeling the $HOCOR_3$ or $WOCOR_4$. In identification test, the final score is the linear combination of the log-likelihood scores given by the source feature ($HOCOR_3$ or $WOCOR_4$) and the tract feature MFCCs, i.e.

$$\begin{aligned} L &= w_t L_t + w_s L_s \\ with \quad & w_t + w_s = 1 \end{aligned} \tag{5.2}$$

where w is the weighting parameter, L is the log-likelihood score, subscripts t and s refer to the vocal tract and vocal source, respectively.

Similarly, the score fusion for verification is given by

$$\begin{aligned} \Lambda &= w_t \Lambda_t + w_s \Lambda_s \\ with \quad & w_t + w_s = 1 \end{aligned} \tag{5.3}$$

where Λ_t and Λ_s are the log-likelihood ratio scores given by source and tract features, respectively.

In this experiment, w_t and w_s are experimentally determined in the development stage using the last three sessions of training data (session 4 to 6). Figure 5.4 and 5.5 show the identification and verification performances with various w_t. The w_t giving the best recognition results will be selected as the fusion weight in the testing. According to Figure 5.4 and 5.5, for identification, w_t is 0.82 for fusion of MFCCs and $HOCOR_3$ and 0.8 for fusion of MFCCs and $WOCOR_4$. For verification, w_t is 0.6 and 0.7, respectively.

5.3.2 Identification results

Table 5.2 compares the identification results with source and tract features and that with fused information. It is clear that although the vocal source features perform worse than the vocal tract feature (39.64% with $HOCOR_3$ and 23.5% with $WOCOR_4$ vs. 1.44% with MFCCs), they reduce the identification error as supplementary features to MFCCs. The system combining MFCCs and $HOCOR_3$ results in 1.17% identification error rate, relatively improves the MFCCs performance by 18.7%. Information fusion with MFCCs and $WOCOR_4$ achieves an identification error rate of 0.94%, a relative improvement of 34.6%.

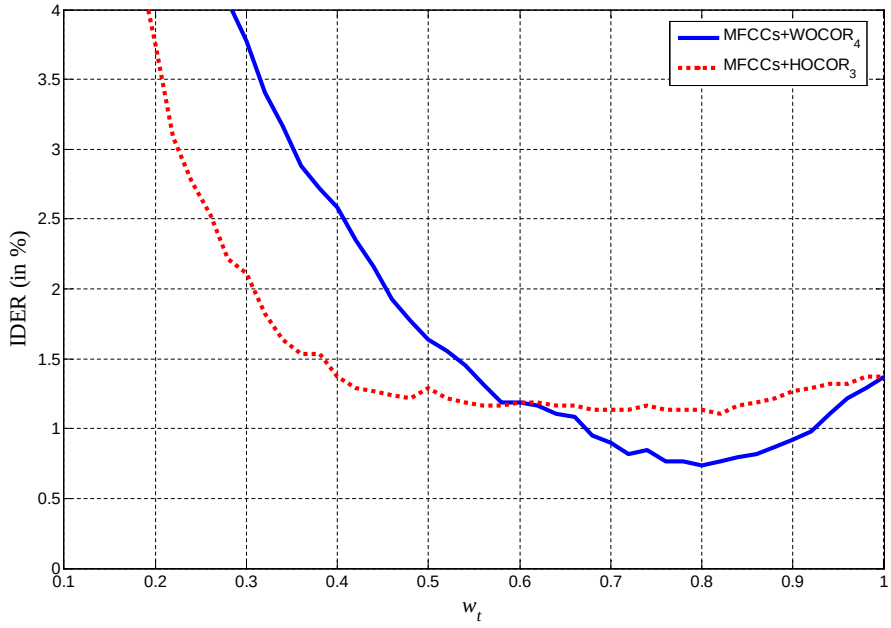


Figure 5.4: Identification error rate with various information fusion weights

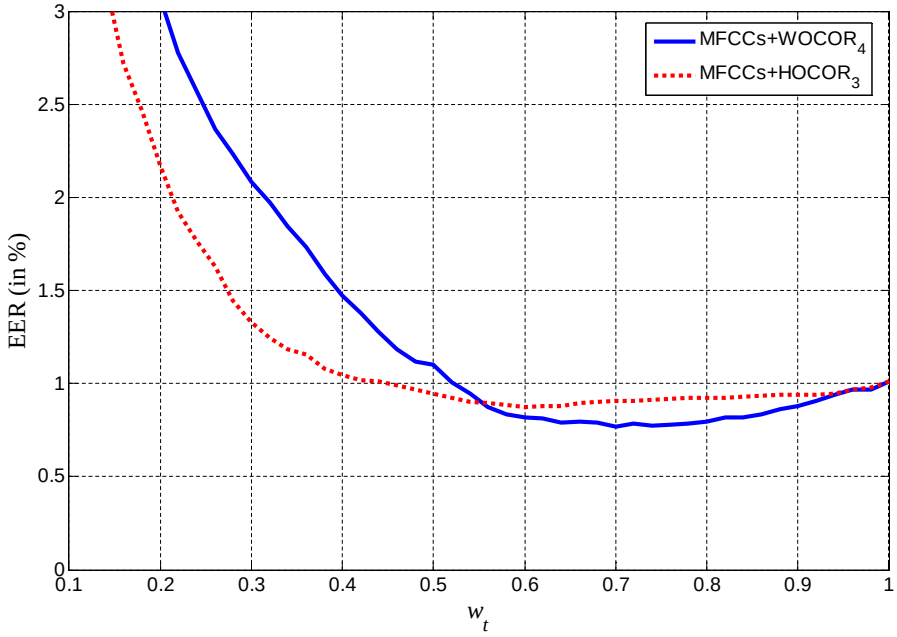


Figure 5.5: Verification equal error rate with various information fusion weights

Table 5.2: Speaker identification error rate

Feature	IDER (in%)
MFCCs	1.44
$HOCOR_3$	39.64
$WOCOR_4$	23.50
$HOCOR_3$+MFCCs	1.17
$WOCOR_4$+MFCCs	0.94

5.3.3 Verification results

The speaker verification performances are illustrated via DET curves as in Figure 5.6. It is clear from the DET curves that although the verification performances of $HOCOR_3$ and $WOCOR_4$ are not convincing in comparison with that of MFCCs, source and tract information fusion really improves the overall performance. Table 5.3 shows the EERs of the different features and the source-tract information fusion. As illustrated, with fusion of MFCCs and $HOCOR_3$, EER is reduced from 1.06% to 0.94%, a relative improvement of 11.3%. Fusion of MFCCs and $WOCOR_4$ reduces EER from 1.06% to 0.81%, a relative improvement of 23.6%.

Table 5.3: Speaker verification equal error rate

Feature	EER (in %)
MFCCs	1.06
$HOCOR_3$	12.42
$WOCOR_4$	7.39
$HOCOR_3$+MFCCs	0.94
$WOCOR_4$+MFCCs	0.81

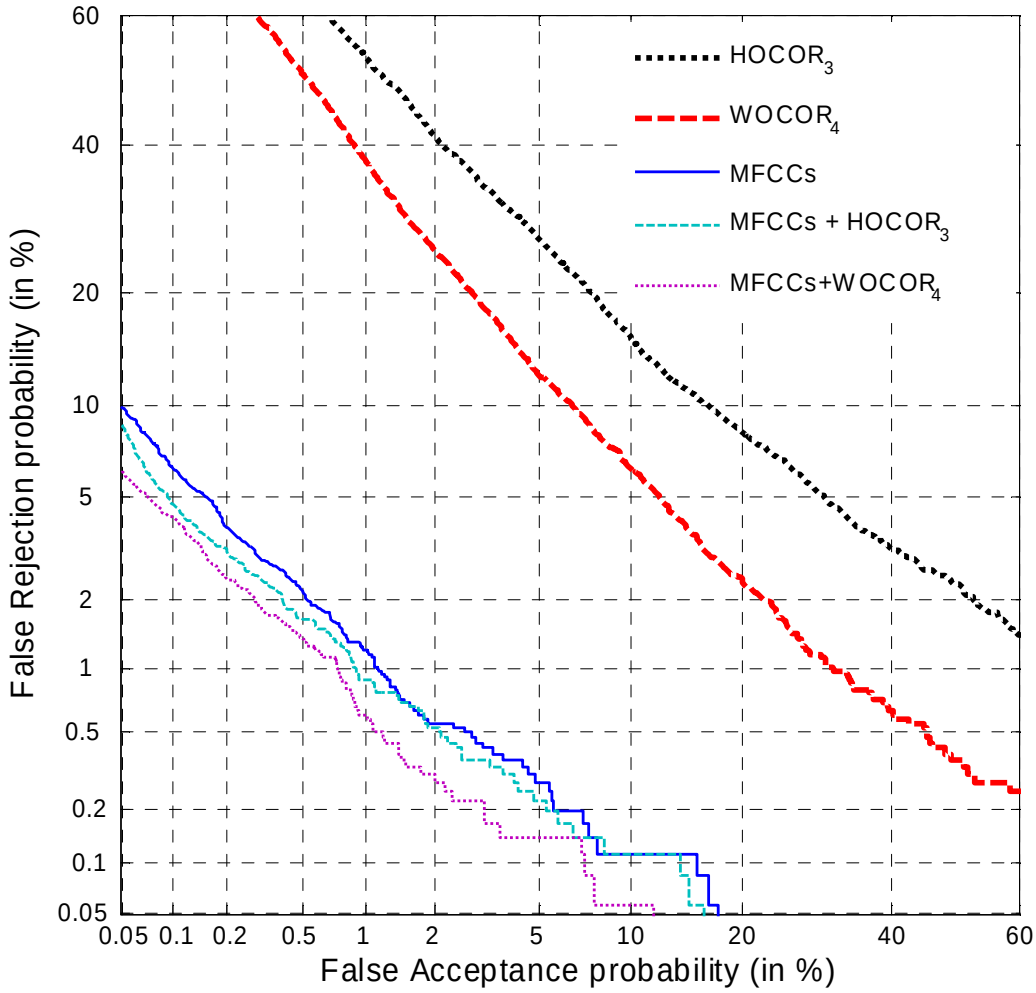


Figure 5.6: DET curves with different features and source-tract info-fusion

5.4 Concluding Remarks

From the experimental results demonstrated in this Chapter, we can draw the following conclusions:

- The temporal information is very important in characterizing the vocal source excitation and is useful for speaker recognition.
- The vocal source features performs worse than the vocal tract features, because the information from vocal cords phonation is ultimately not as rich as that from the vocal tract articulation. What's more, the glottal phonation can be easily altered at will, which results in larger intra-speaker variation of the vocal source features than that of the vocal tract features.
- Fusion of vocal tract and vocal source features improves the system performance. The complementarity between these two kinds of feature is therefore verified.

Chapter 6

Optimized Information Fusion with Discriminative Analysis

The previous chapter demonstrated that improvement of speaker recognition accuracy can be obtained by incorporating vocal source features to supplement the vocal tract ones. A fixed weighting parameter, which was experimentally determined using the training data, was used for the source-tract information fusion. Generally, fusion of the two complementary features will result in a better recognition performance than using the vocal tract feature only. However, as to a specific recognition trial, the effect of incorporating source information may be somewhat uncertain. As demonstrated, vocal tract features generally perform better in speaker recognition than vocal source features. For those cases where vocal source information have errors in recognizing a speaker, incorporating such confusing information could end up with worse results, whilst the vocal tract feature itself might have a correct recognition. To avoid this problem, discrimination-dependent weighting schemes have been exploited for the optimized fusion of source and tract speaker information. In this chapter, rather than the fixed weight, a varying fusion weight, the selection of which is based on the discrimination ratio of the two features, is derived on the fly to take full advantage of the complementarity of the source and tract information for speaker identification.

For speaker verification, a text-dependent fusion scheme is also developed

for improved verification performance. The fixed weighting scheme ignores the diversity of speech production, which has been found to result in different speaker discrimination power across different sounds. Therefore, a text-dependent weighting scheme could be more desirable for the fusion of source and tract features if the spoken text is known. This chapter will also investigate the discrimination power of different Cantonese digits, with which a digit-dependent weighting scheme is developed for speaker verification.

The work in this chapter has been shown in [123][124]

6.1 Information Fusion with Confidence Measure for Speaker Identification

In speaker identification, the confidence measure is a metric to evaluate the ability of a specific feature (here, either MFCCs or WOCOR_4) to discriminate speakers. Previous recognition results showed that the MFCCs feature generally has larger discrimination power than WOCOR_4. Thus, MFCCs is more confident for speaker identification than WOCOR_4. However, to a specific trial, as discussed in Section 3.4, two speakers may have very similar vocal tract configurations while quite different vocal source phonation styles, and the resulting WOCOR_4 feature will be more confident than MFCCs. Incorporating WOCOR_4 to supplement MFCCs can therefore reduce the identification errors. Contrarily, if WOCOR_4 has less confidence, extra identification errors could arise with information fusion. Table 6.3 shows the identification errors with MFCCs only and that with both MFCCs and WOCOR_4. As illustrated, 24 of the total 52 errors with MFCCs have been corrected by incorporating WOCOR_4. However, 6 new errors, which are correctly identified by MFCCs, are introduced when fusing WOCOR_4 and MFCCs. Therefore, it is necessary in each of the identification trial to emphasize the more confident feature and avoid using the confusing counterpart.

Table 6.1: Identification errors with fixed weight info-fusion

MFCCs	MFCCs+WOCOR$_4$			
	error reduced	error remained	new errors	total
52	24	28	6	34

6.1.1 Derivation of confidence measure

Analysis of the matching score of the identification tests shows that, generally, in a correct identification, the difference of the matching scores between the identified speaker and his closest competitor is relative larger than that in an incorrect identification. The score difference can therefore be adopted for measuring the discrimination power, i.e.

$$D = \frac{\max\limits_i\{L_i\} - second\max\limits_i\{L_i\}}{\left|\max\limits_i\{L_i\}\right|} \tag{6.1}$$

where L_i is the log-likelihood score of the i-th speaker. The normalization of the difference over the maximum log-likelihood score can reduce the effect of the dynamic range of L.

Figure 6.1 shows the histogram of D of MFCCs and WOCOR$_4$ features. It is clear that a correct identification is generally associated with a larger D than an incorrect identification. Therefore, a larger D tells that the corresponding feature has a higher confidence for speaker identification. Obviously, it is preferable to take into account D for information fusion in each test instead of using the fixed weight. Although the optimal way to fuse the two scores with knowledge of the discrimination powers of the two features is not known, we found that a confidence measure which embeds the discrimination ratio into the sigmoid function improved the identification performance, especially reduced the extra errors introduced by information fusion with fixed weight. The confidence measure is defined as

$$CM = -\log\frac{1}{1+e^{(-\alpha\cdot DR)}} \tag{6.2}$$

where $DR = D_t/D_s$ is the discrimination ratio between the vocal tract and vocal source features, α is the scaling parameter which controls the slope of the

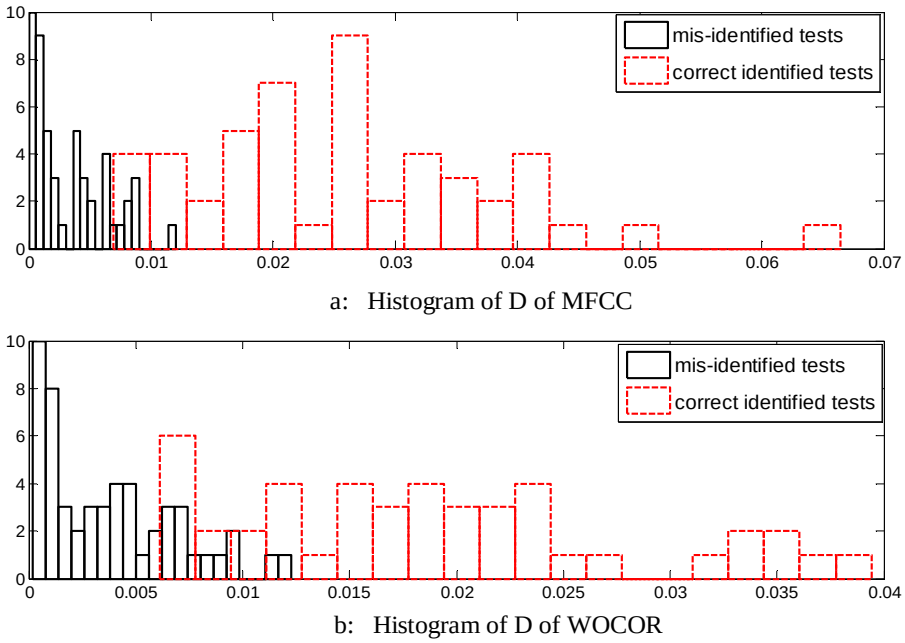


Figure 6.1: Histogram of discriminative power of two features

mapping contour from DR to CM.

Figure 6.2 illustrates the mapping from DR to CM with the log-sigmoid function with various α. With larger α, CM goes to zero more quickly as DR increases, so that the confident feature will be more emphasized.

The score fusion with confidence measure for each identification trial is now as

$$L_i = L_t + L_s \cdot CM \tag{6.3}$$

where the subscripts t and s refer to the vocal tract feature and the vocal source feature, respectively. With the confidence measure CM, the fused score combines better weighted L_t and L_s . As illustrated in Figure 6.2, when DR has a larger value, CM tends to zero, the final decision will not be heavily affected by the vocal source score. Contrarily, small DR corresponds to a large CM, which will introduce more impact of vocal source for final decision.

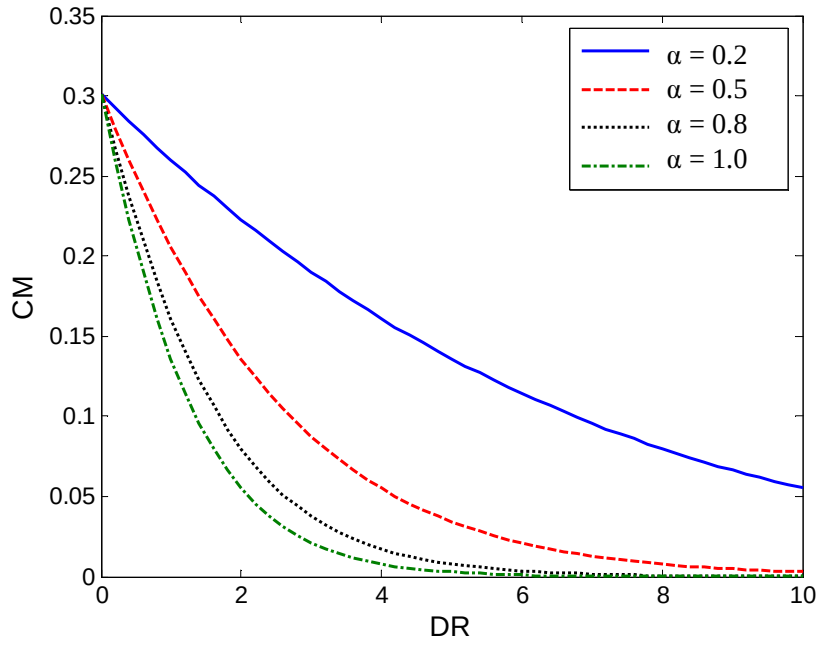


Figure 6.2: Confidence measures CM with various α

6.1.2 Identification results

Speaker identification error rate (IDER) is given as in Table 6.2. The speech corpus and the experimental procedure is the same as that in Section 5.1. The values of α is trained using the training data. It is clear that information fusion with confidence measure further reduces the IDER from 0.94% to 0.78%, a relative improvement of 17% compared with that using fixed weight. In comparison with that using vocal tract feature only, a relative improvement of 46% has been achieved.

Table 6.2: Speaker identification error rate in matched conditions

Feature	MFCC	WOCOR	Info-fusion	
			$w_s = 0.2$	CM $(\alpha = 0.8)$
IDER (%)	1.44	23.5	0.94	0.78

Table 6.3 elaborates how the complementary features and the optimized information fusion can reduce the identification errors. By fusion with fixed

weight, 24 of the total 52 errors with only vocal tract feature have been corrected. However, 6 new errors, which are correctly identified by the vocal tract feature, are introduced. These errors are avoided in the optimized information fusion with confidence measure. Table 6.4 shows the discrimination ratio and the confidence measure for the 6 new introduced errors. Compared with the fixed weight $w_s = 0.2$, which corresponds to a fixed CM of 0.25, it is clearly that with the proposed confidence measure, the vocal source impact in these trials is greatly reduced.

Table 6.3: Comparison of the errors with two info-fusion methods

MFCC	52			
Info-fusion	reduced errors	remained errors	new errors	total errors
with w_s	24	28	6	34
with CM	24	28	0	28

Table 6.4: DR and CM for the *new errors* in Table 6.3

Error No.	1	2	3	4	5	6
DR	1.2	3.2	1.8	8.3	1.3	10
CM	0.14	0.03	0.09	0.00	0.13	0.00

6.2 Text-dependent Information Fusion for Speaker Verification

speech production is a complicated process incorporating the joint effect of vocal cords phonation and vocal tract articulation. As mentioned, the vocal cords phonation contributes little to phoneme classification. It is therefore relative stable for a speaker uttering different sounds. The vocal tract configuration, however, contributes primarily for phoneme classification and secondary for speaker

discrimination. The vocal tract structure varies significantly for uttering different sounds. It is therefore reasonable to infer that the speaker discrimination power of vocal tract features might vary significantly across different sounds. Information fusion with text-independent weight as applied in Chapter 5 ignores the difference of discrimination power between different sounds. To take full advantage of the complementary vocal source and vocal tract features, we developed a text-dependent weighting scheme for fusion of these two features. As an example, this section analyzes the digit-dependent discrimination power of the 10 Cantonese digits. Then the digit-dependent source-tract fusion weights are trained for optimized information fusion for speaker verification.

6.2.1 Analysis of digit-dependent source and tract speaker discrimination power

The digit-dependent speaker discrimination power of vocal source and vocal tract features are investigated by comparing verification score distributions attained with the respective features. For a speaker verification system, the performance is determined by the distributions of the log-likelihood ratio (LLR) scores of the claimant and impostor tests, as shown in Figure 6.3, where $D(\Lambda_i)$ and $D(\Lambda_c)$ are respectively the distributions of LLRs of impostor and claimant tests and the overlap area of $D(\Lambda_i)$ and $D(\Lambda_c)$, P_e is the minimum total error probability including both false rejection and false acceptance errors. A good verification system should has small P_e. That is, $D(\Lambda_i)$ and $D(\Lambda_c)$ should be as separable as possible, which requires: (1) large distance between the mean values of the two distributions (the *inter-score variation*); and (2) small variances of the two distributions (the *intra-score variation*).

To analyzed the speaker discrimination power of the source and tract feature of each digit, time segmentation at digit level of each testing utterance is obtained by HMM forced alignment. Each digit in the utterance is used as a single test. Figure 6.4 shows $D(\Lambda_i)$ and $D(\Lambda_c)$ given by MFCCs, WOCOR_4 and their fused score for each of the Cantonese digits "0"-"9". The P_e values of each subplot are given in Table 6.5.

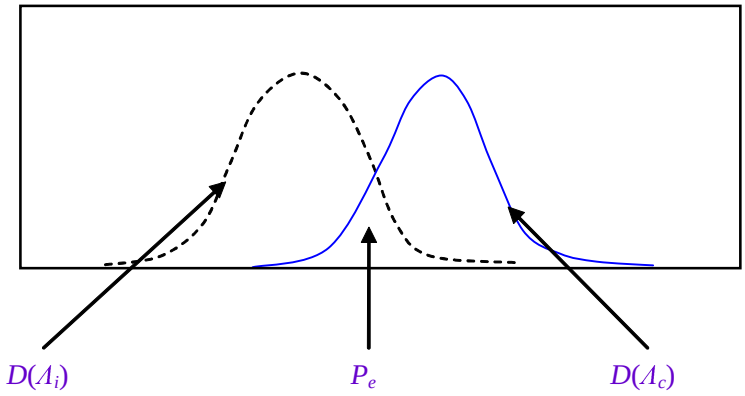


Figure 6.3: Distributions of claimant and impostor LLR scores of a SV system

For each digit, we define the source-tract discrimination ratio to be

$$r = \frac{1/P_e \text{ of source feature}}{1/P_e \text{ of tract feature}} = \frac{P_e \text{ of tract feature}}{P_e \text{ of source feature}} \tag{6.4}$$

Larger r corresponds to more contribution of vocal source feature for speaker discrimination, and vice versa. Table 6.5 shows the source-tract discrimination ratios as defined above.

Table 6.5: P_e and r values of Cantonese digits

Digit	P_e of WOCOR	P_e of MFCC	P_e of Info-fusion	r
0	0.40	0.17	0.12	0.42
1	0.41	0.20	0.13	0.49
2	0.43	0.24	0.18	0.56
3	0.34	0.13	0.09	0.38
4	0.35	0.15	0.12	0.43
5	0.39	0.29	0.18	0.75
6	0.44	0.24	0.15	0.52
7	0.45	0.17	0.12	0.38
8	0.35	0.17	0.11	0.49
9	0.37	0.15	0.11	0.41

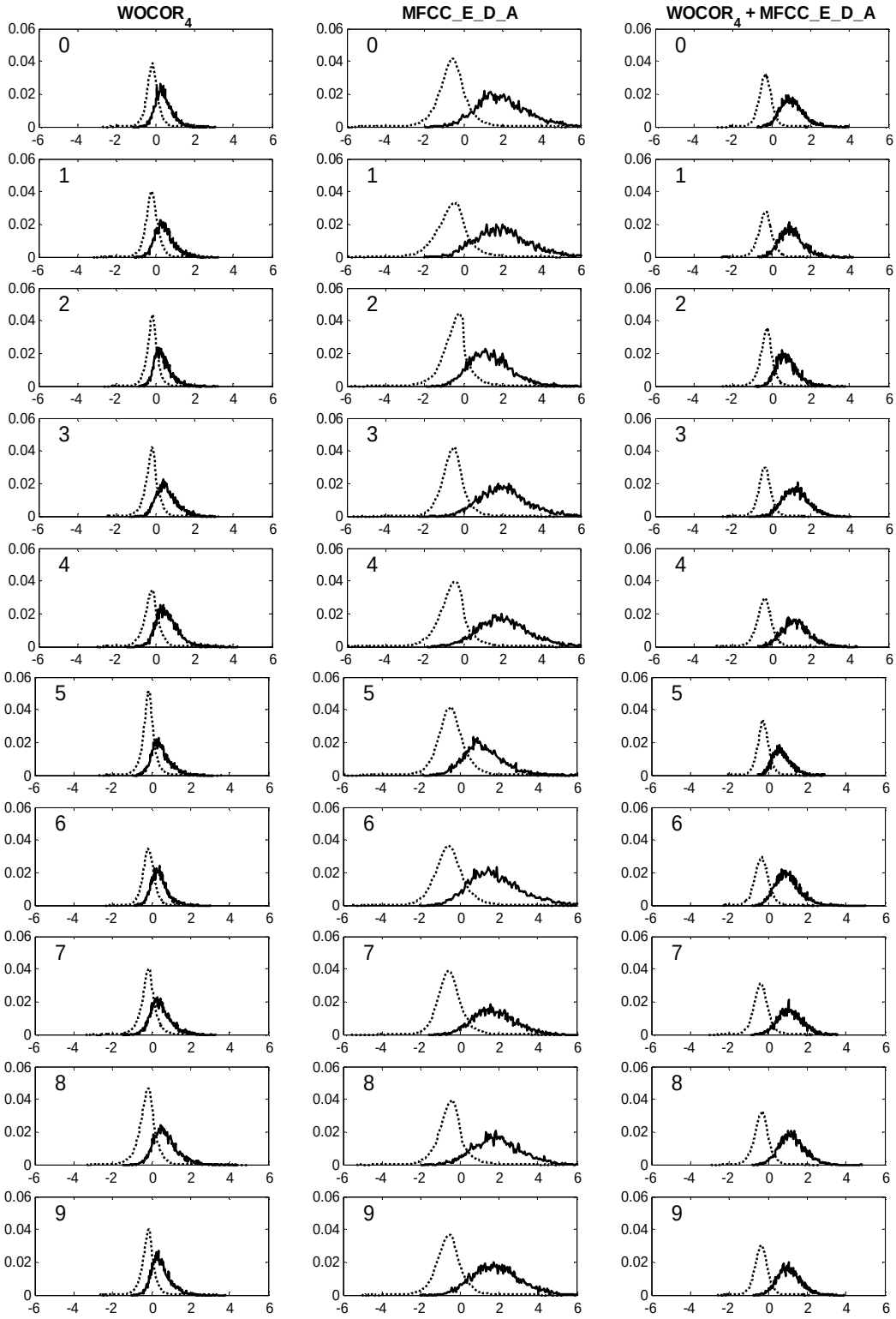


Figure 6.4: Claimant-impostor LLR distributions of source and tract features of 10 Cantonese digits

From Figure 6.4 and Table 6.5 we can see that

- LLRs of MFCCs have large inter-score variation and also large intra-score variation. LLRs of WOCOR_4 have relative small inter-score and intra-score variations. Information fusion gives both moderate inter- and intra-score variations and the overall errors are reduced. This phenomenon partially verifies the complementarity of the two features.

- P_e of MFCCs varies more significantly across different digits than P_e of WOCOR_4. This phenomenon is consistent with our inference that the vocal tract features has larger variation of speaker discrimination power due to the significant variation of vocal tract configuration for uttering different sounds.

- For all digits, P_e of WOCOR_4 is greater than P_e of MFCCs and $r < 1$, which means that the vocal tract feature has higher discrimination power than the vocal source feature.

- r varies significantly across different digits. For example, r is 0.75 for digit "5" while 0.38 for digit "3" and "7". The digit-dependent discrimination power of the two features calls for digit-dependent weights to maximize the benefits obtained from the fusion of vocal source and vocal tract features.

6.2.2 Verification results with digit-dependent information fusion

The digit-dependent source-tract weights are empirically determined. That is, for verification test with each digit, the final LLR is calculated as the weighted combination of the two LLRs

$$\Lambda = w_s\Lambda_s + (1 - w_s)\Lambda_t \tag{6.5}$$

where w_s is the fusion weight for the source feature. EERs are calculated with w_s varying form 0 to 1 in a step size of 0.01. The weight giving the least EER is selected as the optimal weight.

Table 6.6: EERs and the optimal w_s for the digits

Digit	EER (in%)			Optimal w_s
	WOCOR	MFCC	Info-fusion	
0	20.6	8.73	7.19	0.57
1	20.6	10.4	7.92	0.60
2	21.9	12.8	10.8	0.56
3	17.6	6.78	5.57	0.42
4	18.0	8.12	6.47	0.49
5	21.7	16.9	13.2	0.65
6	22.5	11.5	9.41	0.51
7	23.7	8.42	7.38	0.41
8	17.7	8.80	6.81	0.52
9	18.8	8.48	5.98	0.51

Table 6.6 gives the EERs of WOCOR_4, MFCCs and information fusion with the optimal weights of 10 Cantonese digits. Note that w_s greater than 0.5 does not necessary mean that WOCOR_4 contributes more than MFCCs to speaker discrimination. It is also related to the relative ranges of Λ_s and Λ_t. From Table 6.6, it is also clear that EER varies significantly across digits even with the optimal weighted information fusion. This means that different digits contributes differently to speaker discrimination. Thus, when calculating the overall LLR score of an entire testing utterance, LLR score of each digit should be sum up with different weights. This will be the future work and is not currently covered in this thesis.

The efficiency of the proposed digit-dependent weights for source-tract information fusion has been evaluated over the Cantonese database described in Chapter 5. The source and tract features are WOCOR_4 and MFCCs, respectively. For each verification test, the boundary of digit is detected by forced alignment with the known spoken text. Then the LLR scores of each digit is calculated by fusion the LLRs given by WOCOR_4 and MFCCs with the digit-

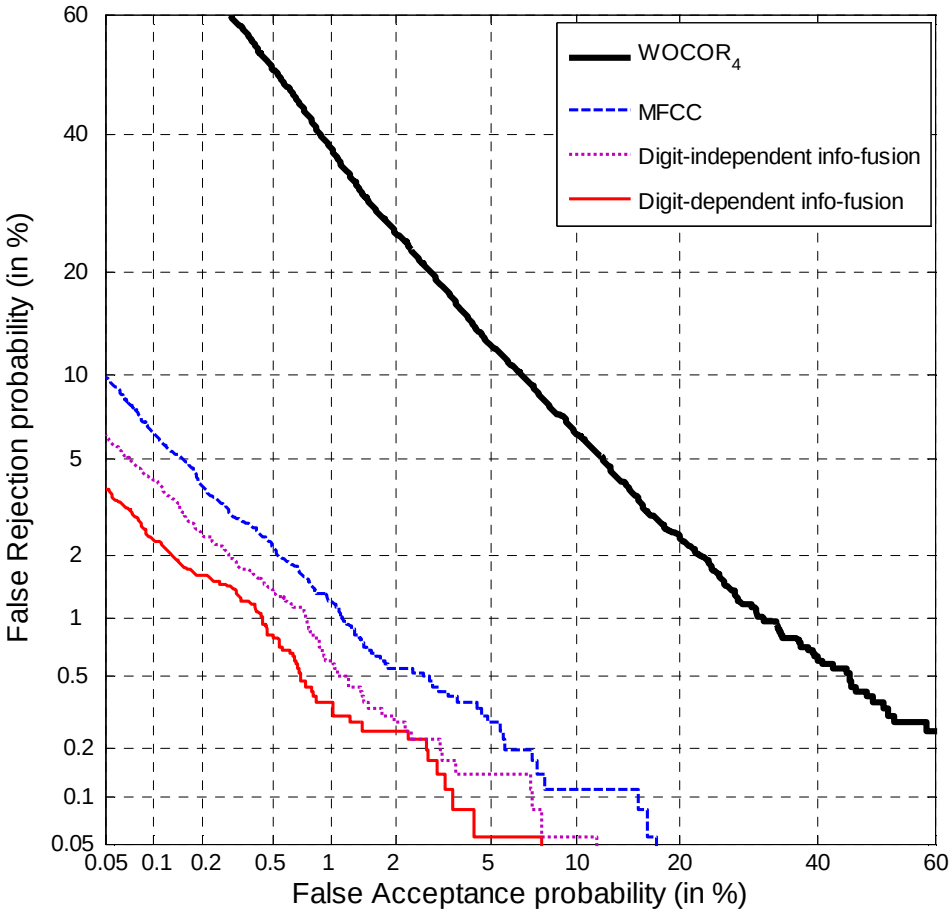


Figure 6.5: DET curves for speaker verification with digit-independent and digit-dependent information fusion

dependent fusing weights listed in Table 6.6. Finally the utterance level LLR score is calculated by summing up the LLRs of all the digits. The verification performance is illustrated via DET curves in Figure 6.5. As shown, The performance is further improved with digit-dependent information fusion. The EER is further reduced from 0.81% with digit-independent fusion weight to 0.64% with digit-dependent weight, as illustrated in Table 6.7. As a whole, compared with EER of MFCCs, a relative improvement of 39.6% is achieved.

Table 6.7: EERs of different information fusion methods

Feature	MFCCs	Digit-indepen. Info-fusion	Digit-depen. Info-fusion
EER (%)	1.06	0.81	0.64

6.3 Conclusion

Although fusion of the complementary vocal source and vocal tract information generally improves the recognition performance, discrimination dependent fusion weight is necessary for maximizing the benefits through the information fusion. We developed a varying weighting scheme for speaker identification, in which the fusion weight is online derived based on the discrimination ratio between the source and tract features. Text-dependent fusion weights were also derived for speaker verification. Both methods further increase the recognition accuracy.

Chapter 7

Discussions

This chapter discusses a few other issuses related to the speaker recognition performance. Section 7.1 illustrates the robust performance of the proposed source-tract information fusion for speaker recognition in mismatched conditions and gives some further considerations on robust vocal source feature extraction. Section 7.2 discusses the long-term variation in vocal source excitation and the modeling of such temporal characteristics for speaker recognition. Section 7.3 compares several vocal source features and their effectiveness in speaker recognition. Finally we discuss some further considerations including training UBM using different database, different methods in MFCC feature extraction, and the expected performance of source-tract information fusion with very large speaker population size.

7.1 Robust Speaker Recognition with Complementary Vocal Source and Vocal Tract Features

Chapter 5 and 6 demonstrated the effectiveness of the proposed source-tract information fusion scheme for speaker recognition in matched conditions. In this section, the usefulness of the proposed method for robustness speaker recognition in mismatched condition is evaluated.

7.1.1 Experimental setup

The database adopted for robust speaker recognition evaluation is an English corpus that had been used in the NIST 2001 speaker recognition evaluation program[112]. Only the *One Speaker Detection* subset of this corpus is used in this evaluation. There are 74 male speakers and 100 female speakers in this subset. Each speaker has a training utterance containing speech data with duration around 2 minutes. There are totally 850 testing utterances from the male speakers and 1188 testing utterances from the female speakers. The duration of testing utterances varies from 15 to 45 seconds. All the utterances are extracted from telephone conversational speech between two speakers. The conversations happened in the indoor, outdoor and car background environments. Various phones produced by different corporations, e.g. Motorola, Bell, Nokia, etc. had been used by the speakers. The speech data were collected over three telephone channels including fixed-line, GSM and cellular networks. Therefore, it is a text-independent speaker recognition task with various background noise and handset and channel distortions.

The vocal source and vocal tract features to be evaluated are WOCOR_4 and MFCCs, respectively. The cepstral mean normalization (CMN) technique is applied on the static MFCC parameters to eliminate the time-invariant channel and handset distortions. GMM-UBM method is adopted for model training and score normalization. The recognition performances are evaluated on male and female data separately. All the testing utterances are used for identification test. For verification, each utterance is tested over 11 speakers within whom one is the target speaker and the other ten are the impostors.

7.1.2 Recognition results

The identification and verification performances on male and female subsets are elaborated in Table 7.1 and 7.2. We can see that the performances of both vocal source and vocal tract features degrade significantly in mismatched conditions. Similar to that observed in matched conditions, the vocal source

features perform worse than the vocal tract feature. Nevertheless, fusion of these two features, where the score level fusion with fixed weight as described in Section 5.3 is adopted, improves the recognition performances. For male speaker evaluation, fusion of $WOCOR_4$ and MFCCs reduces IDER from 24.71% to 20.94% and EER from 8.47% to 7.40%, relative improvements of 15.3% and 12.6% have been achieved. For female speaker evaluation, IDER and EER are reduced from 32.33% and 6.74% with MFCCs only to 28.61% and 5.80%, relative improvements of 11.5% and 14.0%, respectively.

Table 7.1: Male speaker recognition performance in mismatched conditions

Feature	IDER (in %)	EER (in %)
MFCCs	24.71	8.47
$WOCOR_4$	44.71	19.65
$WOCOR_4$+MFCCs	20.94	7.40

Table 7.2: Female speaker recognition performance in mismatched conditions

Feature	IDER (in %)	EER (in %)
MFCCs	32.33	6.74
$WOCOR_4$	56.88	23.11
$WOCOR_4$+MFCCs	28.61	5.80

7.1.3 Further considerations on robust vocal source feature extraction

The preliminary experimental results as given in Table 7.1 and 7.2 show that our information fusion scheme also improves the robustness of speaker recognition systems in such mismatched conditions. It is expected that the performance can be further improved with some additional processing for robust vocal source feature extraction. For example, it has been found that different time-frequency

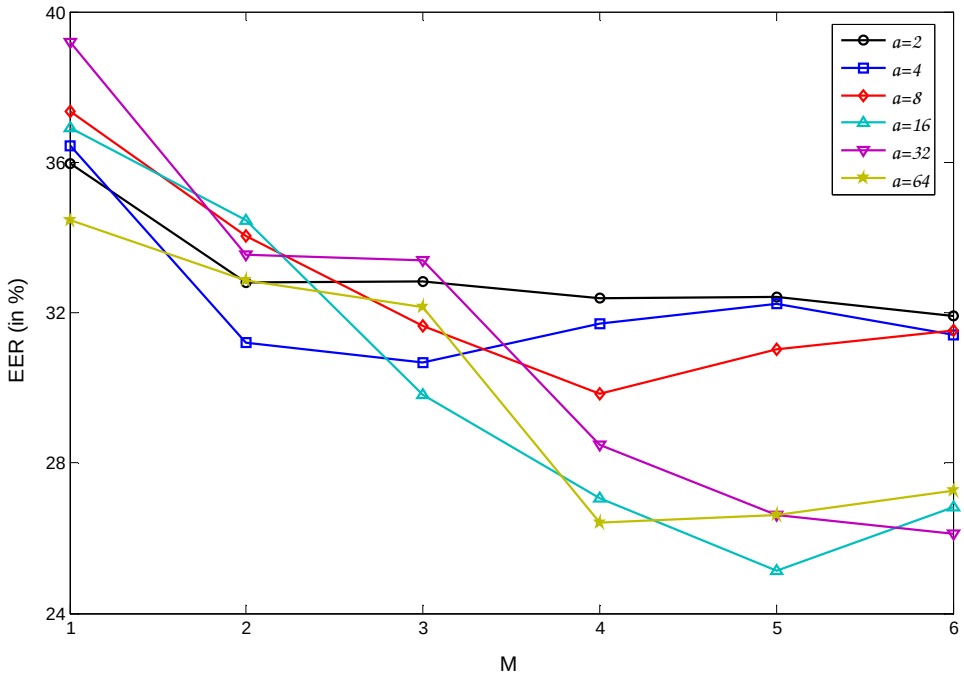


Figure 7.1: EERs of time-frequency components within WOCOR_M parameters

components of WOCOR_M feature set demonstrate different speaker discrimination ability as illustrated in Figure 7.1, in which EERs of time-frequence feature components from a specific octave group (e.g. $a = 2, 4, ..., 64$) are drawn in the same curve, with increasing temporal detail incorporated (i.e. M increase from 1 to 6). Figure 7.1 tells that:

- For the higher frequency components (i.e. $a = 2$ and 4), only roughly temporal detail ($M = 2$) is required. As known, the high frequency components contains lots of noise information. Their temporal details does not contribute much to speaker characterization. Therefore, the amount of high frequency information within each pitch cycle as a whole is enough.

- For those lower frequency components, more detailed temporal information can further improve the recognition performance. For example, it is demonstrated in the figure that the optimal value of M for $a = 8, 16, 32$ and 64 are 4, 5, 6 and 4, respectively.

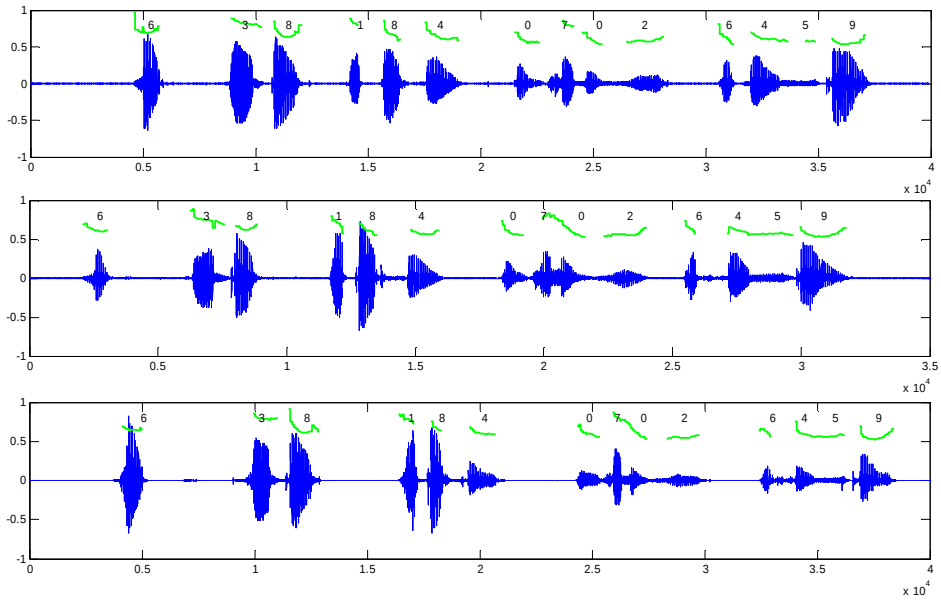
- The lower frequency component contributes more to speaker recognition,

since they capture pitch and harmonics related information and they are also less affected by the noise.

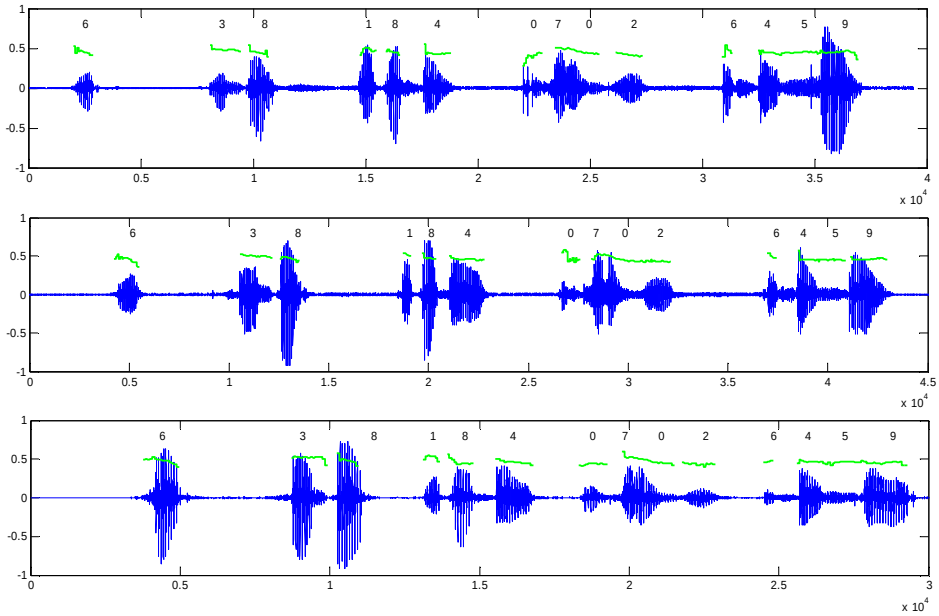
Therefore, time-frequency feature selection according to their performance illustrated in Figure 7.1 is expected to further improve the recognition accuracy and robustness. This topic will be the future work and is not covered in the thesis.

7.2 Modeling the Multi-scale Temporal Information

In Section 4.3, we have described how time-frequency acoustic features can be generated by applying wavelet transform on pitch synchronized segments of LP residual signals. The new feature set WOCOR_M is capable of capturing the multi-bank spectral properties and their temporal variations within each pitch cycle and between the consecutive two pitch cycles (the short-term variation). This kind of spectro-temporal characteristics has been verified to be useful for speaker recognition. However, temporal variation across several pitch cycles (the long-term variation), has been ignored. The long-term variation of glottal phonation always exists in real speech, and it could be significant for some speakers. The variation of vocal cords phonation is primarily represented as the change in pitch periods (or the fundamental frequency F0). Particularly, in tonal language, F0 could change significantly within a phoneme, either to manifest the tone label or to be consistent with the leading and following phonemes in co-articulation in continuous speech. Figure 7.2 illustrates the long-term variation (pitch contour) of two male speakers speaking the same utterance at 3 different time. It is clear that the two speakers have different pitch variations in speaking the same digit string, whilst the utterances from the same speaker at different time have similar pitch contours. It is found that pitch variation has a great impact on the resulting LP residual signal. For example, great pitch variation (*jitter*) results in a strong degree of inter-pulses noise and great pulse amplitude variation (*shimmer*) in the LP residual signal.



a. Speaker A



b. Speaker B

Figure 7.2: Speech waveform and the F0 contour (normalized by 200 Hz) of two speakers speaking the same utterance at three different time

To exploit the usefulness of such pitch variation related long-term information for speaker recognition, the dynamic feature is derived from the static feature parameters WOCOR_4 using the following regression formula [117]

$$D_t = \frac{\sum_{\theta=1}^{\Theta} \theta(S_{t+\theta} - S_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2} \tag{7.1}$$

where S_t and D_t are the t-th static and dynamic features, respectively, and $\Theta = 2$.

To extend further so as to exploit the temporal variation during the pronunciation of a digit, Hidden Markov Models (HMM) is applied for modeling the static and dynamic features. HMM is useful in text-dependent speaker recognition since it models not only the distribution of features generated from each speech segment, but also their temporal variations. In this case, spectro-temporal characteristics in multiple scales, i.e. within every pitch cycle, between consecutive cycles, across several cycles, and among the whole time span of pronouncing a digit as well, can be represented.

To evaluate the contribution of the multi-scale temporal information modeling, speaker recognition experiments are carried out on the Cantonese database as described in Chapter 5. The training of HMMs for each speaker is similar to the UBM-GMM training technique. First, 10 universal background HMMs (UBHMM), each corresponding to a digit, are trained using all the training data from the enrolled speakers. Then, for each speaker, HMMs for each digit are adapted from the corresponding UBHMMs using his own data. Each UBHMM and HMM has 6 states including the entering and ending states. The Gaussian mixture number of each state is 128, which is experimentally determined according to recognition results.

The usefulness of such long-term variation of source excitation in speaker characterization has been confirmed by the recognition results elaborated in Table 7.3 and 7.4, where identification and verification performances of WOCOR_4 and its dynamic coefficients with both GMM and HMM modeling are compared. As expected, modeling the long-term temporal variation with HMM significantly reduces the recognition errors. The dynamic feature, though performs not as good as the static feature, provides a certain degree of speaker discrimination

capability. However, in this experiment, fusion of static and dynamic features (that is, concatenate the two 24 dimensional feature parameters together to form a 48 dimensional feature set) does not perform better than using only the static feature in both GMM and HMM modeling. A better information fusion scheme, e.g. using different weights for combination of static and dynamic scores in decoding as proposed in [116], or with score level fusion as proposed in Chapter 5 and 6, is necessary to obtain better performance. Furthermore, the dynamic feature has been demonstrated to be more robust to noise distortions than the static features [116]. It could therefore be useful for robust speaker recognition in mismatched conditions.

Table 7.3: EERs (in %) with static and dynamic features and GMM/HMM modeling

Feature	$WOCOR_4$	Dynamic	$WOCOR_4$_D
GMM	7.39	9.66	8.44
HMM	4.73	7.50	5.33

Table 7.4: IDERs (in %) with static and dynamic features and GMM/HMM modeling

Feature	$WOCOR_4$	Dynamic	$WOCOR_4$_D
GMM	23.5	29.3	23.2
HMM	13.5	21.6	15.3

7.3 Comparison of Different Vocal Source Features for Speaker Recognition

The contribution of vocal source excitation on speaker characterization has been acknowledged for more than 40 years. Miller and Mathews in their work in 1963 showed that the glottal waveform has more variation among different speakers

than among different utterances of the same speaker [72]. From then on, many efforts have been devoted to exploit efficient vocal source features for speaker recognition. The possible candidates for vocal source features include pitch, harmonics, cepstrum and the glottal modelling parameters, as addressed in Section 3.3.2.

Among all these source features, pitch for speaker recognition has been most widely studied. As one of the most important acoustic features characterizing the vocal cords phonation, pitch carries rich speaker information and has been demonstrated to be useful for speaker discrimination. However, pitch value generally has large intra-speaker variations including the variation within an utterance and that in different utterances spoken in different time by the same speaker. The large intra-speaker variation dramatically degrades its speaker discrimination ability, especially when the population size is large. Another drawback of pitch is that although pitch itself is very robust to noise distortions, accurate pitch tracking is most of time very difficult due to the sub-multiple and multiple errors, especially in noisy speech [104].

The harmonic features including the harmonic richness and the degree of inter-harmonics noise have been shown to be good parameters to characterize different voice types [25]. However, accurate harmonic amplitude detection is also very difficult due to the impact of formant modulation and the restriction of frequency resolution in Fourier analysis. We have developed a harmonic feature extraction technique with linear prediction and pitch prediction analyses. The harmonics features, i.e. HRF (Eqt. 3.17) and NHR (Eqt. 3.18) demonstrated a certain degree of speaker discrimination power [120]. Similar to the pitch feature, the HRF and NHR parameters are also very vulnerable to noise distortions.

The cepstrum of the LP residual signal is a noise like signal with a bursts at quefrency of about one pitch period. Therefore, it also carries pitch related information. As to the glottal pulse modeling parameters, although it is claimed to be useful for speaker recognition, it is difficult to be applied for real-time applications as addressed in Section 3.3.2.

As known, the speaker-specific vocal source information resides on both time and frequency domain. However, none of these features represents such time-frequency properties. The short-term and long-term pitch variations, which is a very important speaker characteristics, are not fully represented. This thesis is aimed to exploit the speaker-specific time-frequency characteristics embedded in the vocal source excitation. The previous experiments have demonstrated that our feature sets, HOCOR_α and WOCOR_M, are effective in capturing such kind of vocal source characteristics and they greatly improve the recognition accuracy in both matched and mismatched conditions.

An advantage of WOCOR parameters is that it is very efficient in capturing temporal variations between consecutive pitch cycle due to the pitch-synchronous analysis, which makes it very successful in characterizing speakers with large degree of pitch variation. The disadvantage of WOCOR parameters is that its performance relies on the accuracy of pitch tracking. An error in pitch estimation could result in mis-detected pitch epochs, and thus the generated WOCOR parameters can not capture the desired spectro-temporal properties correctly. We have tested several pitch tracking methods including the cepstral analysis [74], pitch prediction adopted in CELP coding [65], and the RAPT algorithm [104], with increasing accuracy and also increasing computational complexity. As illustrated in Table 7.5. the speaker identification accuracy increases as the pitch tracking accuracy increases.

Table 7.5: Impact of pitch tracking accuracy in speaker identification performance

Pitch tracking methods	Cepstral analysis	pitch prediction	RAPT
ID rate (%)	68	74	77

The HOCOR parameters, on the other hand, are generated without pitch-synchronous analysis. Therefore, the temporal variation between consecutive pitch cycles can not be explicitly represented. Nevertheless, the time-frequency analysis makes it still effective in capturing the spectro-temporal characteristics of every segment of the analyzed LP residual signal. In particular, the

rectangular basis function of Haar transform is very effective in detecting the pitch epochs from the noisy signal. What's more, the low computational complexity also makes HOCOR a competitive candidate for vocal source feature in real-world applications.

The robustness of the proposed vocal source features have been partially verified by the robust speaker recognition experiments described in Section 7.1. Further exploration on this topic and robust feature selection will be carried out in the next stage.

To compare the effectiveness of these vocal source features in speaker recognition, their performances in speaker identification are listed in Table 7.6. It is clear that $HOCOR_3$ and $WOCOR_4$ perform much better than pitch, harmonics and cepstrum. As to the glottal modeling parameters (GMP), we did not replicate the experiments therefore cannot compare their results fairly. According to the original work [78], its identification rate with TIMIT database is about 69% compared with the 92% identification rate with the 14 dimensional static LPCC parameters. Its performance in matched conditions is comparable to our features. However, its performance on NTIMIT database greatly degrades mainly due to the errors in estimating the glottal flow waveform from noisy speech.

Table 7.6: Comparison of the identification rate of different vocal source features

Feature	Pitch	Harmonics		Cepst.	GMP	T-F Features	
		HRF	NHR			$HOCOR_3$	$WOCOR_4$
ID Rate (%).	18.5	38.3	17.4	20.0	NA	60.3	76.5

7.4 Other considerations

7.4.1 Training the UBM and GMM using different database

In Chapter 5 and 6, the UBM-GMM scheme was adopted for training the universal background model and the target models, where we used the same speech data for training the UBM and adapting the GMMs. However, in real-world applications, the amount of available training data is always restricted and generally there is not enough data to train a UBM with large number of Gaussian mixtures as what we have done in the previous experiments. In this case, as described in Section 2.2, one can train a UBM using another publicly available database with the same language as that used for adapting the target GMMs. And the two database should have similar data collection conditions.

In this section, instead of training the UBM and GMM using the same database, an existing database CUCALL is used for training the UBM. This database was originally developed for Cantonese speech recognition applications [57]. We use the digit-subset of the CUCALL database, which contains speech data from 339 male speakers, each speaker has several (less than 10) utterances consisting of 16 Cantonese digits. The speech data were collected with different handsets over fixed-line telephone channels. To reduce the effect of handset and channel distortions, CMN is applied in MFCC feature extraction.

Figure 7.3 illustrates the speaker verification performances of WOCOR_4, MFCCs and the fusion of these two features with two UBMs. The UBM0 is trained using the same data as that for GMM adaptation. The UBM1 is trained using the CUCALL database. Similar to what we have observed in the performances with UBM0, with UBM1, WOCOR_4 performs much worse than MFCCs, and fusion of WOCOR_4 and MFCCs obtains better performance than that using MFCCs only. It is also clear from the figure that the systems with UBM0 perform better than the corresponding ones with UBM1. This is because:

- UBM1 does not contain the speaker information since it is trained with a

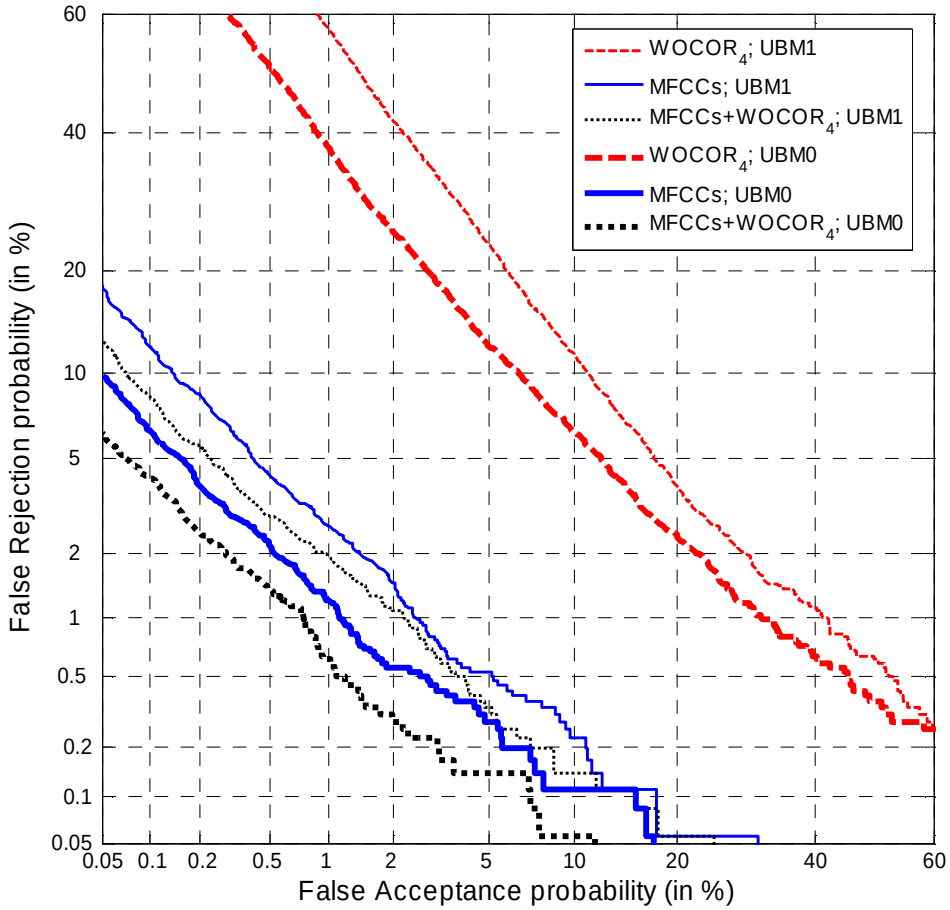


Figure 7.3: Comparison of verification performances with different UBMs. UBM0: the training data is the same as for GMM; UBM1: the training data is different from that for GMM

different database.

- the CMN processing reduces the handset and channel distortion, however it might also remove some speaker-specific information. That is, the individual personality in speech production can be partially characterized by a time-invariant frequency modulation system.

7.4.2 Comparison of the conventional MFCC parameters and that derived in this thesis

In this thesis, to maximize the complementarity of vocal source and vocal tract features, the MFCC parameters are generated from Fourier spectrum derived from the LP coefficients via Eqt. 5.1. The MFCC parameters are different from the conventional ones in which the Fourier spectrum is calculated by applying Fourier transform on the speech signal. Section 5.1.2 has addressed the advantages and disadvantages of these two MFCC parameters. Table 7.7 compares the speaker verification EERs of these two MFCC parameters.

Table 7.7: Comparison of the EERs (in %) with two MFCCs derived with different methods

Gender	Conventional MFCC		MFCC in this thesis	
	MFCCs	Info-fusion	MFCCs	Info-fusion
Male	8.32	6.91	8.47	7.40
Female	8.60	7.50	6.74	5.80

As illustrated, for male speakers, the conventional MFCCs performs slightly better than the proposed MFCCs (8.32% vs. 8.47%). Fusion of source feature and the conventional MFCCs also gives better verification results (6.91% vs. 7.40%). However, as to the female speakers, our proposed MFCCs significantly outperforms the conventional one (about 2% absolute EER reduction).

As known, the conventional MFCC parameters contains both vocal tract and vocal source information. For the high pitched female speakers, the formant structure, particularly the first formant, is more easily distorted by F0 and first some harmonics. When there is a large variation of F0 in training and testing speech, the performance of the conventional MFCCs will degrade dramatically. In this case, eliminating the effect of F0 and harmonics as what we do on deriving our MFCC parameters will benefit the recognition system.

From the EERs of WOCOR_4 on male and female database as illustrated in Table 7.1 and 7.2, we can see the performance of the vocal source feature on

female speakers is not as good as that on the male speakers (23.11% vs. 19.65% in EER). The less discriminative vocal source information in high pitched female speaker makes the conventional MFCC parameters less effective in female speaker recognition. Our source-tract information fusion scheme is therefore particularly useful for female speakers.

7.4.3 Expected performance in very large population size

The population size is an important factor affecting the speaker recognition performance and should be taken into consideration in the development of speaker recognition system. Generally, as the population size increases, it is more difficult to disambiguate one speaker from the others. Therefore the recognition accuracy declines with large number of speakers. In particular, a speaker identification system needs to make an N (the population size) to 1 decision. The identification accuracy degrades dramatically with very large N. Therefore, it is necessary to analyze the performance of a recognition system with different population sizes.

Table 7.8: SI performance (IDER, in %) with various population size

Population size	$WOCOR_4$	MFCCs	Info-fusion
15	31.7	15.8	11.5
25	39.2	14.4	10.3
50	45.9	20.8	16.6
74	44.7	24.7	21.5

Table 7.8 illustrates the speaker identification performance on the male subset of the NIST 2001 SRE database as described in Section 7.1 with different population sizes. It is clear that the population size has a great impact on the identification performances of both $WOCOR_4$ and MFCCs features. For example, as N increases from 15 to 74, IDER increases from 15.8% to 24.7% for MFCCs, and from 31.7% to 44.7% for $WOCOR_4$. Nevertheless, the fusion of

these two source and tract features reduces the identification error in each case, although its performance also degrades as N increases.

Due to the lack of data, we have not evaluated the performances of different features in very large population size, e.g. N>1000. Even though, it is expected that our proposed source-tract information fusion scheme will still improve the system performance with very large N.

Chapter 8

Conclusions and Future Work

8.1 Conclusions

We have investigated the feasibility of using the complementary vocal source and vocal tract information to achieve a better speaker recognition performance. Conventional speaker recognition systems typically employ the the vocal tract features with which the speaker-specific information associated with the vocal tract articulation is represented. We have addressed the speaker-specific characteristics in vocal cords phonation and proposed effective feature extraction techniques to generate the speaker discriminative vocal source features. We have also exploited efficient information fusion scheme to take the full advantage of the complementarity of the vocal source and vocal tract features for speaker recognition.

This research was motivated by the improved speaker recognition performance by fusion of different information sources from the speech signal. The separation of information sources was achieved by the linear predictive inverse filtering, which results in two orthogonal information components, i.e. the LP coefficients and the LP residual signal. The complementarity of vocal source and vocal tract features derived from these two orthogonal components is therefore maximized. The LP residual signal, though not giving the true glottal pulse waveform, is a good representative of the source excitation. Besides the source excitation related characteristics, other information that has not been covered

by the LP coefficients, e.g. zeros due to nasal voice, the source-trance interaction, are also retained in the residual signal, in which rich speaker-specific information is embedded.

The speaker-specific information within the LP residual signal resides on both time and frequency domain. To extract efficient speaker discriminative source features, two time-frequency transforms, i.e. Haar transform and wavelet transform, have been applied for analyzing the LP residual signal. The resulting features, HOCOR and WOCOR, effectively capture the speaker-specific spectro-temporal characteristics of the residual signal. Particularly, by applying pitch-synchronous wavelet transform on every two pitch periods of residual signal, the WOCOR parameters is capable of capturing the pitch-related low frequency properties and the high frequency information associated with the pitch epochs, as well as their temporal variations within a pitch period and over consecutive periods. The proposed vocal source features are believed, and have been verified to be robust to noise and channel distortions since the pitch period and pitch epochs are relative resistant to these distortions.

The complementary contributions of vocal source and vocal tract features in speaker recognition have been evaluated on speech database with both matched and mismatched training and testing conditions. Experimental results showed that although the performances of the vocal source features are not convincing compared with that of the vocal tract features. Fusion of these two kinds of features did improve the overall speaker recognition performance. Taking the WOCOR_4 parameters as an example, with matched conditions, the identification error rate was reduced from 1.44% with only MFCCs to 0.94% with both WOCOR_4 and MFCCs, a relative improvement of 34.5%. The verification equal error rate was reduced form 1.06% to 0.81%, a relative improvement of 23.6%. Fusion of these two features also improved the robustness of the system in mismatched condition. For example, The identification error rate was reduced from 24.71% with MFCCs only to 20.94 with WOCOR_4 and MFCCs. The verification EER was also reduced from 8.74% to 7.40%.

To maximize the benefit through information fusion so as to further improve

the system performance, we developed two optimized weighting schemes with discriminative analysis. For speaker identification, instead of using a pre-trained fixed fusion weight, a varying weighting scheme is developed. Analysis of identification results shows that generally a correct identification is associated with a larger discrimination power (D, see Eqt. 6.1) than the incorrect identification. Therefore, a confidence measure was online derived based on the discrimination ratio between the two features in each identification trial. Information fusion with confidence measure combines better weighted scores given by the two features and avoids the errors introduced by incorporating source feature with fixed weight. Identification error rate was further reduced from 0.94% to 0.78% in matched conditions and from 20.94% to 19.41% in mismatched conditions.

Due to the diversity of vocal system configurations in speech production, different sounds hold different source and tract discrimination power. We have analyzed this kind of text-dependent discrimination power for the 10 Cantonese digits. The analysis results showed that there are significant variation of source-tract speaker discrimination ratio across the 10 digits. Therefore, a digit-dependent source-tract weighting scheme was developed for speaker verification. The fusion weights for each digit were trained by performing speaker verification with only one digit. Information fusion with such digit-dependent weights further improved the verification performance.

To summary, the major contribution of this research includes:

- The role of vocal source characteristics for speaker recognition was systematically discussed. The usefulness of vocal source feature for supplementing the vocal tract feature in speaker recognition has been confirmed.

- A novel method for feature extraction from vocal source signal was proposed. The new features are effective in capturing the speaker-specific spectro-temporal vocal source characteristics.

- Efficient information fusion techniques were developed for speaker recognition. Fusion of the complementary vocal source and vocal tract features improves the recognition performance and the system robustness as well.

8.2 Perspectives of Future Work

We have demonstrated the effectiveness of using vocal source features to supplement vocal tract features for improved speaker recognition performances. However, the performance of the proposed vocal source features are still far from reliable compared with that by conventional vocal tract features. It will be therefore appealing to exploit more efficient vocal source feature extraction techniques, which currently confronts two major difficulties. The first difficulty comes from the lack of a reliable method for estimating the true vocal source signal. The second difficulty is due to that the existing vocal source models can not fully account for the irregular source excitation styles in natural speech which is a very important acoustic cues for speaker discrimination. Further researches in these problems will be, though very difficult, highly demanded and will benefit not only speaker recognition, but also many other speech applications such as speech synthesis, speaking emotion recognition, etc.

As the dominant mode of information exchange for human beings and for human-machine communications, speech signal contains a great diversity of speaker information. The successful applications of source and tract information fusion for robust speaker recognition demonstrated in this thesis suggests that fusion of multiple information sources merits further research. Future work in this direction can extend this method to fuse more sources of information, such as the high level information as described in Table 2.1. The efficient information retrieval and fusion techniques have long been the goal of many researches. Most of these investigations, including our work, are still in the preliminary stages. Further exploration is definitely necessary.

Bibliography

[1] A. Acero. *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Kluwer Academic Pub., Dordrecht, 1992.

[2] P. Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118, 1992.

[3] T. A. Ananthapadmanabha and B. Yegnanarayana. Epoch extraction from linear prediction residual for identification of closed glottis interval. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(4):309–319, 1979.

[4] T. V. Ananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *Speech Communication*, 1(3-4):167–184, 1982.

[5] F. J. Anthony and S.-S. Joseph. *Physiology of the Ear*. San Diego, CA : Singular/Thomson Learning, 2001.

[6] B. S. Atal. Automatic speaker recognition based on pitch contours. *J. Acoust. Soc. Am.*, 52:1687–1697, 1972.

[7] B. S. Atal. Efectiveness of linear prediction characteristics of the speech wave for automatic speaker identication and verication. *J. Acoust. Soc. Am.*, 55(6):1304–1312, 1974.

[8] B. S. Atal. Automatic recognition of speakers from their voices. *Proc. IEEE*, 64:460–475, 1976.

[9] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.

[10] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *Ann. Mathemat. Stat.*, 37:1554–1563, 1966.

[11] K. G. Beauchamp. *Walsh Functions and Their Applications*. London: Academic Press, 1975.

[12] C. M. Bishop. *Neural Networks for Pattern Recognition*. New York : Oxford University Press, 1995.

[13] R. H. Bolt, F. S. Cooper, E. E. David, P. B. Denes, J. M. Pickett, and K. N. Stevens. Identification of a speaker by speech spectrograms: How do scientists view its reliability for use as legal evidence? *Science*, 166:338–343, 1969.

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

[15] H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach*. Boston : Kluwer Academic Publishers, 1994.

[16] D. M. Brookes and D. S. F. Chan. Speaker characteristics from a glottal airflow model using robust inverse filtering. *Proc. Inst. of Acoustics*, 16(5):501–508, 1994.

[17] J. P. Campbell. Testing with the YOHO CD-ROM voice verification corpus. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 341–344, 1995.

[18] J. P. Campbell. Speaker recognition: a tutorial. *Proc. IEEE*, 85(9):1437–1462, 1997.

[19] W. M. Campbell and K. T. Assaleh. Polynomial classifier techniques for speaker verification. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 321–324, 1999.

[20] W. M. Campbell and K. T. Assaleh. Polynomial classifier techniques for speaker verification. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 321–324, 1999.

[21] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek. High-level speaker verification with support vector machines. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 73–76, 2004.

[22] Y. T. Chan. *Wavelet Basics*. Kluwer Academic Publishers Group, 1996.

[23] C. Che and Q. Lin. Speaker recognition using HMM with experiments on the yoho database. In *Proc. Eurospeech*, pages 625–628, 1995.

[24] D. G. Childers, D. M. Hicks, G. P. Moore, L. Eskenazi, and A. L. Lalwani. Electroglottography and vocal fold physiology. *Journal of speech and hearing research*, 33:245–254, 1990.

[25] D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *J. Acoust. Soc. Am.*, 90:1991, 2394-2410.

[26] D. G. Childers and K. Wu. Quality of speech produced by analysis-synthesis. *Speech Communication*.

[27] J. Colombi, D. Ruck, S. Rogers, M. Oxley, and T. Anderson. Cohort selection and word grammer effects for speaker recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 85–88, 1996.

[28] S. B. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust., Speech, Signal Processing*, 28(4):357–366, 1980.

[29] V. Digalakis, D. Ritischev, and L. Neumeyer. Speaker adaptation using constrained reestimation of gaussian mixtures. *IEEE Trans. Speech Audio Processing*, 3(5):357–366, 1995.

[30] G. R. Doddington. Speaker recognition-identifying people by their voices. *Proc. IEEE*, 73:1651–1664, 1985.

[31] J. Eatock and J. S. Mason. Automatically focusing on good discriminating speech segments in speaker recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 133–136, 1990.

[32] J. P. Egan. *Signal Detection Theory and ROC Analysis*. New York: Academic Press, 1975.

[33] L. Eskenazi, D. G. Childers, and D. M. Hicks. Acoustic correlates of vocal quality. *J. Speech Hear Res.*, 33(2):298–306, 1990.

[34] H. Ezzaidi, J. Rouat, and D. O'Shaughnessy. Towards combining pitch and MFCC for speaker identification systems. In *Proc. Eurospeech*, pages 2825–2828, 2001.

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

[36] G. Fant. Glottal flow: models and interaction. *J. Phonet.*, 14:393–399, 1986.

[37] A. J. Fourcin and E. Abberton. First applications of a new laryngograph. *Med. Biol. Illus.*, 21:172–182, 1971.

[38] S. Furui. Cepstral analysis technique for automatic speaker verification. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-29(2):254 – 272, 1981.

[39] S. Furui. Speaker independent isolated word recogntion using dynamic features of speech spectrum. *IEEE Trans. Acoust., Speech, Signal Processing*, 34, 1986.

[40] S. Furui. Recent advances in speaker recognition. *Pattern Recognition Letters*, 18:859–872, 1997.

[41] S. Furui. Comparison of speaker recognition methods using statistical features and dynamic features. *IEEE Trans. Acoust., Speech, Signal Processing*, 29(3):1981, 342-350.

[42] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Boston : Kluwer Academic Publishers, 1992.

[43] H. Gish, M. Krasner, W. Russell, and J. Wolf. Methods and experiments for text-independent speaker recognition over telephone channels. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 865–868, 1986.

[44] H. Hermansky and N. Morgan. Rasta processing of speech. *IEEE Trans. Speech Audio Processing*, 2(4):578–589, 1994.

[45] H. Hermansky, N. Morgan, and H. G. Hirsch. Recognition of speech in additive and convolution noise based on RASTA spectral processing. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 83–86, 1995.

[46] A. L. Higgins and R. E. Wohlford. A new method for text-independent speake recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 869–872, 1986.

[47] B. Imperl, Z. Kacic, and B. Horvat. A study of harmonic features for speaker recognition. *Speech Communication*, 22(4):385–402, 1997.

[48] Q. Jin and A. Waibel. Application of LDA to speaker recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2000.

[49] I. T. Jolliffe. *Principal Component Analysis.* New York : Springer-Verlag, 2002.

[50] B.-H. Juang, W. Chou, and C.-H. Lee. Minimum classification error rate methods for speech recognition. *IEEE Trans. Speech Audio Processing*, 5(3):257–265, 1997.

[51] L. G. Kersta. Voiceprint identification. *Nature*, 196:1253–1257, 1962.

[52] D. S. Kim, S. Y. Lee, and R. M. Kil. Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Trans. Speech Audio Processing*, 7:55–69, 1999.

[53] S. Kim, T. Eriksson, H. G. Kang, and D. H. Youn. A pitch synchronous feature extraction method for speaker recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 405–408, 2004.

[54] D. H. Klatt and L. C. Klatt. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Am.*, 87:820–857, 1990.

[55] S. Y. Kung, M. W. Mak, and S. H. Lin. *Biometric Authentication: A machine Learning Approach*. Prentice Hall, 2004.

[56] Q. Le and S. Bengio. Client dependent GMM-SVM models for speaker verification. *Lecture Notes in Computer Science*, 2714:443–451, 2003.

[57] T. Lee, W. K. Lo, P. Ching, and H. Meng. Spoken language resources for cantonese speech processing. *Speech Communication*, 36:327–342, 2002.

[58] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(4):806–814, 1995.

[59] P. Lieberman. *Human Language and Our Reptilian Brain: the Subcortical Bases of Speech, Syntax, and Thought*. Harvard University Press, 2000.

[60] Y. Linde, A. Buzo, and R. M. Gray. An algorithm for vector quantizer design. *IEEE Trans. Comm.*, C-28(1):84–95, 1980.

[61] M. W. Mak, W. Allen, and G. Sexton. Speaker identification using multilayer perceptrons and radial basis function networks. *Neurocomputing*, 6:1994, 99-117.

[62] J. Makhoul. Linear prediction: A tutorial review. *Proc. IEEE*, 63(4):561 – 580, 1975.

[63] J. D. Markel. *Linear Prediction of Speech*. New York : Springer-Verlag, 1976.

[64] J. D. Markel and S. B. Davis. Text-independent speaker recognition from a large linguistically unconstrained time-spaced data base. *IEEE Trans. Acoust., Speech, Signal Processing*, 27:74–82, 1979.

[65] J. S. Marques, I. M. Trancoso, J. M. Tribolet, and L. B. Almedia. Improved pitch prediction with fractional delays in celp coding. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 665–668, 1990.

[66] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech*, pages 1895–1898, 1997.

[67] A. Martin and M. Przybocki. The NIST 1999 speaker recognition evaluation: An overview. *Digital Signal Processing*, 10:2000, 1-18.

[68] T. Matsui and S. Furui. Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. *Speech Communication*, 17:1995, 109-116.

[69] T. Matsui and S. Furui. Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 157–160, 1992.

[70] G. J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition*. New York : Wiley,, 1992.

[71] H. M. Meng, P. C. Ching, T.-Y. Fung, Y.-C. Li, M.-C. Ho, C.-K. Keung, W.-K. Lo, T.-H. Lo, K.-F. Low, and K.-C. Siu. The 'author once, present

anywhere' (AOPA) software platform. In *Proc. of the 2003 Hong Kong International Computer Conference*, 2003.

[72] J. E. Miller and M. V. Mathews. Investigation of the glottal waveshape by automatic inverse filtering. *J. Acoust. Soc. Am.*, 35:1876, 1963.

[73] J. Naik. Speaker verification: a tutorial. *IEEE Commun. Mag.*, 28:42–48, 1990.

[74] A. M. Noll. Cepstrum pitch determination. *J. Acoust. Soc. Am.*, 41:293–309, 1967.

[75] Y. Normandin. Maximum mutual information estimation of hidden markov models. In C.-H. Lee, F. K. Soong, and K. Paliwal, editors, *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer, 1996.

[76] J. Oglesby and J. Mason. Radial basis function networks for speaker recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 393–396, 1991.

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

[78] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Trans. Speech Audio Processing*, 7(5):569–585, 1999.

[79] C. Qin. Verbal information verfication for high-performance speaker authentication. Master's thesis, The Chinese University of Hong Kong, 2005.

[80] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice Hall, 2001.

[81] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[82] L. R. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall, 1993.

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

[84] M. G. Rahim and B. H. Juang. Signal bias removal by maximum likelihood estimation for robust telephone speech recognition. *IEEE Trans. Speech Audio Processing*, 4(1):19–30, 1996.

[85] D. A. Reynolds. Comparison of background normalization methods for text-independent speaker verification. In *Proc. Eurospeech*, pages 963–966, 1997.

[86] D. A. Reynolds. An overview of automatic speaker recognition technology. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 4072–4075, 2002.

[87] D. A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones1, and B. Xiang. The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 784–787, 2003.

[88] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 10(1-3):19–41, 2000.

[89] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary, and B. A. Carlson. The effects of telephone transmission degradations on speaker recognition performance. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*.

[90] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds. Integrated models of signal and background with application to speaker identication in noise. *IEEE Trans. Speech Audio Processing*, 2(2):245–257, 1994.

[91] A. E. Rosenberg. Automatic speaker verification: a review. *Proc. IEEE*, 64:475–487, 1976.

[92] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong. The use of cohort normalized scores for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 599–602, 1992.

[93] A. E. Rosenberg and F. K. Soong. Recent research in automatic speaker recognition. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*. New York: Marcel Dekker, 1992.

[94] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, 26(1):43–49, 1978.

[95] A. Sankar and C. H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. Speech Audio Processing*, 4(3):190–202, 1996.

[96] M. Schmidt and H. Gish. Speaker identification via support vector machines. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 105–108, 1996.

[97] A. Schmidt-Nielsen and T. H. Crystal. Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. *Digital Signal Processing*, 10(1-2):249–266, 2000.

[98] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 3189–3192, 1998.

[99] M. K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg. A lognormal tied mixture model of pitch for prosody based speaker recognition. In *Proc. Eurospeech*, pages 1391–1394, 1997.

[100] F. K. Soong and A. E. Rosenberg. On the use of instantaneous and transitional spectral information in speaker recognition. *IEEE Trans. Acoust., Speech, Signal Processing*, 36(6):871–879, 1988.

[101] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang. A vector quantization approach to speaker recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 387–390, 1985.

[102] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods. Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material. *J. Acoust. Soc. Am.*, 44:1596–1607, 1968.

[103] G. Strang and T. Nguyen. *Wavelets and Filter Banks*. Wellesley-Cambridge Press, 1996.

[104] D. Talkin. A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*. Elsevier, 1995.

[105] P. Thevenaz and H. Hugli. Usefulness of the LPC residue in text-independent speaker verification. *Speech Communication*, 17(1-2):145–157, 1995.

[106] D. W. Thomas. Burst detection using the Haar spectrum. In *Proc. Theory and Application of Walsh and Other Non-sinusoidal Functions*, 1973.

[107] O. Thyes, R. Kuhn, P. Nguyen, and J.-C. Junqua. Speaker identification and verification using eigenvoices. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, 2000.

[108] V. N. Vapnik. *Statistical Learning Theory*. New York: Wiley, 1998.

[109] D. E. Veeneman and S. L. Bement. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE Trans. Acoust., Speech, Signal Processing*, 33(2):369–377, 1985.

[110] V. Wan. *Speaker Verification using Support Vector Machines*. PhD thesis, University of Sheffield, 2003.

[111] J. Wayman. The functions of biometric identification devices. *San Jose, CA: National Biometric Test Center*, 2000, 2000.

[112] Web Resource:. The NIST 2001 speaker ID evaluation protocol. *http://www.nist.gov/speech/tests/spk/2001/index.htm*.

[113] Web Resource:. The NIST speaker recognition evaluations. *http://www.nist.gov/speech/tests/spk/*.

[114] D. B. Webster. *Neuroscience of Communication*. San Diego, Calif. : Singular Pub. Group, 1995.

[115] D. J. Wong, J. D. Markel, and A. H. Gray. Least squares glottal inverse filtering from the acoustic speech wave. *IEEE Trans. Acoust., Speech, Signal Processing*, 27(8):350–355, 1979.

[116] C. Yang, F. K. Soong, and T. Lee. Static and dynamic spectral features: Their noise robustness and optimal weights for ASR. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 241–244, 2005.

[117] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK Book*. Cambridge University Engineering Department, 2001.

[118] W. M. Yu, M. W. Mak, and S. Y. Kung. Speaker verification from coded telephone speech using stochastic feature transformation and handset identification. In *Proc. Pacific-Rim Conference on Multimedia*, pages 598–606, 2002.

[119] W. M. Yu, M. W. Mak, C. H. Sit, and S. Y. Kung. Speaker verification based on G.729 and G.723 coder parameters and handset mismatch compensation. In *Proc. Eurospeech*, pages 1681–1684, 2003.

[120] N. H. Zheng. Pitch prediction for glottal spectrum estimation with applications in speaker recognition. unpublished.

[121] N. H. Zheng and P. C. Ching. Using Haar transformed vocal source information for automatic speaker recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 77–80, 2004.

[122] N. H. Zheng, P. C. Ching, and T. Lee. Time frequency analysis of vocal source signal for speaker recognition. In *Proc. Int. Conf. on Spoken Language Processing (ICSLP)*, pages 2333 –2336, 2004.

[123] N. H. Zheng, T. Lee, and P. C. Ching. Fusion of vocal source and vocal tract features for robust speaker identification. In *submitted to ICASSP 2006*.

[124] N. H. Zheng, T. Lee, and P. C. Ching. Comparative analysis of discrimination power of the vocal source and vocal tract features for speaker verification. In *Proc. National Conference on Man-Machine Speech Communication (China)*, 2005.

[125] N. H. Zheng, C. Qin, T. Lee, and P. C. Ching. CU2C: A dual-condition cantonese speech database for speaker recognition applications. In *Proc. Oriental-COCOSDA*, 2005.

[126] R. D. Zilca, J. Navratil, and N. Ramaswany. Depitch and the role of fundamental frequency in speaker recognition. In *Proc. IEEE Int. Conf. on Acoust., Speech, Signal Processing (ICASSP)*, pages 81–84, 2003.