Subband Spectral Features for Speaker Recognition

TAM Yuk Yin

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

July 2004

Abstract of thesis entitled:

Subband Spectral Features for

Speaker Recognition

Submitted by **Tam Yuk Yin**

for the degree of **Master of Philosophy**

in **Electronic Engineering**

at **The Chinese University of Hong Kong**

in **July 2004**.

Speaker recognition is an attractive field of research nowadays, which makes use of one of the most natural and the least obtrusive biometric measures. It refers to the use of a speaker's spoken words to identify his or her identity. In real applications, the bandwidth of speech data used is usually limited. This may be one of the factors that degrade the recognition performance. In this thesis, we focus on the speaker-dependent information in different frequency subbands.

The use of spectral envelope features extracted from narrowband NB (0 – 4kHz) and higher frequency band HB (4 – 8kHz) in text-dependent speaker identification (SID) is studied. It is shown that the importance of features from NB and HB in speaker recognition is text-dependent. Fusing features from these two bands based on their importance is investigated with two different approaches. They are model level fusion and feature level fusion.

We extend our study on subband spectral envelope features to utterance-level SID. The result of text-dependent SID indicated that some words are more reliable in discriminating speakers than the others. Based on this result, text-dependent weights used in linear combination of likelihood scores from individual word are designed. Recognition accuracy can be improved by this method.

With a better understanding on the contribution of features from NB and HB, we can fuse them together in a proper way so as to maximize the benefits from these two bands. Similarly, we can assign heavier weights to those words that show higher reliability in recognizing speakers for utterance-level SID. The approach of applying confidence weights in speaker recognition provides a possible way to improve the performance.

摘要

說話人識別是一個引人矚目的研究領域，它利用了最自然最簡單的生物鑒定術。本文將針對利用語音來識別說話人的身份。在實際的應用中，語音數據的帶寬通常是被限制的，這會導致識別性能的降低。在本文中我們重點研究說話人相關的子帶信息。

在文本相關的說話人鑒別實驗中，我們研究了從窄頻帶和高頻帶提取出來的頻譜包絡特徵參數。我們發現窄頻帶和高頻帶特徵參數的重要性是文本相關的。根據他們的重要性，我們可以把這兩種特徵參數融合在一起加以綜合利用。我們詳細討論了兩種融合的辦法。他們是特徵參數層次的融合和模型層次的融合。

我們把對子帶頻譜包絡特徵參數的研究拓展到語句層次的說話人鑑定實驗上。以文本相關的說話人鑑定實驗結果顯示某些文字區分說話人的性能比其他文字更加可靠。根據不同文字可靠性的表現，我們用文本相關的權重對不同文字的似然度得分進行綫性加權組合。用這種辦法我們提高了識別性能。

在更好的了解窄頻帶和高頻帶特徵參數的重要性之後，我們進而可以尋找到使性能提高最大化的辦法。類似的，在語句層次的說話人識別實驗中，我們給更可靠的語句加以更高的權重。這個施加可信度權重的辦法，給說話人鑑定性能的提高帶來了可能。

Acknowledgment

I would like to thank my supervisor, Prof. Tan Lee for his supervision and insightful advice throughout this research. He has given me lots of advices and suggestions on my thesis. I would also like to thank Prof. P.C. Ching and Prof. Y.T. Chan for their valuable suggestions.

Special thanks are given to Dr. F. K. Soong for fruitful discussion and encouragement.

I would like to thank all the colleagues and friends in DSP laboratory. Especially, I would like to thank Zhu Yu for providing speech recognizer for me to do forced alignment and having many helpful discussions with me. Also, thanks are given to Qian Yao, Yang Chen, Yvonne Lee, Yuan Meng, Zheng Neng Heng and Joyce Chan for their comments and suggestions on my work, and Arthur Luk for his technical support. Thanks are also due to my friends: Maggie Lee, Fred Wong, Jam Ku and Ryan Lai for their encouragement and support, especially during the time of writing thesis. I would also thank those people who helped me to record speech data. Without their help, this study cannot be completed.

Finally, I would like to express sincere gratitude to my parents and my sisters for their love and support, although they may not know what I have done in these two years.

Thank God for giving me strength to finish this work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In modern society, there are a wide variety of occasions requiring reliable identity recognition. Many commercial applications involve the process of user authentication so that services are provided only to authorized users. An example is the personal identification number (PINs) for automated teller machines (ATMs). Recently, authenticating humans by biometrics receives great attraction as it can provide convenience to users while maintaining a high degree of security. Biometrics authentication refers to the automatic recognition of individuals based on their physiological and/or behavioral characteristics [1]. Recognizing users by voice is one of the most commonly used techniques. It offers one of the most natural and the least obtrusive biometrics measures. What the user has to do is speaking a few words. In this research, we focus on the speaker-dependent information in different frequency subbands.

User authentication of speaker's voice is done either by confirming whether one is the claimed person or determining one's exact identity. A wide variety of occasions require user authentication so that only authorized users can access the provided services. Examples are secure access to buildings and computer systems, telephony-based transactions, ATMs and e-commerce.

Traditionally, user authentication is performed with knowledge-based techniques (e.g. passwords) or token-based techniques (e.g. smart cards and keys). However, both

types of systems have their drawbacks and limitations. Passwords may be stolen by imposters. Typically, for the convenience of memorizing, people set their passwords based on words or digits that they can remember easily (e.g. birthday of family members). This increases the risk because the intruder can easily conjecture the passwords if they can somehow obtain the users personal information such as birthday. Therefore, such system is not secure enough. Using smart cards and keys in access control is not reliable enough as they can be duplicated, lost or stolen.

1.1. Biometrics for User Authentication

Biometrics-based recognition refers to the automatic recognition of individuals based on their physiological and/or behavioral characteristics [1]. Examples of biometrics include face, voice, iris and fingerprint. Among various types of biometrics, fingerprint has been accepted as an effective way of identity verification for a long time. In recent years, many applications start to use other biometrics. For example, the Amsterdam Airport Schiphol began to use iris recognition in the border passage system in October 2001 (the Schiphol Privium scheme) [2].

In general, biometrics can be divided into two types: physiological and behavioral (see Figure 1-1). Physiological characteristics, include fingerprint and iris pattern, are physical features and they are basically invariant without trauma to the individual. On the other hand, behavioral characteristics, including signature and voice, may not remain invariant. They are influenced by physical and emotional conditions.

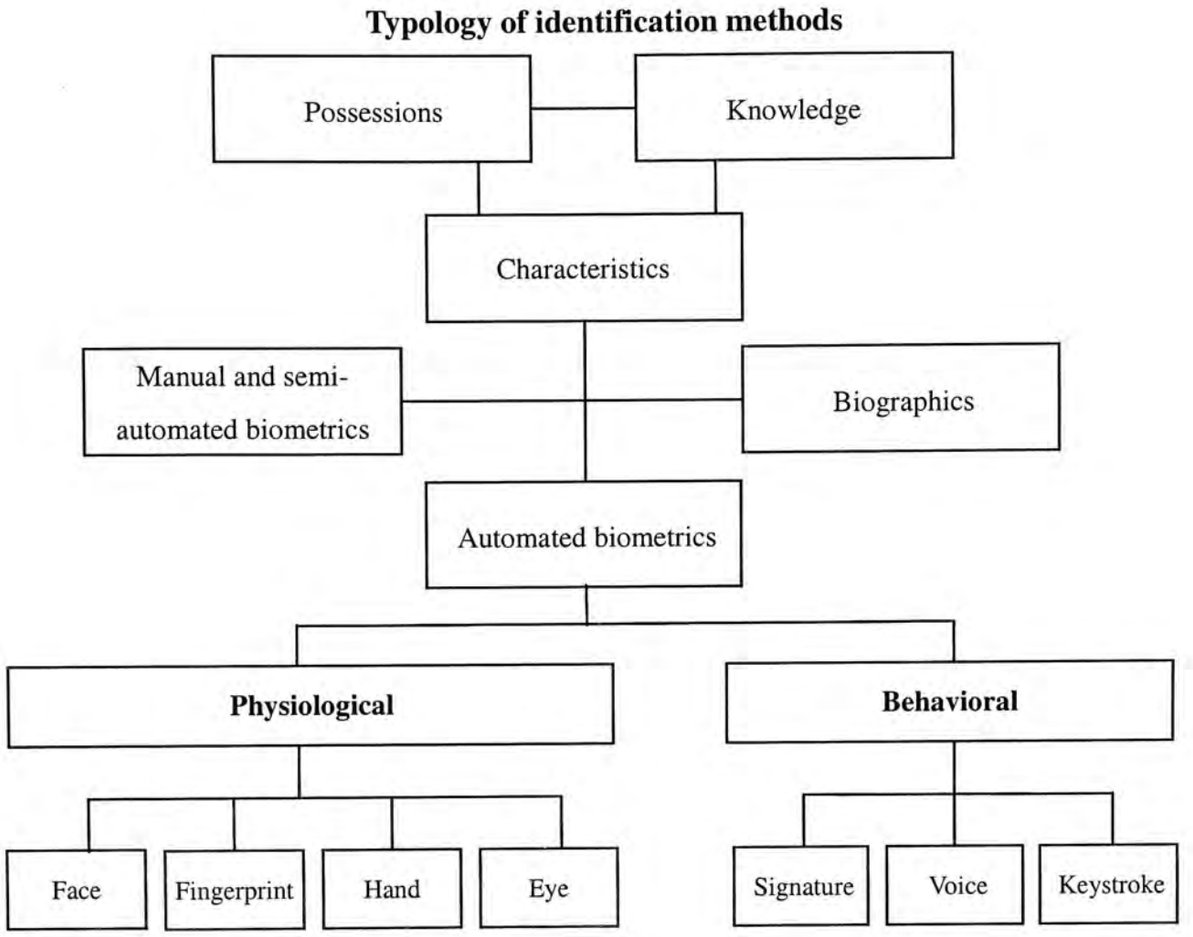


Figure 1-1 Typology of identification methods with examples for the biometrics type [3]

Strengths and Weaknesses of using Biometrics-based Recognition

By using biometrics-based recognition, a user is no longer recognized by what he /she remembers (i.e. password) or what he/she possesses (i.e. smart cards). Instead, user authentication is done by recognizing the user's biometrics characteristics. Compared with the traditional knowledge-based and token-based techniques, biometrics system can provide more convenience to users and a sufficiently high degree of security at the same time. This is because biometrics cannot be forgotten. Users no longer need to remember long and complicated passwords. Also, biometrics are supposed to be permanent and unchangeable. It is relatively difficult to copy, steal or forge biometrics with as much ease as passwords and keys. Therefore, it can maintain a relatively high

degree of security.

Nevertheless, there are many applications that choose to use traditional techniques for user authentication. Indeed, the accuracy of biometrics-based systems with current technologies still needs improvement [1]. For example, when a user's voice is seriously altered due to sickness or verification is processed under noisy environment, it will degrade the recognition performance. These types of problems concern the need of more research study on the technology area for robustness.

General Considerations on Using Biometrics

In order to be used in user authentication, the physiological or behavioral characteristics must meet the following requirements [1]:

1. *Universality*: obtainable from every individual.
2. *Distinctiveness*: distinctive between different people.
3. *Permanence*: sufficiently invariant over a period of time.
4. *Collectability*: quantitatively measurable.

There are also some other considerations on choosing the type of biometrics for practical applications. They include the accuracy that the system needs and the availability of equipments (e.g. sensor used to collect the biometrics data) so as to attain the required level of performance. Another consideration is the willingness of users to provide a particular type of biometrics feature to use. Lastly, the required level of security is also needed to consider. This is because some types of biometrics are easier to be forged by imposters than the others (e.g. signature versus fingerprint).

Advantages and Disadvantages of Various Biometrics

Each type of biometrics has its own advantages and disadvantages. The choice is application-dependent. Among the various types of biometrics, fingerprint, iris, face and voice are most commonly used. Their strengths and weaknesses are listed in Table 1-1.

Biometrics	Advantages	Disadvantages
Fingerprint	♦ Highly distinctive. Different for different people, even different on each finger of the same person.	♦ Not applicable to some people who always have a number of cuts and bruises on their fingers due to their occupation.
Iris	♦ Highly distinctive. ♦ Difficult to change the texture of iris artificially.	♦ Not suitable to use for people with visual impairment. ♦ Inconvenient to use as it requires equipment to scan the iris pattern.
Face	♦ Applicable to all people. ♦ Easy to capture face images by camera. ♦ Non-intrusive and convenient to use.	♦ Not highly distinctive between different people (e.g. identical twins). ♦ Impose a number of restrictions on how the facial images are obtained (e.g. in a fixed and simple background)
Voice	♦ Easy to collect speech data by microphone or telephone. ♦ Low equipment cost. ♦ Non-intrusive and convenient to use.	♦ Not applicable to people with speech impairment. ♦ Not permanent. A person's voice changes over time due to aging, medical conditions or emotional state. ♦ Not very distinctive between people.

Table 1-1 Comparison of several biometrics technologies

1.2. Voice-based User Authentication

Voice-based user authentication, or named as speaker recognition, is the focus of this research. Using voice in user authentication is one of the most natural and unobtrusive measures, compared with those using iris and fingerprint. Users are more willing to adopt this technique of authentication. Regarding on the practical considerations, speaker recognition does not require special measuring equipment. Only microphone or telephone is needed. Compared with other biometrics, the equipment cost for speaker recognition is relatively low. It provides an economical means for user authentication. Therefore, speaker recognition is suitable in many different areas, especially for telephone-based applications.

Figure 1-2 describes the general operation of a speaker recognition system. Given an input utterance from an unknown speaker, the acquired speech signal is analyzed to extract features of the speaker's voice. Afterwards, the measured features are compared with the prototypes of a set of known speakers in the system. A decision is made through one of two possibilities. Either the system verifies the claimed speaker, or it identifies a person as one of the known speakers in the system. General overview of speaker recognition can be found in [4]-[7].

Speaker recognition has been studied for a long time. However, there are many factors that degrade the recognition performance in real applications. It is mainly related to the variability of speech signal [6] [8]. This includes speaker-generated variability and variability induced by recording channels and conditions. For example, there are researches studying the robustness against the corruption of the speech signal by channel and noise [9] [10] and mimicry by humans and computer-aided voice [11] [12]. On the other hand, features carrying vocal tract characteristics are usually used. In order to improve recognition performance, researchers have also

investigated other speaker-dependent features, such as features from excitation source [13] [14].

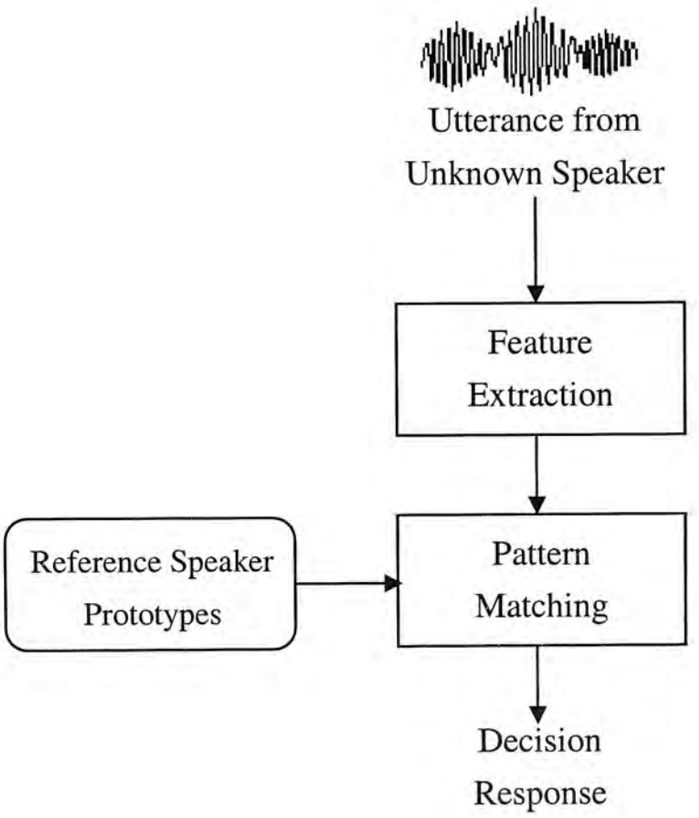


Figure 1-2 General operation of a speaker recognition system

1.3. Motivation and Focus of This Work

In order to improve recognition accuracy, speaker recognition research has been focused mainly on exploring speaker-dependent features from speech signal or enhancing the robustness against noise and channel variations. In this thesis, we focus on the speaker-dependent information in different frequency subbands.

The bandwidth of human speech is approximately up to 7 kHz [15]. Without bandwidth limitation on the speech signal, speaker recognition performed by human listeners cannot give perfect accuracy. It is common that we may recognize a person wrongly by only listening his/her voice. In real applications, the bandwidth of speech

data used is usually limited. It may degrade the performance of a speaker recognition system.

An important application of speaker recognition is on telephone network, which has a bandwidth of 0 – 4 kHz. Such a bandwidth is used based on the consideration that the transmitted speech signal can be reasonably perceived and understood. Speaker recognition is not the major concern. The underlying assumption is that the useful speaker information is mostly found at the frequency below 4 kHz.

However, researchers showed that there are important speaker-dependent characteristics beyond telephone bandwidth [16]–[18]. In [16], the speaker characteristics in the frequency band between 4 kHz and 10 kHz were investigated. The results showed that information in this band is useful to improve speaker recognition performance. In the study on the use of independent processing and recombination of subbands in speaker recognition with TIMIT database [17] [18], it was shown that speaker-specific information is not equally distributed over subbands. More speaker-dependent information is found in frequency subbands above 3 kHz.

From these findings, we have with a number of questions about the contribution of features in different frequency subbands. For example, what are the contributions from the features in the narrowband and higher frequency band for speaker recognition? How to fuse the decision from the features of these two bands in a proper way based on their relative importance? All these questions motivate us to study the importance of features in the narrowband and higher frequency band.

Previous works on subband processing in speaker recognition were focused mainly on two issues. They are optimal division of the frequency domain (e.g. [18]) and recombination of classification results from different subbands (e.g. [19]). We believe that a better understanding on the contribution of features from narrowband and higher frequency band is essential to these two issues. For example, if features

from higher frequency band are found to be more important in speaker recognition for some words, we may consider partitioning that region into finer subbands.

Speaker recognition over telephone is required in many applications. Therefore, we focus our study on the features from telephone bandwidth and higher frequency band. To be more precisely named the frequency bands, frequency ranged between 0 – 8 kHz, 0 – 4 kHz and 4 – 8 kHz are called wideband (WB), narrowband (NB) and higher frequency band (HB) in the thesis.

State-of-the-art speaker recognition systems use features that carry mainly vocal tract characteristics. These features are namely mel-frequency cepstral coefficient (MFCC) [20]. In this research, we focus our study on spectral envelope features that carry features of vocal tract.

We choose to study the contribution of features from NB and HB on word basis. Text-dependent speaker identification is performed. Text-dependent systems mean that training and testing data come from the same word. In many real applications, utterances containing digit strings are used. There has been no similar study on the contribution of features from subbands using Cantonese digits. Therefore, utterances containing digit strings are used in this study.

1.4. Thesis Outline

The thesis will be divided into seven chapters. A brief introduction of the fundamentals of speaker recognition will be given in the next chapter. In Chapter 3, the details of data collection will be described. We will also talk about the baseline of text-dependent speaker identification (SID) using MFCC features. Afterwards, text-dependent SID using spectral envelope features in NB, HB and WB will be studied in Chapter 4. In Chapter 5, combining the use of spectral envelope features

from NB and HB in feature level and model level will be investigated. In Chapter 6, the use of text-dependent weights for linear combination of scores from individual words in utterance-level SID will be studied. Also, applying text-dependent subband feature weights in utterance-level SID will be discussed. Finally, conclusions and some suggested future work will be given.

References

[1] A. K. Jain, A. Ross and S. Prabhakar, "An introduction to biometric recognition", *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4 – 20, Jan. 2004.

[2] (2002) Schiphol Backs Eye Scan Security, CNN World News [On-line], Available: http://www.cnn.com/2002/WORLD/europe/03/27/schiphol.security/

[3] B. Miller, "Vital signs of identity", *IEEE Spectrum*, vol. 31, issue: 2, pp. 22 – 30, Feb. 1994.

[4] B. S. Atal, "Automatic recognition of speakers from their voices", in *Proceedings of the IEEE*, vol. 64, no. 4, pp. 460 – 475, 1976.

[5] G. R. Doddington, "Speaker recognition – identifying people by their voices", in *Proceedings of the IEEE*, vol. 73, no. 11, pp. 1651 – 1664, 1985.

[6] A. E. Rosenberg and F. K. Soong, "Recent research in automatic speaker recognition", *Advances in Speech Signal Processing*, by S. Furui and M. M. Sondhi (Ed.), Marcel Dekker, New York, pp. 701 – 738, 1992.

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

[8] S. Furui, "Recent advances in speaker recognition", in *Proceedings of the Audio and Video based Biometric Person Authentication,* 1997, pp. 237 – 252.

[9] R. J. Mammone, X. Zhang and R. P. Ramachandran, "Robust speaker recognition – a feature-based approach", *IEEE Signal Processing Magazine*, vol. 13, issue: 5, pp. 58 – 71, 1996.

[10] H. A. Murthy, F. Beaufays, L. P. Heck and M. Weintraub, "Robust text-independent speaker identification over telephone channels", *IEEE Transactions on Speech and Audio Processing*, vol. 7, issue: 5, pp. 554 – 568, 1999.

[11] A. E. Rosenberg, "New techniques for automatic speaker verification", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-23, no. 2, pp. 169 – 176, April 1975.

[12] B. L. Pellom and J. H. L. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 1999, vol. 2, pp. 837 – 840.

[13] C. R. Jankowski, T. F. Quatieri and D. A. Reynolds, "Fine structure features for speaker identification", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 1996, vol. 2, pp. 689 – 692.

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

[15] D. O'Shaughnessy, *Speech Communications: Human and Machine*, 2nd Ed., Institute of Electrical and Electronics Engineers Press, 2000.

[16] S. Hayakawa and F. Itakura, "Text-dependent speaker recognition using the information in the higher frequency band", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 1994, vol. 1, pp. 137 – 140.

[17] L. Besacier and J. Bonastre, "Subband approach for automatic speaker

recognition: optimal division of the frequency domain", in *Proceedings of the Audio and Video based Biometric Person Authentication,* 1997, pp. 195 – 202.

[18] L. Besacier and J. Bonsatre, "Subband architecture for automatic speaker recognition", *Signal Processing*, vol. 80, issue 7, pp. 1245 – 1259, 2000.

[19] P. Sivakumaran, A. M. Ariyaeeinia and J. A. Hewitt, "Sub-band based speaker verification using dynamic recombination weights", in *Proceedings of International Conference on Spoken Language Processing,* 1998, vol. 3, pp. 551 – 554.

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

Chapter 2

Fundamentals of Automatic Speaker Recognition

This chapter provides the background theory of speech production and introduces the general principles of automatic speaker recognition. It also briefly describes the conventional approach of statistical speaker recognition based on short-time spectral features.

2.1. Speech Production

A schematic diagram of the human speech production mechanism is shown in Figure 2-1 [1]. Lungs, trachea, larynx (organ of voice production), throat, oral cavity (mouth) and nasal cavity are involved in speech production. The throat and oral cavities are usually grouped together and named as vocal tract. It begins at the output of the larynx, and terminates at the input to the lips. Vocal folds or vocal cords, soft palate or velum, tongue, teeth and lips are known as articulators. They move to different positions to produce various speech sounds. Here, only a brief description of speech production will be given. The reader is referred to [1] for more details.

In brief, speech is produced as follows. Air is expelled from the lungs through the trachea. Then the air flow passes through the vocal tract and is modulated in frequency by the resonances of the vocal tract, which its shape is changed by the articulators. Speech is produced as a sequence of sounds. The positions of articulators (e.g. jaw, tongue, lips and mouth) change over time to produce different speech sounds.

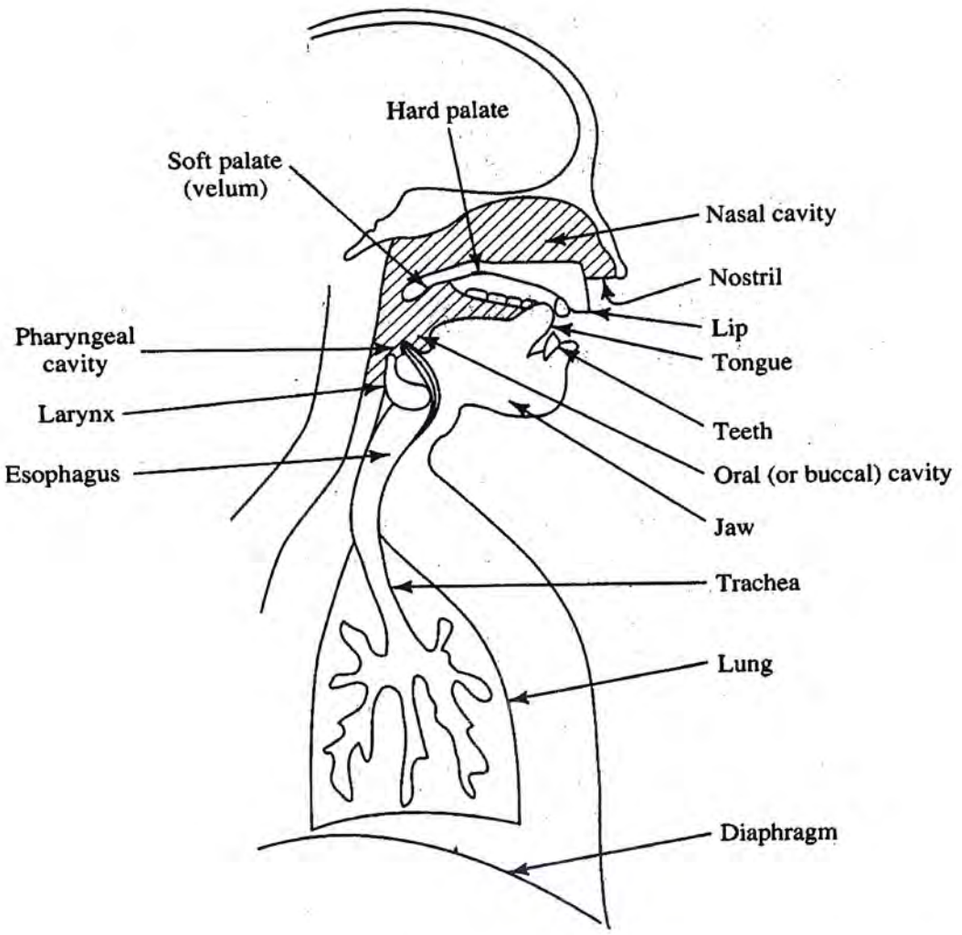


Figure 2-1 A schematic diagram of the human speech production mechanism [1]

Basically, speech can be classified into two types: voiced and unvoiced. They are produced with different excitation sources. When voiced speech is produced, the tensed vocal cords within the larynx are caused to vibrate periodically by the air flow. As a result, the air flow is chopped into quasi-periodic pulses and its period equals to the fundamental frequency (i.e. pitch). For unvoiced speech, the vocal cords are relaxed and do not vibrate. The air flow passes through a constriction in the vocal tract and becomes turbulent noise which is aperiodic in nature.

Based on the place and manner of articulation, speech sounds can be further categorized into several types, such as fricatives and nasals. For details, the reader is

referred to [2].

2.2. Features of Speaker's Voice in Speech Signal

Generally, features of speaker's voice can be physiologically or behaviorally based. In the process of speech production, physiological characteristics exhibit in several areas. For example, the shape of vocal tract contains speaker's information that can be estimated from the spectral shape (e.g. formant location and spectral tilt) of the voiced signal. The excitation source, which drives the human vocal mechanism, also contains speaker-dependent information. It can be characterized by the fundamental frequency of oscillation which depends on the length, tension and mass of the vocal folds. Also, variations in the velum and size of nasal cavities give different spectral spectrum when nasal sounds are produced. Other physiological speaker-dependent characteristics include maximum phonation time (the maximum duration that a syllable can attain) and glottal air flow (amount of air going through vocal folds) [3].

On the other hand, speaker-dependent behavioral characteristics include speaking rate, intonation and other speaking styles, such as preference in the choice of words.

However, some of the features mentioned above are difficult to be measured from the speech signal. For example, except for fundamental frequency, voice source characteristics, such as glottal source waveform, are not easy to extract from the speech signal.

Based on the practical consideration, features that can be captured from the speech signal are used in speaker recognition, such as short-term and long-term spectral energy, overall energy, and fundamental frequency [4]. The details of feature extraction will be discussed later in this chapter.

Source-Filter Model

As we have mentioned above, speech production can be viewed as air is forced through vocal cords and then filtered by the shape of the vocal tract. From an engineering perspective, usually a source-filter model is used to model the speech production process. Speech can be modeled as a quasi-periodic pulse (periodic over individual frames) when producing voiced speech or a noise-like turbulent flow of air when producing unvoiced speech followed by a linear time-invariant filter representing the vocal tract (see Figure 2-2).

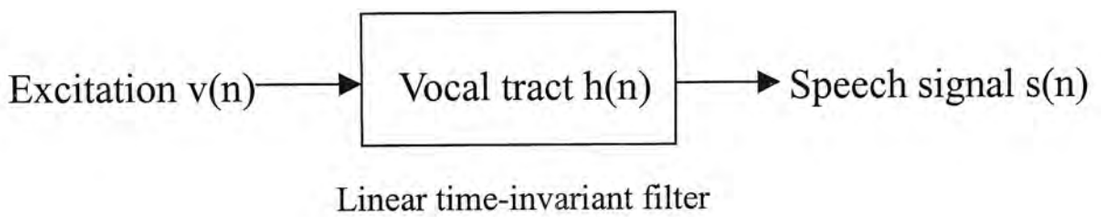


Figure 2-2 Source-filter model for speech production process

Let s(n) denotes the speech signal, h(n) is the impulse response of vocal tract and v(n) is the excitation source. Speech is composed of a convolution of an excitation source and the impulse response of the vocal tract. In frequency domain, speech can be represented as

$$S(\omega) = H(\omega)V(\omega) \tag{2.1}$$

By using source-filter model to represent the process of speech production, it can help us to understand the physical meaning of the speech spectrum. For most speech sounds, the shape of vocal tract varies slowly compared to the variations in the excitation source. So in the speech spectrum, the spectral envelope denotes the slowly varying shape of vocal tract (see Figure 2-3). As the shape of vocal tract contains speaker-dependent information, spectral envelope features are extracted and used in

speaker recognition.

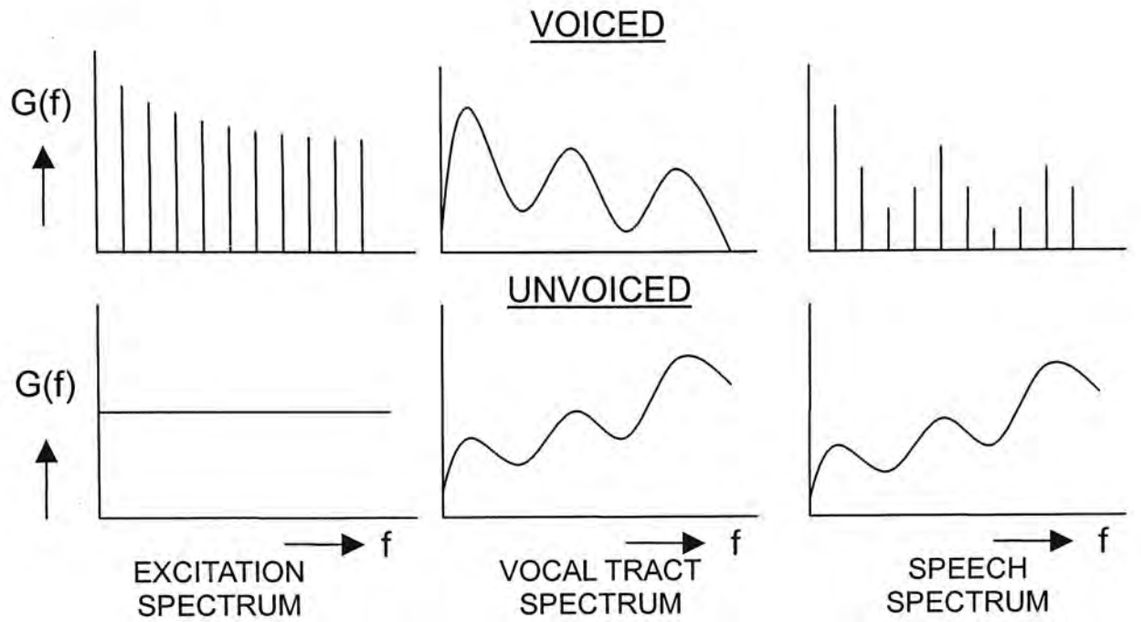


Figure 2-3 Spectra of voiced and unvoiced sounds. The spectral envelope gives the vocal tract spectrum [5]

Cepstral Analysis

The goal of feature extraction is to extract the vocal tract transfer function $H(\omega)$ from the speech spectrum $S(\omega)$. Since $H(\omega)$ and $V(\omega)$ in equation (2.1) are combined by multiplication, we cannot separate them directly to get $H(\omega)$. However, by taking logarithms of the spectral magnitude on both sides, we have

$$|\log S(\omega)| = |\log V(\omega)| + |\log H(\omega)| \tag{2.2}$$

In this way, the two individual components are combined by addition. The more important vocal tract shape information can then be separated from the less informative pitch information.

Cepstral analysis was firstly applied in speech signal processing [9]. It is performed based on the output from equation (2.2) to extract the vocal tract transfer function. The steps are shown in Figure 2-4.

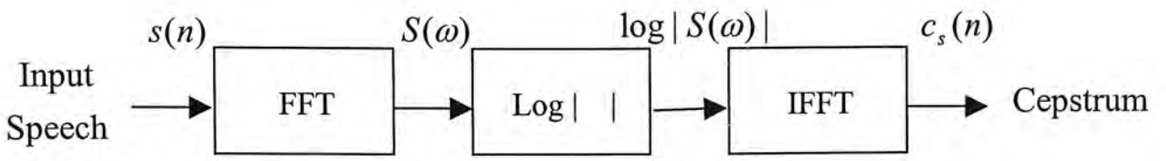


Figure 2-4 Steps of finding cepstrum

After finding the magnitude spectrum $|S(\omega)|$ computed by FFT for each speech frame, logarithm of the magnitude spectrum is computed and let $C_s(\omega) = \log|S(\omega)|$. By taking inverse FFT on $C_S(\omega)$, the cepstrum $c_s(n)$, which is spectrum of the log spectrum, is found. The whole process of finding cepstrum can be represented as follows:

$$\text{cepstrum (frame)} = \text{FFT}^{-1}(\log|\text{FFT (frame)}|)$$

Afterwards, the slowly varying component, which corresponds to the spectral envelope, produces the low-time part of the cepstrum. The component with fast variations, which corresponds to the excitation, results at larger values on the time axis of the cepstrum. Therefore, the low-order cepstral values are used to give the spectral envelope features.

2.3. Basics of Speaker Recognition

There are two different tasks of interest in speaker recognition: speaker identification (SID) and speaker verification (SV). In speaker identification, an input utterance is analyzed and compared with the models of a set of known speakers in the system. The speaker of the input utterance is then identified as one of the speakers in the system whose model best matches with the input utterance.

For speaker verification, the task is to determine if the input utterance is from the

claimed speaker. After analyzing the input utterance, it is compared with the model of the claimed speaker. A score that quantitatively measures the matching with the claimed speaker model is found. If the score is larger than a pre-determined threshold, the input utterance will be accepted as from the claimed speaker, otherwise, it will be rejected.

On the other hand, speaker recognition can also be classified based on the constraints on the materials used to train and test the system. It can be divided into three categories: text-dependent, text-independent and vocabulary-dependent. In text-dependent system, the training and testing data come from same phrase or word, while text-independent system has no such constraint on data. For vocabulary-dependent system, the training and testing data come from a limited vocabulary such as digits.

2.4. Existing Approaches of Speaker Recognition

A speaker recognition system mainly consists of two components: feature extraction and speaker modeling. For feature extraction, it concerns extracting a set of discriminative features so as to retain or enhance the inter-speaker variation while minimizing the intra-speaker variation. Speaker modeling involves constructing speaker models from the features extracted from the speech signal. These two components determine the success of a speaker recognition system. The approaches currently employed in these two components will be described in the following sections.

2.4.1. Feature Extraction

2.4.1.1 Overview

Indeed, there exist various choices of speaker-dependent features that can be derived from the speech signal. They include pitch, speech intensity and formant frequencies [5]. Usually features from spectral envelope are used in speaker recognition. They include cepstral coefficients derived from linear predictive (LP) analysis [5] and filter-bank analysis [2]. Nowadays, mel-frequency cepstral coefficient (MFCC) [6] is commonly used due to its reported good performance [8]. The details of MFCC will be discussed in the next section.

2.4.1.2 Mel-Frequency Cepstral Coefficient (MFCC)

In practice, cepstral coefficients derived from a mel-frequency filterbank are used and this type of feature is called mel-frequency cepstral coefficient (MFCC). Figure 2-5 shows the steps of extracting MFCC features.

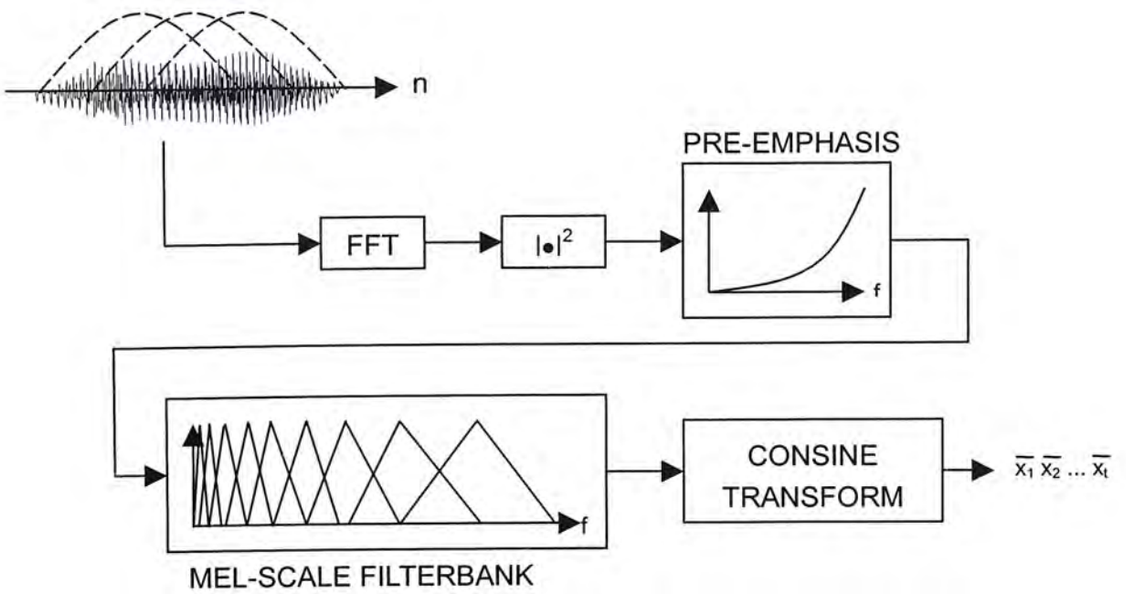


Figure 2-5 Steps of extracting MFCC features [10]

After finding the magnitude spectrum computed by FFT for each speech frame, it is processed in a simulated mel-scale filterbank.

Psychophysical studies have shown that the human ear resolves frequencies non-linearly across the audio spectrum [11]. A mel is a unit of measured of perceived pitch or frequency of a tone [1] and it can be approximated by [12]

$$Mel(f) = 2595\log_{10}\left(1+\frac{f}{700}\right) \tag{2.3}$$

Figure 2-6 shows the mapping between the linear frequency scale (Hz) and the perceived frequency scale (mel). The mapping is approximately linear below 1 kHz and logarithmic above [13].

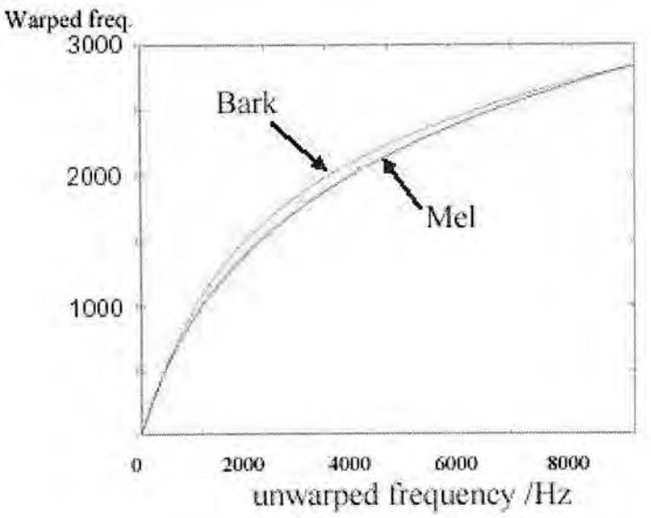


Figure 2-6 Warping the linear frequency scale to two commonly used non-linear frequency scale: Mel-scale and Bark-scale [19]

Mel-scale models the sensitivity of the human ear in a closer way than a purely linear scale. It is suggested that designing a front-end to operate in a similar manner as human auditory system can improve recognition performance. Nowadays, mel-scale frequency analysis has been widely used in modern speech recognition [14]. It is also commonly used in speaker recognition.

On the other hand, it is found that the perception of a particular frequency by the

auditory system, say f_0, is influenced by energy in a critical band of frequencies around f_0 [1]. Therefore, it is suggested to use the log total energy in critical bands around the mel frequencies to compute cepstral coefficients. Figure 2-7 shows the mel-scale filterbank used. The filters are triangular shape and they are equally spaced along the mel-scale in the Nyquist range.

Therefore, each magnitude coefficient of the spectrum is multiplied by the corresponding mel-scale triangular filter gain and the results are accumulated. Then the log-energy filter outputs are cosine transformed to produce the cepstral coefficients C_i using the following formula:

$$C_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right), \mathrm{i} = 0,1,... \tag{2.4}$$

where N is the number of filterbanks used and m_j is log filterbank amplitude.

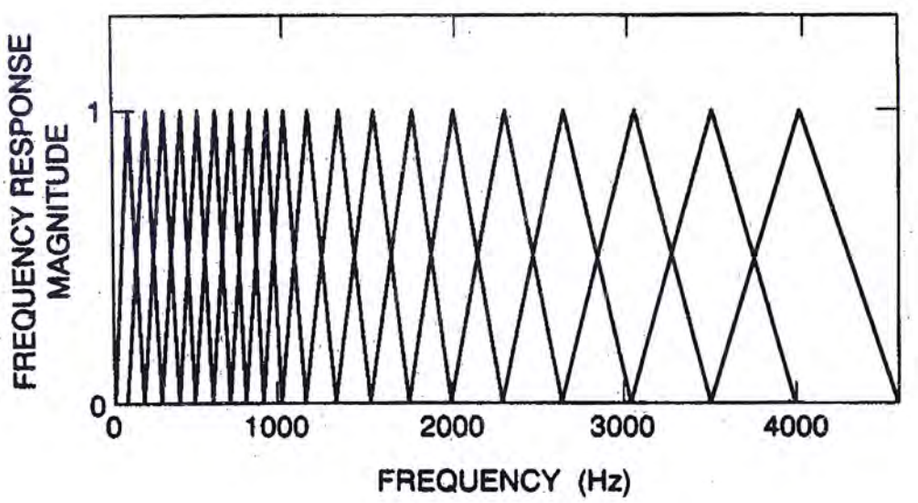


Figure 2-7 The triangular mel-scale filterbank distributed in the Nyquist range [2]

2.4.2. Speaker Modeling

2.4.2.1 Overview

There are mainly two approaches for speaker modeling: template models and stochastic models. With template models, the test utterance is compared with a collection of templates developed for each of the speakers in training and decision is made based on the distances to the templates. Example of template model is vector quantization (VQ) codebook [15]. It makes use of standard clustering procedures on the training data and stores multiple templates in term of a VQ codebook to characterize frames of speech. However, this approach is particularly sensitive to variation in background noise and it is not robust enough to use.

Instead of modeling speakers by templates, stochastic modeling uses probability distribution to represent features of a speaker's voice. The conditional probability distribution is estimated for each speaker from a set of training data. The speaker's identity of the test utterance is determined by measuring the likelihood of the input utterance given the speaker model in the system. This approach is widely used nowadays as it can offer good flexibility and result in a theoretically meaningful probabilistic likelihood score [3].

One of the most popular stochastic techniques for modeling is Hidden Markov Model (HMM). It models both the underlying speech sounds and the temporal sequencing among these sounds. However, the temporal change of speech does not contain too much speaker-dependent information in text-independent task. Recently, Gaussian mixture model (GMM) [10] is commonly used in speaker recognition because it is computationally inexpensive and provides high recognition accuracy [16].

2.4.2.2 Gaussian Mixture Model (GMM)

The distribution of feature vectors extracted from a person's speech is modeled by a Gaussian mixture density. Let $\vec{x}$ be a D-dimensional feature vector, the mixture density for a speaker is computed by

$$p(\vec{x} \mid \theta) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{2.5}$$

It is a weighted sum of M unimodal Gaussian densities, $b_i(\vec{x})$. $b_i(\vec{x})$ can be expressed as,

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} \left|\sum i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x} - \vec{\mu}_i)' \left(\sum i\right)^{-1} (\vec{x} - \vec{\mu}_i)\right\} \tag{2.6}$$

where $\vec{\mu}_i$ is a D x 1 mean vector and $\sum i$ is a D x D covariance matrix. p_i is the mixture weight and satisfy the constraint $\sum_{i=1}^{M} p_i = 1$. The entire GMM model can be represented by the notation $\theta = \{ p_i, \ \vec{\mu}_i, \ \sum i \}$, where $i = 1, \ldots, M$.

Theoretically, full covariance matrix is required. Diagonal covariance matrices are used in practice because they provide more computational efficiency.

We can interpret the physical meaning of using GMM to model the distribution of features of a speaker's voice in the following way. Speech is made up of different phonetic classes. Each individual mixture component may correspond to a phonetic class. Also, the Gaussian mixture density is used to model the underlying long-term sample distribution of observations obtained from utterances of the speaker. GMM not only models the distribution in individual phonetic class, but also models the probability of having these phonetic classes in speaker's utterance.

2.4.3. Speaker Identification (SID)

This research is focused on the task of speaker identification (SID). The implementation details on how to train GMM and how to make decision based on the input utterance in SID will be discussed.

Training of GMM

The goal of training is to estimate the parameters of GMM that best matches the distribution of the training feature vectors. The most popular and well-established approach is known as the maximum likelihood (ML) estimation [10]. It is aimed at finding the model parameters that maximize the likelihood of the GMM given the training data (see equation (2.7)).

$$\theta_{MLE} = \arg\max_{\theta} p(\vec{x} \mid \theta) \tag{2.7}$$

where θ and $\vec{x}$ denote model parameters of the GMM and training data respectively.

Figure 2-8 shows the steps of training GMM model. To begin, the model parameters are initialized by partitioning all speech frames into K clusters, where K is the number of mixture components. This is done by clustering algorithm such as clustering them in random. Then, the feature vectors in each cluster gives the mixture weight, while means and covariances are derived directly from the vectors in each cluster.

The estimation of ML parameters can be obtained iteratively using the expectation-maximization (EM) algorithm [18]. The basic idea of EM algorithm is to estimate a new set of model parameters from the initial one such that the value of the model likelihood increases monotonically. It is implemented by choosing the Gaussian component with the maximum likelihood from the estimated mixture model.

A new set of model parameters can then be found. Afterwards, the new model becomes the initial model for the next iteration and this process is repeated iteratively until the model parameters converge. This is a critical stage as if the parameters of a GMM model are not well estimated, they cannot reflect the actual distribution of the speaker's features. It will affect the recognition performance.

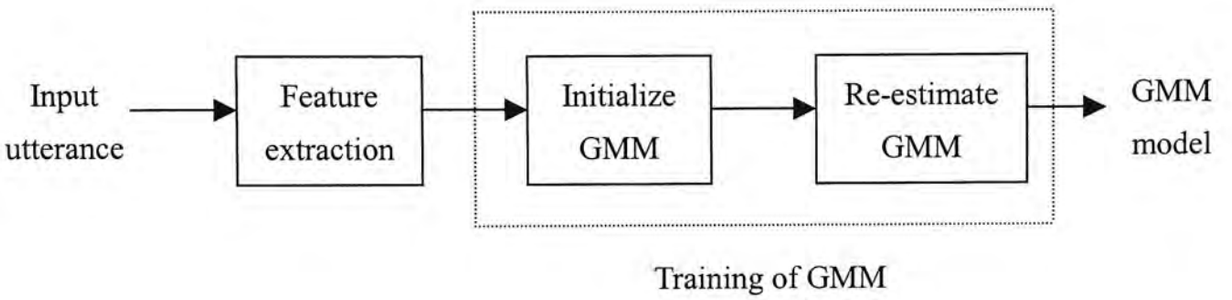


Figure 2-8 Steps of training GMM model

Identification Process

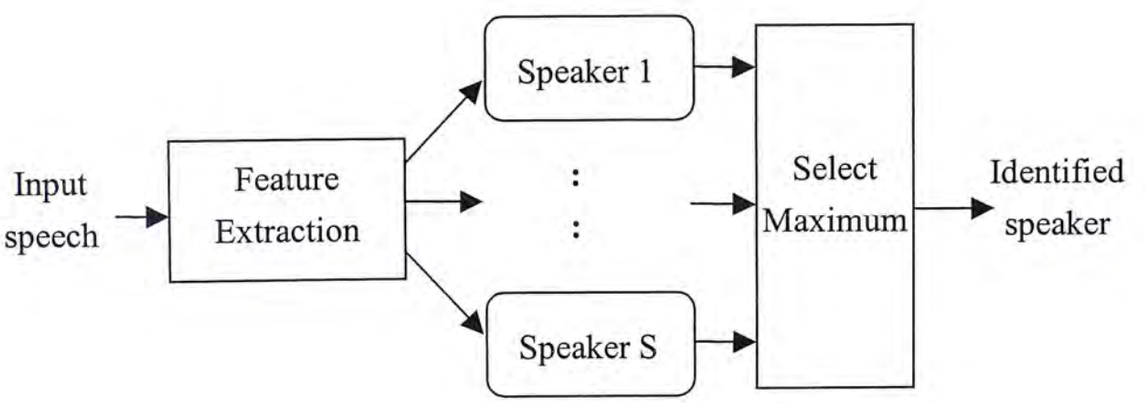


Figure 2-9 Block diagram of a speaker identification system

Figure 2-9 shows the block diagram of a speaker identification system. Suppose there are S speakers in the system. After extracting features from the input utterance with feature vector sequence $X = \{\vec{x}_1,...,\vec{x}_T\}$, then the a posteriori probability of X, $Pr(\theta|X)$,

from each of the S speakers in the system is computed. The decision rule of identification is to find the speaker model which has the largest value of $Pr(\theta|X)$ (equation (2.8)).

$$\hat{S} = \arg\max_{1\le s\le S} \Pr(\theta_s \mid X) \tag{2.8}$$

By applying Bayes' rule, it is equal to

$$\hat{S} = \arg\max_{1\le s\le S} \frac{p(X \mid \theta_s)}{p(X)} \Pr(\theta_s) \tag{2.9}$$

Assuming equal prior probabilities of speakers, the term $p(X)$ and $Pr(\theta_s)$ are constant for all speakers. Therefore, equation (2.9) is simplified to

$$\hat{S} = \arg\max_{1\le s\le S} p(X \mid \theta_s) \tag{2.10}$$

Using logarithms and the independence between observations, the decision rule becomes

$$\hat{S} = \arg\max_{1\le s\le S} \frac{1}{T} \sum_{t=1}^{T} \log p(x_t \mid \theta_s) \tag{2.11}$$

The likelihood score for each frame, $p(x_t \mid \theta_S)$, is found by equation (2.5). In order to normalize the utterance duration, the log-likelihood value is divided by T, which is the number of frames of the input utterance.

References

[1] J. R. Deller, Jr., J. G. Proakis, J. H. L. Hansen, *Discrete-time processing of speech signals*, Macmillan, New York, 1993.

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

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

[4] A. E. Rosenberg and F. K. Soong, "Recent research in automatic speaker recognition", *Advances in Speech Signal Processing*, by S. Furui and M. M. Sondhi (Ed.), Marcel Dekker, New York, pp. 701 – 738, 1992.

[5] B. S. Atal, "Automatic recognition of speakers from their voices", in *Proceedings of the IEEE*, vol. 64, no. 4, pp. 460 – 475, 1976.

[6] S. Furui, "Cepstral analysis technique for automatic speaker verification", *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-29, no. 2, April 1981.

[7] S. Furui, "Recent advances in speaker recognition", in *Proceedings of the Audio and Video based Biometric Person Authentication*, 1997, pp. 237 – 252.

[8] D. A. Reynolds, "Experimental evaluation of features for robust speaker identification", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 639 – 643, 1994.

[9] A. M. Noll, "Cepstrum pitch determination", *Journal of the Acoustical Society of America*, vol. 41, pp. 293 – 309, Feb. 1967.

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

[11] S. S. Stevens and J. Volkman, "The relation of pitch to frequency", *American Journal of Psychology*, vol. 53, pp. 329, 1940.

[12] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, Dec. 2001.

[13] W. Koenig, "A new frequency scale for acoustic measurements", *Bell Telephone Laboratory Record*, vol. 27, pp. 299 – 301, 1949.

[14] X. Huang, A. Acero and H. W. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*, Prentice Hall, 2001.

[15] F. K. Soong, A. E. Rosenberg, L. R. Rabiner and B. H. Juang, "A vector quantization approach to speaker recognition", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1985, pp. 387 – 390.

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

[17] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital Signal Processing*, vol. 10, pp. 19 – 41, 2000.

[18] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm", *Journal of the Royal Statistical Society*, vol. 39, pp. 1 – 38, 1977.

[19] Tan Lee, *Course Notes of Automatic Speech Recognition, The Chinese University of Hong Kong*, 2004.

Chapter 3

Data Collection and Baseline System

3.1. Data Collection

Design Considerations

The general considerations on the design of data collection include the constraints on the recording materials and do recording in single or multiple sessions [3]. In this research, we focus our study on text-dependent speaker identification. On the other hand, speakers are required to do recording in multiple sessions. It is known that the voice for the same speaker will be changed with time (e.g. changing speaking behavior and aging [1]) and it is called intra-speaker variation. A speaker recognition system should be able to accommodate natural and expected modifications in speech signal characteristics due to this type of variation [2]. Therefore, speech data should be collected in multiple sessions.

In this research, we focus our study on the contribution of features from different frequency bands. We choose to implement text-dependent SID on digit basis. Therefore, the recording materials only consist of digit strings.

Usually, the difference of voice between male and female is significant. To focus on our study, other factors that might affect the recognition performance should be eliminated. To eliminate the effect of gender difference, speech data is collected from male speakers only.

Recording Set-up

Speakers are required to do recording inside a confined room that provides a closed silent recording environment. At the beginning, speakers are asked to read an instruction on recording procedures. They are required to read out the prompted texts from the computer screen to the microphone using the set-up as shown in Figure 3-1. Speaker can choose from the menu to record a particular utterance again if he has spoken it wrongly.

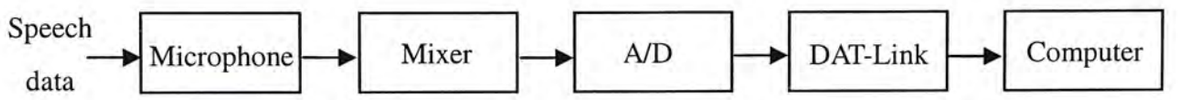


Figure 3-1 Set-up for recording

The microphone used for recording is a head mount microphone (Figure 3-2). It works in close-talk operation. The speech data collected in this way can be in high quality.

Figure 3-2 Head mount microphone [4]

The analog signal passes through a mixer and is A/D converted at 44.1kHz, 16 bit using Digital Audio Tape (DAT) recorder. Then, the digital data is down-sampled

in real time to 16 kHz by DAT-Link using its built-in digit signal processor and sent through a SCSI interface to the hard disk of the computer. Therefore, the output is 16 bit signed with the most significant byte stored first. It is sampled at 16 kHz and quantized using linear Pulse Code Modulation (PCM). The output signal is in mono type and only has single channel.

Recording Materials

Table 3-1 summarizes the details of recording materials. Speech data from 20 male speakers are collected. The time span between sessions ranges from 3 days to 2 weeks. The entire data collection is lasted for 2.5 months.

Number of recording sessions	12
Time separation between sessions	S01 – S08: 3 days
	S09 – S10: 1 week
	S11 – S12: 2 weeks
Types of speech content and number of utterances for each type	10 utterances of singe digit, i.e. digit '0' – '9'
	5 utterances of digit string (2 digits), e.g. '36'
	5 utterances of digit string (8 digits), e.g. '9408 4513'
Total number of utterances	100
Change of the use of recording materials	Change for every 2 consecutive sessions

Table 3-1 Details of recording materials

The speech content consists of three types: single digit, digit string containing two digits and eight digits (see Table 3-1). The digit string is read in the following way: e.g. '36' is read as 'three-six', instead of 'thirty-six' in Cantonese. Each type of digit string has five different utterances in the same session. For each utterance, it is required to record repeatedly for five times. Therefore, there are totally 100 utterances

in a single session.

The digit strings used in recording are different for every two consecutive sessions. They are designed in such a way that the total number of occurrences in the 12 sessions for each digit is approximately the same. Table 3-2 listed the total number of occurrences for the 10 Cantonese digits in the 12 recording sessions. The number of occurrences is similar among the 10 digits and their average is equal to 360. The speech content used in recording is listed in Appendix 1.

Digit	Total number of occurrences in the 12 sessions
0	370
1	350
2	360
3	390
4	360
5	330
6	340
7	390
8	350
9	360

Table 3-2 Total number of occurrences in the 12 sessions for the 10 Cantonese digits

Post-processing of Speech Data

Speech data are randomly selected to verify if the speakers utter the designated texts correctly. Also, all collected speech data are processed to check if there is overflow in recording. If so, the corresponding data will be discarded to use.

3.2. Baseline System

3.2.1. Experimental Set-up

Figure 3-3 shows the steps of baseline system. It mainly consists of 2 steps: training of speaker model and evaluation of the system. Each of them will be introduced one by one.

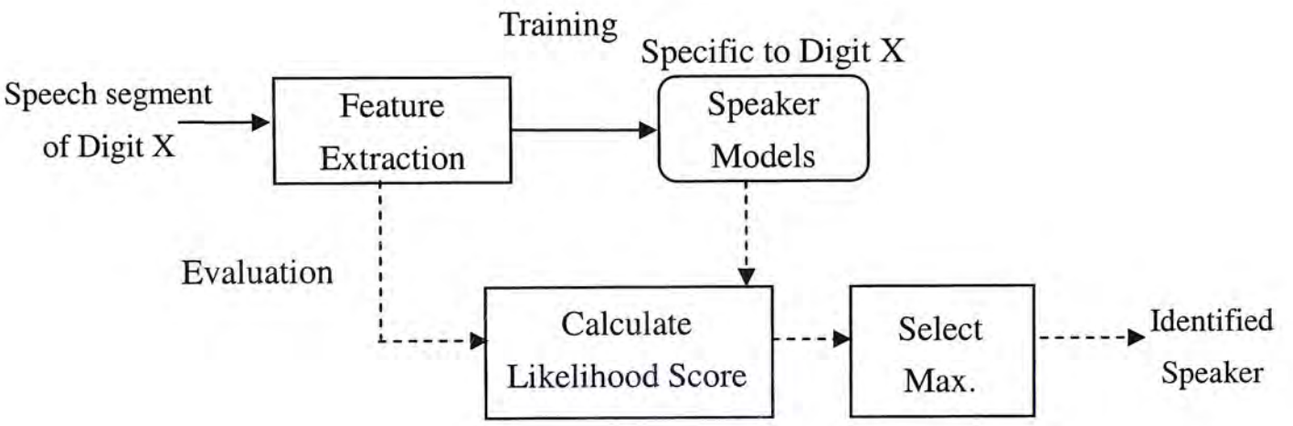


Figure 3-3 Steps of baseline system

Forced Alignment

Text-dependent SID is implemented in this study. Segments of speech speaking the same digit are used to train digit-specific speaker models. The collected utterances may contain a sequence of digits. Hence, alignment is required to find the duration of each digit spoken in the utterance.

The step of forced alignment is not shown in Figure 3-3. It is performed before feature extraction. Assuming that the content of the input utterance is known, speech recognizer of a particular digit is used to find the duration for that digit within the utterance. The alignment result is then used in feature extraction.

The speech recognizer[1] we used is trained by using data from the male speakers set in the CUDIGIT corpus [5], which consists of microphone speech of continuous Cantonese digit strings sampled at 16 kHz. The speech recognizer is trained with features containing 12 mel-cepstral coefficients with 1 energy parameter, and their first and second derivatives, and its recognition accuracy is 97.24%.

It is important to remove silence/noise frames from both the training and testing data to avoid modeling and detecting the environment rather than the speaker [6]. Since the speech recognizer we used also contains silence model, performing forced alignment can also serve the purpose to find out the duration of silence within the utterance.

Throughout this study, it is assumed that the content of input speech is known. Forced alignment will be performed on all speech data, including single-digit utterances.

Training of Speaker Model

Gaussian mixture model (GMM) is used in the baseline system. With the speech segment of digit X spoken by a particular speaker, features will be extracted and used to train the model that is specific to digit X for that speaker. The steps of training have been discussed in Chapter 2.

Evaluation

After feature extraction, the log-likelihood scores, or simply called likelihood scores, are computed by using speaker models specific to the input digit. The one with the highest score will be identified as the speaker of the input speech (equation (2.11)).

---

[1] The speech recognizer is provided by Ms. Zhu Yu, M. Phil candidate in the DSP and Speech Technology Laboratory, Department of Electronic Engineering, the Chinese University of Hong Kong.

Configuration of Baseline System

In our study, data from the first six sessions (S01 – S06) are used for training the speaker models while data from the remaining sessions (S07 – S12) are used for evaluating the speaker recognition system (see Table 3-3).

Data	Use
S01 – S06	Train digit-specific speaker models
S07 – S12	Evaluate the system

Table 3-3 Use of data in the baseline system

Mel-frequency cepstral coefficient (MFCC) is used in this baseline system (Digit Baseline 1). In generating these features, each digit segment is segmented into 20 ms frames at intervals of 10 ms using a Hamming window and it is pre-emphasized with coefficient of 0.97. Energy normalization is performed before extracting cepstral coefficients. Such configuration is commonly used in speaker recognition [8]. [8] also mentioned that usually the zeroth cepstral coefficient is not included in the feature vector. This coefficient only represents the average energy of the input speech and it is expected not to contain too much information on the speaker's voice. Therefore, in the baseline system, the first 22 cepstral coefficients, except the zeroth cepstral coefficient, will be used to compose the feature vector.

As mentioned in Chapter 2, MFCC is found from a mel-frequency filterbank. It is required to determine the number of mel-frequency filters. In [9], 24 mel-filters were used to extract features for SID. It used the TIMIT database [3] with sampling rate of 16 kHz and it is similar with the speech data we collected before. On the other hand, using more mel-filters to extract MFCC from wideband speech may give better performance. Therefore, using 24 and 32 mel-filters in feature extraction will be performed.

Also, the number of Gaussian mixtures used in speaker model will be determined by trying with different number of Gaussian distributions.

3.2.2. Results and Analysis

Table 3-4 and Table 3-5 listed the overall identification rate for the 10 Cantonese digits using 24 and 32 mel-filters in feature extraction respectively. From the results, it is seen that the identification rate does not depend too much on the number of filterbanks. But on average, using 32 mel-filters to extract MFCC gives slightly better performance. Therefore, 32 mel-filters are chosen to use in feature extraction.

Set 1 Use 24 mel-filters

	Overall SID rate (%)				
Digit	8 mix.	16 mix.	24 mix.	32 mix.	64 mix.
0	97.5	97.55	97.94	98.19	98.16
1	92.71	94.55	95.41	95.72	95.94
2	93.49	94.3	94.77	94.99	95.44
3	96.91	97.55	97.58	97.69	97.72
4	96.61	97.69	98.36	98.53	98.72
5	86.56	86.76	86.92	87.76	86.81
6	91.87	93.9	94.73	95.52	95.94
7	94.94	97.26	97.57	97.79	98.02
8	92.66	93.46	94.1	94.44	94.27
9	95.3	96.33	97.02	97.39	97.5
Average	93.86	94.94	95.44	95.80	95.85

Table 3-4 Overall SID rate (%) of using 24 mel-filters in feature extraction

Set 2 Use 32 mel-filters

	Overall SID rate (%)				
Digit	8 mix.	16 mix.	24 mix.	32 mix.	64 mix.
0	97.58	98.03	98.08	98.33	98.33
1	93.3	94.6	95.33	96.05	96.19
2	94.41	94.74	95.19	95.69	95.8
3	97.69	98.14	98.25	97.97	97.69
4	96.75	97.69	98.08	98.47	98.5
5	86.34	87.67	87.79	87.48	86.25
6	91.49	93.84	94.49	95.52	95.55
7	95.47	96.68	97.52	97.81	98.02
8	92.74	94.02	94.55	94.52	94.24
9	95.44	96.44	97.33	97.36	97.69
Average	94.12	95.19	95.66	95.92	95.83

Table 3-5 Overall SID rate (%) of using 32 mel-filters in feature extraction

The objective of speaker modeling is to choose the most appropriate number of Gaussian mixtures to adequately model a speaker for good SID. On the other hand, the number of Gaussian mixtures used is related to the amount of data available for training. We need to consider if the available training data is sufficient enough to train so many Gaussian mixtures. Table 3-2 gives the total number of occurrences for the 10 digits in the 12 sessions and their average equals to 360. Since data from the first six sessions are used to train speaker models, there are 180 digit segments available for training on average. Suppose the average number of speech frames for a digit segment is 20, so approximately 3600 speech frames are used to train a speaker model. As a rule of thumb, a Gaussian distribution should require about 100 speech frames to train it. Therefore, with the available training data, not more than 64 Gaussian mixtures should be used in a speaker model. From Table 3-5, we can find that the average identification rate kept increasing with the number of Gaussian mixtures used.

But the identification rate by using digit '3', '5' and '8' started to decrease when the number of Gaussian mixtures increased from 24 to 32. Hence, 24 Gaussian mixtures are chosen to use in the baseline.

To conclude, in the baseline system (Digit Baseline 1), 32 mel-filters are used in feature extraction and each speaker model is represented by 24 Gaussian mixtures. The highlight part in Table 3-5 shows the baseline result.

References

[1] S. Furui, "An analysis of long-term variation of feature parameters of speech and its application to talker recognition", *Electron. Commun. Jap.*, vol. 57a, pp. 34 – 42, 1974.

[2] A. E. Rosenberg and F. K. Soong, "Recent research in automatic speaker recognition", *Advances in Speech Signal Processing*, by S. Furui and M. M. Sondhi (Ed.), Marcel Dekker, New York, pp. 701 – 738, 1992.

[3] J. P. Campbell Jr. and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 1999, vol. 2, pp. 829 – 832.

[4] Shure Model SM10A User Guide.

[5] W. K. Lo, Tan Lee and P. C. Ching, "Development of Cantonese spoken language corpora for speech applications", in *Proceeding of the first International Symposium on Chinese Spoken Languages Processing*, 1998.

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

[7] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev and P. Woodland, *The HTK Book (for HTK Version 3.1)*, Cambridge University Engineering Department, Dec. 2001.

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

[9] 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 *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 1995, vol. 1, pp. 329 – 332.

Chapter 4

Subband Spectral Envelope Features

This chapter describes the extraction of spectral envelope features from a set of prescribed subbands. More precisely, we divide the full band into 0 – 4 kHz (NB) and 4 – 8 kHz (HB). The narrowband (NB) is further divided into four non-overlapping bands. Cepstral analysis is performed in each subband to find the corresponding spectral envelope features. The contributions of NB and HB features will be investigated via SID experiments.

4.1. Spectral Envelope Features

General Principles

By applying cepstral analysis, spectral envelope features can be extracted. The underlying principle is briefly introduced in the following. For more details about cepstral analysis, please refer to Chapter 2.

As we have mentioned in Chapter 2, the speech production process can be modeled by a source-filter model. Speech signal can be viewed as a convolution of an excitation source and the impulse response of vocal tract. Taking logarithm on the short-time power spectrum can change the combination of the above two components from multiplication to addition. After applying Discrete Cosine Transform (DCT), cepstrum can be found. Low-order cepstral coefficients represent the slowly varying

shape of vocal tract, i.e. the spectral envelope in the speech spectrum. Those cepstral coefficients are used to represent spectral envelope features. Figure 4-1 shows the general process of extracting short-time spectral envelope features.

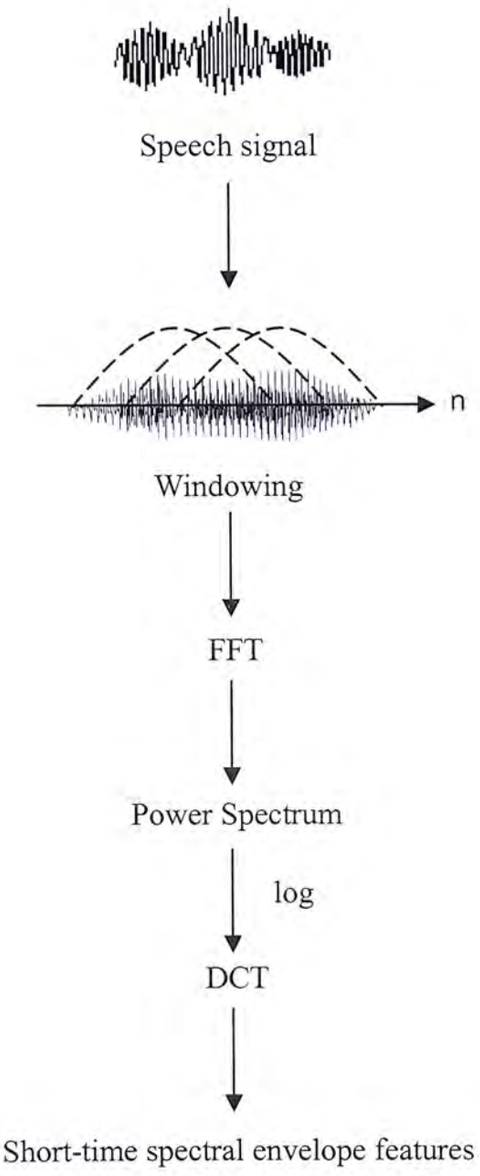


Figure 4-1 Extraction of short-time spectral envelope features

Comparison with MFCC

MFCC is one of the methods to find features of spectral envelope (see details in Chapter 2). The difference between MFCC and spectral envelope features computed by the method described above is that MFCC is computed by filter-bank analysis in mel-scale while spectral envelope features are computed directly from FFT spectrum

in linear frequency scale.

If filter-bank analysis is used in feature extraction, the found feature components describe the global shape of the spectral envelope. On the other hand, feature components that are computed directly from FFT spectrum describe the local spectral envelope. If high-order cepstral coefficients are used, the reconstructed spectrum will also contain fine harmonics. It indicates features from vocal source are also included.

4.2. Subband Spectral Envelope Features

Motivation

In this study, we focus on the contribution of spectral envelope features in NB and HB for speaker recognition. During computation of MFCC from the full band, the log-energy filter outputs undergo a cosine transform to produce the cepstral coefficients. In the cepstral domain, we cannot tell what frequency each cepstral coefficient is representing. Therefore, MFCC computed from full band is not suitable for our intended study.

The spectral envelope computed by filter-bank analysis is smoothed. We want to study the contribution of the real spectral envelope. Therefore, we choose to compute spectral envelope features from FFT spectrum directly.

In cepstral domain, feature components are independent to each other, but they are not in spectral domain. As mentioned in Chapter 2, diagonal covariance matrices in GMM are used in practice for the computational efficiency. The underlying assumption is that the feature components are uncorrelated. Therefore, features found in cepstral domain are used.

If the same number of cepstral coefficients is used, the reconstructed spectral envelope found from a single band is similar with the one found from subbands

partitioned from a single band, except there is discontinuity between consecutive subbands in the latter case. An example is illustrated in Figure 4-2. NB is partitioned into four non-overlapping subbands, each has bandwidth of 1 kHz. 20 cepstral coefficients are used to reconstruct the spectral envelope in NB. We can see that the reconstructed spectral envelope are very similar between the two cases.

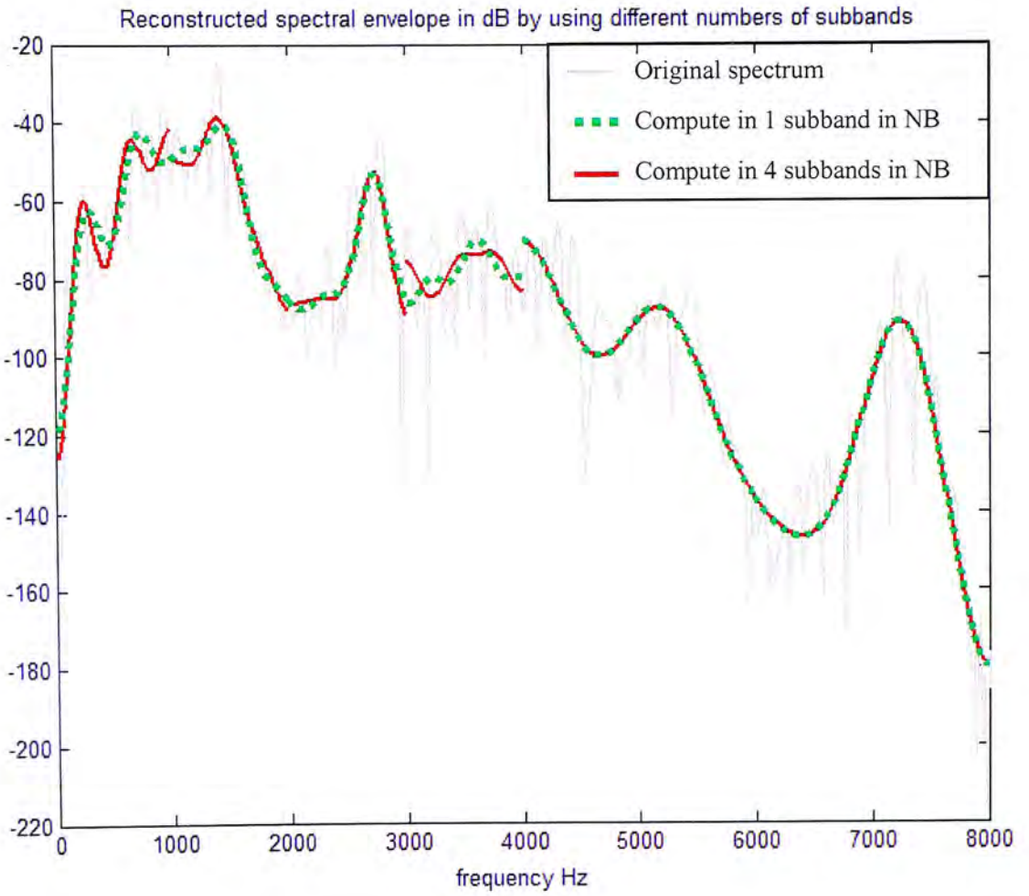


Figure 4-2 An example illustrates that the reconstructed spectral envelope found from a single band in NB is similar with the one found from subbands partitioned from NB

Although spectral envelope features can be computed by either from a single band or from subbands partitioned from a single band, subband spectral envelope features are used in this study. The full band is divided into subbands and spectral envelope features in each subband are obtained. In this way, we can choose to use or

not use the features from any combination of these subbands. It facilitates the goal of our study. It also provides more flexibility in feature extraction. We can freely select the number of cepstral coefficients used in each subband based on the importance of that subband.

Consideration of Subband Design

As described in Chapter 2, the excitation source is modulated in frequency by the resonances of the vocal tract and speech signal results. The resonance frequencies are commonly called formants. The first three formants are usually below 3500Hz [1]. Therefore, NB has a relatively high energy concentration. It is expected that features of a speaker's voice are comparatively rich in that region. Based on this consideration, NB is further partitioned into finer subbands and spectral envelope features are extracted from each subband. It can give a more detailed spectral envelope in this way.

Figure 4-3 summarizes how the full band is divided into five subbands for feature extraction. The higher frequency band (HB) ranges from 4 kHz to 8 kHz while the narrowband (NB) is partitioned evenly into four non-overlapping subbands and each subband has bandwidth of 1 kHz.

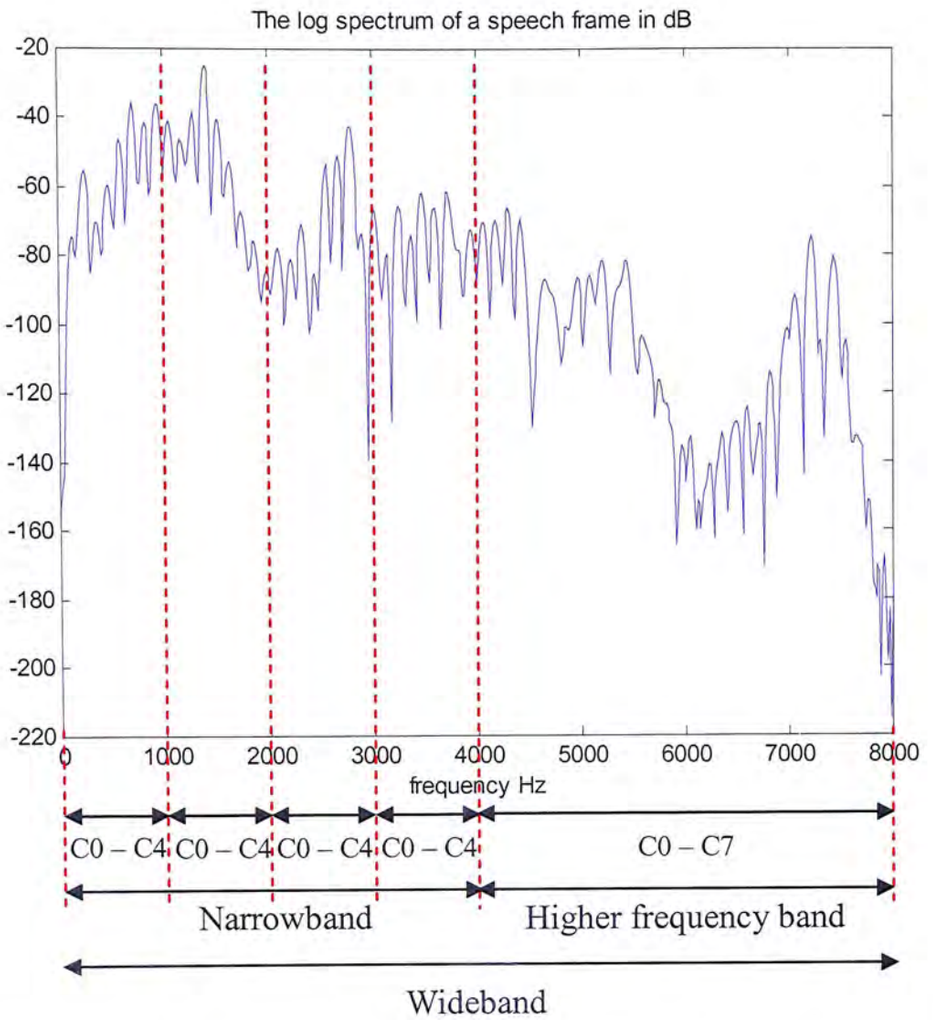


Figure 4-3 Division of frequency bands for extracting subband spectral envelope features

NB is partitioned evenly into four non-overlapping subbands. The reason why NB is not partitioned into fewer subbands (e.g. 2) is related to the number of cepstral coefficients used. This will be further explained in the next section.

On the other hand, NB should not be cut into too many subbands. Otherwise, the discontinuity between consecutive subbands will become significant and it will affect describing the shape of spectral envelope. We can observe this from the example illustrated in Figure 4-4. In this example, NB is partitioned into 16 subbands. The first three cepstral coefficients in each subband are used to reconstruct the spectral envelope.

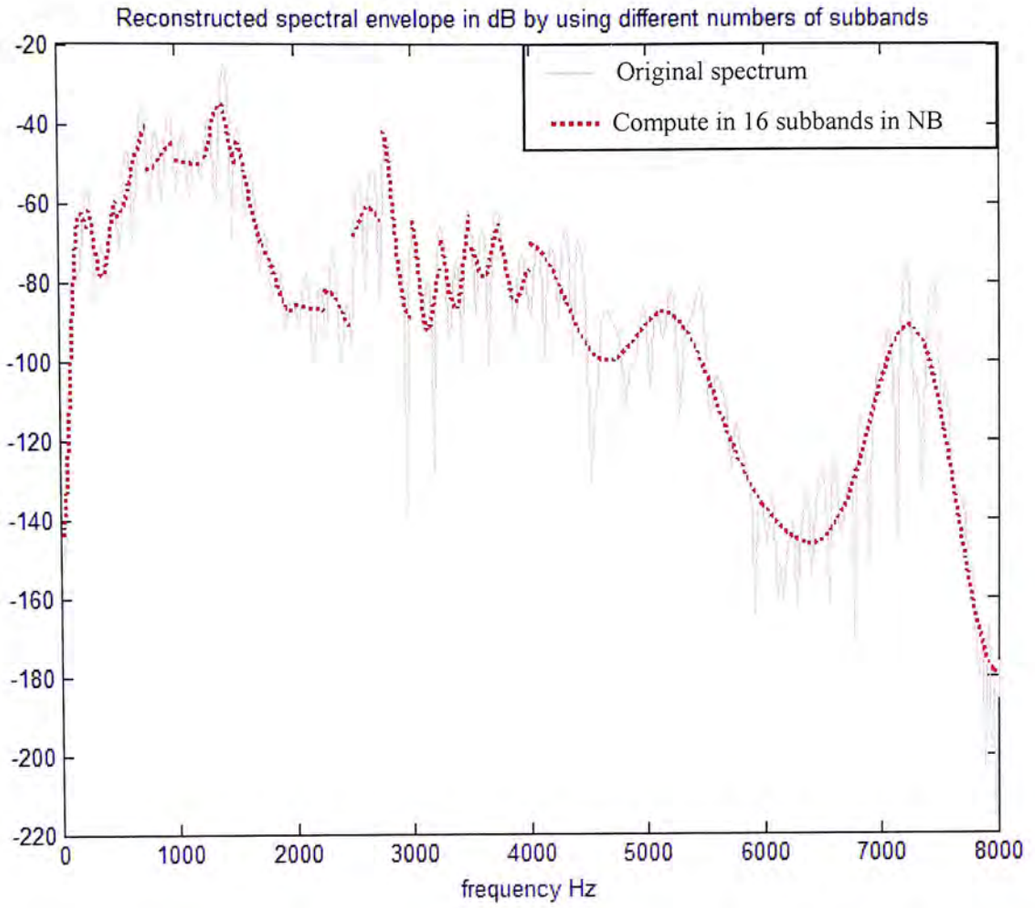


Figure 4-4 An example illustrates that the discontinuity between subbands affect describing the spectral envelope if a single band is partitioned into too many subbands to compute spectral envelope features

The SID result of using MFCC features (Digit Baseline 1) is compared with the one using WB features extracted by the method described above (the result will be shown later in this chapter). Their results are comparable. Based on the above considerations, we believe that the current design of subband structure for feature extraction is appropriate for the intended goal of investigation.

Number of Cepstral Coefficients for Each Subband

The physical meaning of cepstral analysis has been explained in Chapter 2. A more detailed spectral envelope can be reconstructed by using more cepstral coefficients. If

all cepstral coefficients are used, the reconstructed spectrum will be exactly the same as the original one. For each subband, the number of cepstral coefficients used is based on how well the spectral envelope can be represented by using only portions of cepstral coefficients for that subband.

We have discussed the reason why NB is divided into four subbands before. Another reason is suggested here. In attaining similar reconstructed spectral envelope, the number of cepstral coefficients required for a single band is more than that for subbands partitioned from a single band. An example is illustrated in Figure 4-5. As shown in Figure 4-3, the frequency band that is between 0 and 2 kHz is divided into two subbands for feature extraction. In this example, spectral envelope features for that frequency band are found in a single band. In the former case, 10 cepstral coefficients are used to reconstruct the spectral envelope for the frequency band of 0 – 2 kHz (Figure 4-5). However, if feature extraction is performed in the latter case, 12 cepstral coefficients are needed to attain similar reconstructed spectral envelope.

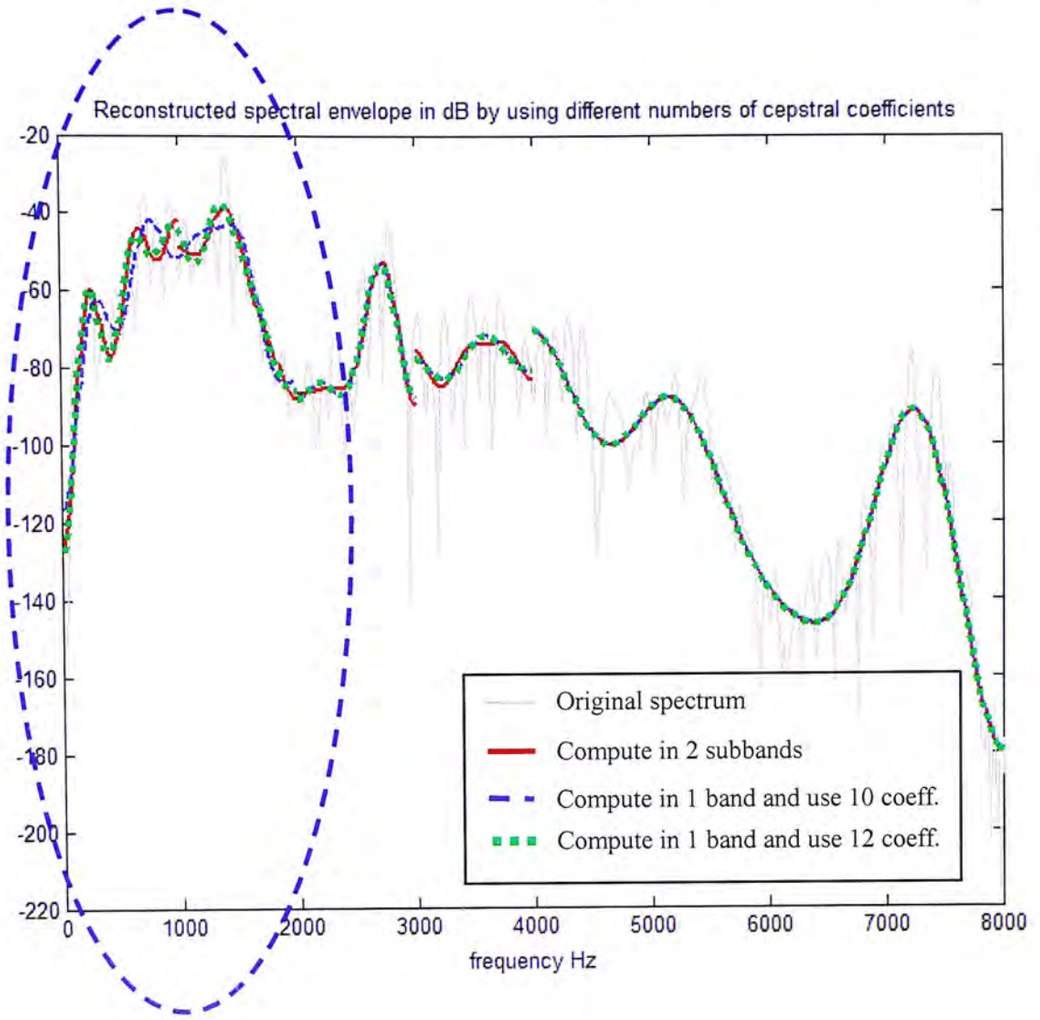


Figure 4-5 An example illustrates that more cepstral coefficients are required if feature extraction is performed in a single band (0 – 2 kHz)

4.3. Feature Extraction Procedures

The process of extracting subband spectral envelope features is roughly shown as in Figure 4-1. Details of the key steps are given below.

The speech signal is segmented into frames by a 20 ms window progressing at a 10 ms frame rate. Each frame of speech is pre-emphasized with a coefficient of 0.97. Then, each speech frame is multiplied with a Hamming window and followed by 1024-point Fast Fourier Transform (FFT). The power spectrum is obtained by the magnitude square of the FFT values.

Different speakers, or even the same speaker, may speak in different loudness in different occasions. As spectral envelope features are extracted directly from the

energy spectrum, the loudness may affect the feature value. Later, the feature distribution will be estimated during speaker model training. Therefore, feature values from different utterances should be kept in similar range and energy normalization is required.

Speaker may speak in different loudness across the whole utterance. Also, spectral envelope features in word level are studied in this work. Therefore, energy normalization is performed on a token basis, i.e. normalized with respect to each digit segment. The average of frame energy over the digit segment is found. Then, magnitudes of the power spectrum are normalized by this value.

After energy normalization, logarithm of the normalized power spectral magnitudes is found. Then DCT is performed in each subband as,

$$C_i(k,l) = \frac{1}{I_i}\sum_{m=1}^{I_i} \log(W(m + \sum_{j=1}^{i-1} I_j, l)) \cos\left(\frac{k\pi}{I_i}(m - \frac{1}{2})\right) \qquad (4.1)$$

$$k = 0,1,..., (I_i - 1),\ i = 1, 2, ..., 5$$

where $W(m,l)$ is the normalized power spectral magnitude at frequency point m of frame l, I_i is the total number of frequency points in the i^{th} subband and $C_i(k,l)$ is the k^{th} cepstral coefficient in the i^{th} subband of frame l.

Different numbers of cepstral coefficients used in each subband have been tried. Finally, it is determined that for the first four subbands, the 0^{th} to 4^{th} cepstral coefficients will be used (i.e. $k = 0 - 4$). In the fifth subband, the 0^{th} to 7^{th} cepstral coefficients will be used. An example of reconstructed spectral envelope is shown in Figure 4-6.

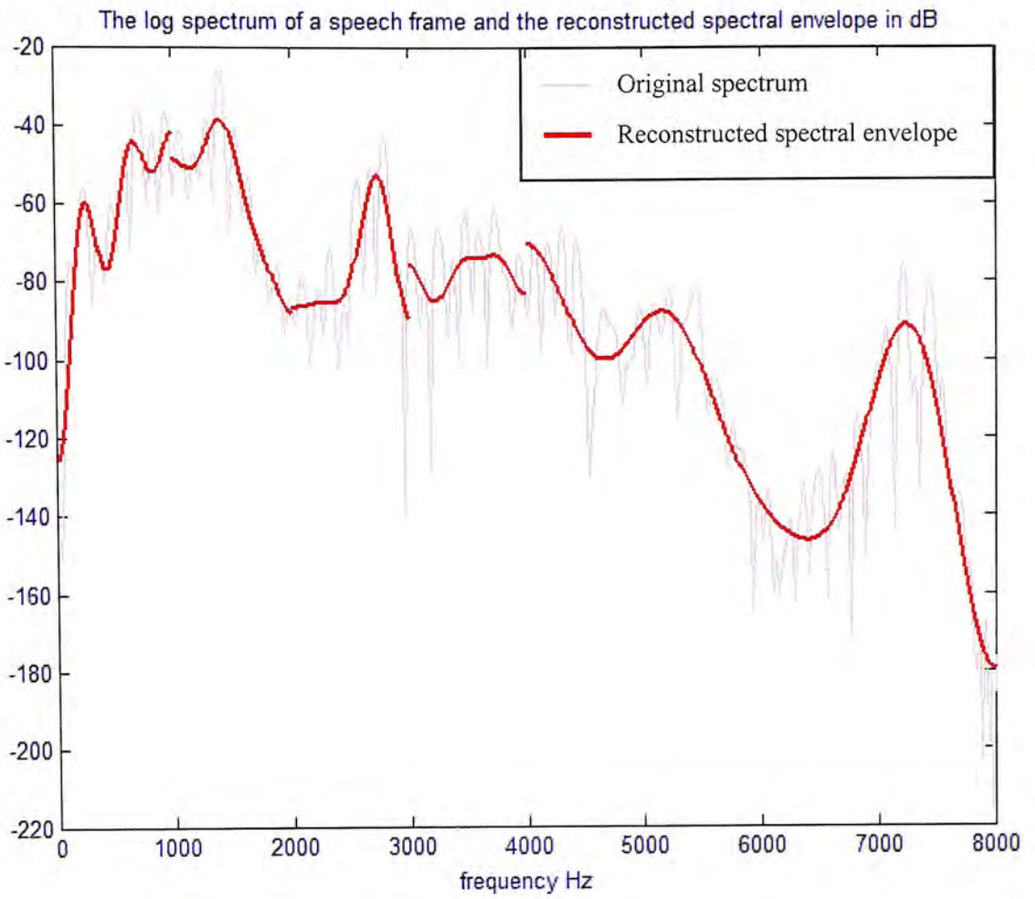


Figure 4-6 An example of reconstructed spectral envelope, in comparison with the original spectrum

The feature vectors extracted from WB, NB and HB are composed as shown in Figure 4-7. The total number of coefficients for features extracted from WB, NB and HB are equal to 28, 20 and 8 respectively.

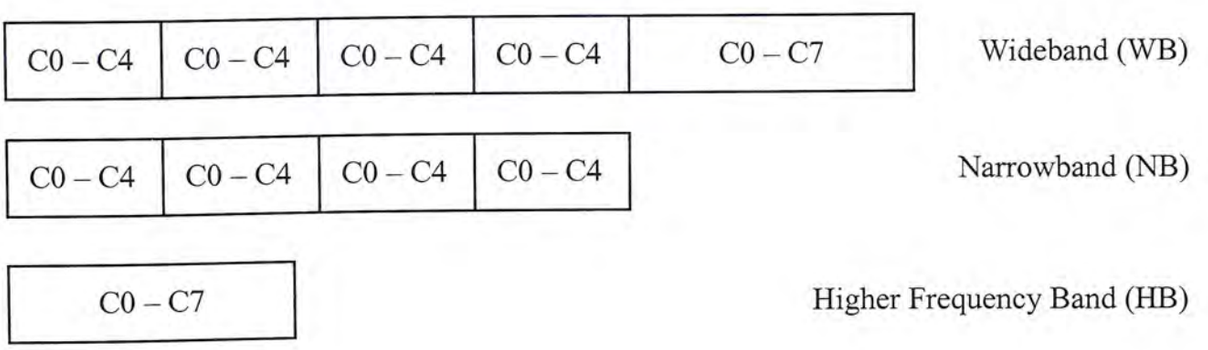


Figure 4-7 Layout of feature vector

4.4. SID Experiments

4.4.1. Experimental Set-up

Spectral envelope features in WB, NB and HB are extracted using the method described in the previous section. Speaker identification (SID) experiments are performed using the features extracted from these three bands individually. The step is the same as the baseline system (Digit Baseline 1) in Chapter 3. 24 Gaussian mixtures will be used in each speaker model throughout the study.

4.4.2. Results and Analysis

Table 4-1 lists the SID results using NB, HB and WB features. Our goal is to study the contribution from NB and HB in speaker recognition. This can be done by comparing the SID results of NB and HB features with the one from WB. In Chapter 3, a baseline system using MFCC (Digit Baseline 1) has been built. The identification result using WB features is first compared with that in Digit Baseline 1. Using the same number of Gaussian mixtures, their results are comparable. The overall SID accuracy differs by less than 1 %. The result with WB features serves as another baseline system for our study. For convenience, it is named as Digit Baseline 2.

From Table 4-1, we can find that WB features outperformed the NB features. It confirms that there is indeed important speaker-specific information outside the narrowband region. To have a more detailed study on the contribution of NB and HB features, these two sets of results are compared with Digit Baseline 2 individually as shown in Table 4-2 and Table 4-3.

	Overall SID rate (%)		
Digit	NB	WB	HB
0	94.83	97.72	73.94
1	90.35	94.99	73.96
2	86.31	95.61	75.91
3	96.72	98.22	86.21
4	94.69	98.89	88.19
5	84.81	87.92	57.07
6	87.16	92.64	63.77
7	91.96	97.15	83.34
8	90.91	94.83	72
9	94.49	96.91	76.36
Average	91.22	95.49	75.08

Table 4-1 Overall SID rate (%) of using spectral envelope features from NB, HB and WB for the 10 Cantonese digits

Digit	Difference of overall SID rates between WB and NB systems (%)	Rank of difference in ascending order
0	2.89	3
1	4.64	7
2	9.3	10
3	1.5	1
4	4.2	6
5	3.11	4
6	5.48	9
7	5.19	8
8	3.92	5
9	2.42	2

Table 4-2 Compare identification results of using WB and NB features

Digit	Difference of overall SID rates between WB and HB systems (%)	Rank of difference in ascending order
0	23.78	8
1	21.03	6
2	19.7	4
3	12.01	2
4	10.7	1
5	30.85	10
6	28.87	9
7	13.81	3
8	22.83	7
9	20.55	5

Table 4-3 Compare identification results of using WB and HB features

By comparing the identification results between using WB features and the NB ones, contribution from HB features can be observed. In Table 4-2, we have ranked the difference of overall SID rates between WB and NB systems on digit basis. The higher ranking the digit obtains, the more relatively important the HB features are. We found that digit '2', '6' and '7' attain the greatest improvement when HB features are used.

Similarly, we can make comparison between the results of using WB features and the HB ones. Contribution from NB features can be observed. From Table 4-3, we can find that digit '5', '6' and '0' show the greatest improvement when NB features are used.

The above observations indicate that different digits rely on the features from different frequency bands for speaker recognition. This is probably related to the different phonetic composition of the digits. Before further discussion on this issue, a brief introduction of Cantonese speech will be given first. For more details, please refer to [2].

Cantonese is a monosyllabic and tonal language [2]. Each Chinese character is pronounced as a single syllable that carries a specific tone. A character may have many pronunciations and a syllable typically corresponds to a number of different characters [3]. A Cantonese syllable is divided into the *Initial* part and the *Final* part. Initials and Finals are composed by phonemes, which concern the manners of articulation.

Phonemes can be classified into four categories: vowels, diphthongs, semi-vowels and consonants. Vowels and diphthongs belong to the category of voiced speech. To produce these speech sounds, vocal cord vibrates periodically to generate quasi-periodic air pulses along the vocal tract. Semi-vowels are described as transitional, vowel-like sounds and are similar in nature to the vowels and diphthongs [1]. For unvoiced sounds, the vocal cord does not vibrate and it is generated by a turbulent flow of air at constriction in the vocal tract and sudden flow of air under the control of some articulators [1]. Consonants can be further divided into four types: fricatives, affricates, nasals and stops. They are classified by their places and manners of articulation.

The phonetic composition of the 10 Cantonese digits using phonetic symbols proposed by the Linguistic Society of Hong Kong (LSHK) [4] and International Phonetic Association (IPA) are listed in Table 4-4. The phonetic features of each component are also given.

Digit	LSHK	IPA	Phonetic components
0	ling4	lɪŋ	Liquid + Vowel + Nasal
1	jat1	jɐt	Glide + Vowel + Stop
2	ji6	ji	Glide + Vowel
3	saam1	sam	Fricative + Vowel + Nasal
4	sei3	sei	Fricative + Diphthong
5	ng5	ŋ	Nasal
6	luk6	lʊk	Liquid + Vowel + Stop
7	cat1	$ts^hɐt$	Affricate + Vowel + Stop
8	baat3	pat	Stop + Vowel + Stop
9	gau2	kɐu	Stop + Diphthong

Table 4-4 Phonetic transcriptions of the 10 Cantonese digits using the LSHK scheme [4] and the IPA scheme

Different phonetic units have different acoustic properties. Since the feature parameters are extracted from a particular digit, the recognition performance is inevitable related to the phonetic composition of the digit.

We expect that features extracted from frequency band with higher energy concentration show more relatively importance in speaker recognition. Some phonemes have more energy concentrated at lower frequency band while other phonemes have more energy concentrated at higher frequency. Therefore, the contribution of features from NB and HB in speaker recognition is text-dependent.

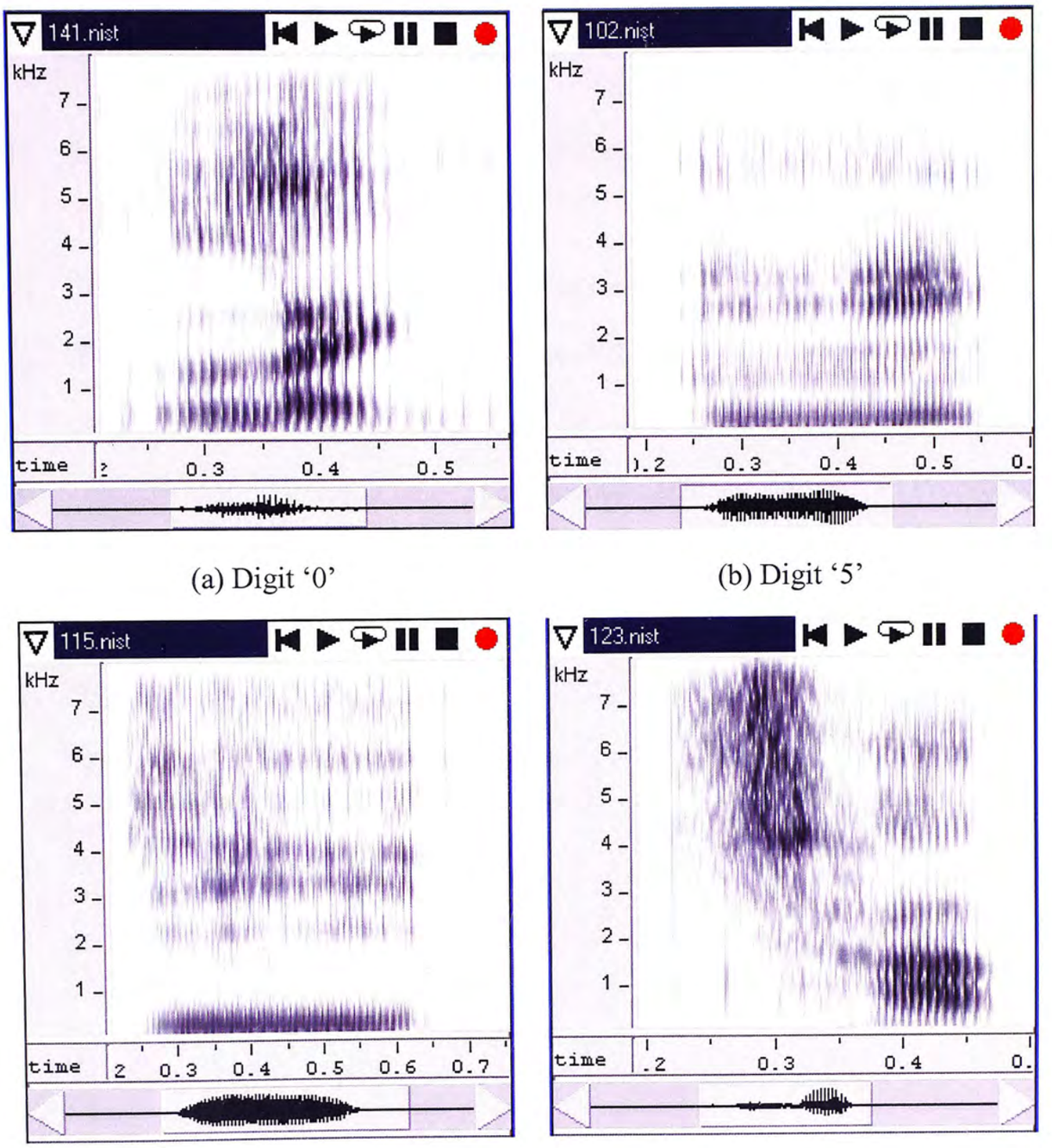
For example, /ng/ is a nasal consonant. It is produced with the velum lowered so that air flows through the nasal tract. Hence, sound is radiated at the nostrils. Acoustically, there is a concentration of low-frequency energy in the speech signal. /ng/ is the core part of digit '5'. This indicates that NB features are relatively important for digit '5' in SID. We can also observe this from the comparison of the SID results in WB and HB systems (Table 4-3).

Vowels are produced with vibration of the vocal cord. The way that the

cross-sectional area varies along the vocal tract determines different vowel sounds. For example, /i/ is a vowel with the position and height of tongue hump are in the front and high respectively, where the tongue hump is the mass of the tongue at its narrowest constriction within the vocal tract [1]. The vowel sound produced in this way has high-frequency resonance. It means there is a relatively high concentration of energy in the higher frequency. This vowel is one of the components of digit '2'. It implies that HB features are relatively important for digit '2' in SID. By comparing the SID results between WB and NB systems, we can also get this observation (Table 4-2).

Figure 4-8 shows examples of spectrograms for some Cantonese digits. Spectrogram is a three-dimensional plot of the signal intensity in different frequency bands over time. Darker color at particular frequency band implies that there is higher energy in that area. In Figure 4-8, we can see that different digits have different concentration of energy in NB and HB. Part (a) and (b) of Figure 4-8 show that digit '0' and '5' have higher concentration of energy at the narrowband, while part (c) and (d) show that digit '2' and '7' have higher concentration of energy at the higher frequency band. It is generally coherent with the results listed in Table 4-2 and Table 4-3.

To conclude, higher frequency band does contain important features of a speaker's voice. Contributions of features from NB and HB in speaker recognition are text-dependent.



(a) Digit '0'

(b) Digit '5'

(c) Digit '2'

(d) Digit '7'

Figure 4-8 Examples of spectrograms for selected Cantonese digits (a) Digit '0'; (b) Digit '5'; (c) Digit '2'; (d) Digit '7'

References

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

[2] W. K. Lo, "Cantonese phonology and phonetics: an engineering introduction", *Internal Documentation, DSP and Speech Technology Laboratory, Department of Electronic Engineering, the Chinese University of Hong Kong*, 2000.

[3] Tan Lee et al, "Modeling tones in continuous Cantonese speech", in *Proceedings of International Conference on Spoken Language Processing*, vol. 4, pp. 2401 – 2404, 2002.

[4] Linguistic Society of Hong Kong (LSHK), *Hong Kong Jyut Ping Characters Table* (粵語拼音字表), Linguistic Society of Hong Kong Press (香港語言學會出版), 1997.

Chapter 5

Fusion of Subband Features

In Chapter 4, it was seen that both NB and HB contain useful information for SID. This importance is text-dependent. Instead of simply lumping the HB and NB features together for SID, features from these two bands can be fused properly based on the results in Chapter 4. In this chapter, fusion performed at model level and feature level will be investigated.

5.1. Model Level Fusion

To use two different types of features to make a joint decision, one of the common ways is to fuse them at the model level. It means that classification of the two sources of information is performed separately and the classification results are combined in a proper way, such as [1]. Fusing of NB and HB features for SID at model level will be investigated in this section.

5.1.1. Experimental Set-up

Figure 5-1 shows the steps of fusing features from NB and HB at model level. Given the input speech, features from NB and HB are extracted. The corresponding likelihood scores from the digit-specific speaker models are calculated. Since the dimensions of feature vectors from NB and HB are different, the dynamic range of the

likelihood scores from these two systems would be different. Before the scores are combined, they are normalized by equation (5.1). With input utterance b,

$$Savg_{a,b} = \frac{1}{20}\sum_{j=1}^{20}\left|S_{a,b,j}\right|, \; a = \text{HB or NB}$$

$$\tilde{S}_{a,b,j} = \frac{S_{a,b,j}}{Savg_{a,b}}, \; \text{j} = 1, ..., 20 \tag{5.1}$$

$S_{a,b,j}$ denotes the likelihood score computed from speaker model j using NB features when symbol a equals to NB. $Savg_{a,b}$ is the average of the absolute values of the scores from the 20 speaker models. This value is then used to normalize $S_{a,b,j}$ to give $\tilde{S}_{a,b,j}$.

Scores that are given by speaker model j using NB and HB features are linearly combined as follows:

$$S_{HB+NB,b,j} = \alpha\tilde{S}_{NB,b,j} + (1-\alpha)\tilde{S}_{HB,b,j}, \; \text{j} = 1, \ldots, 20 \tag{5.2}$$

Speaker model that gives the maximum value of $S_{HB+NB,b,j}$ among the 20 speakers would be recognized as the speaker of the input utterance. Different values of α have been tried and the results are given in Table 5-1.

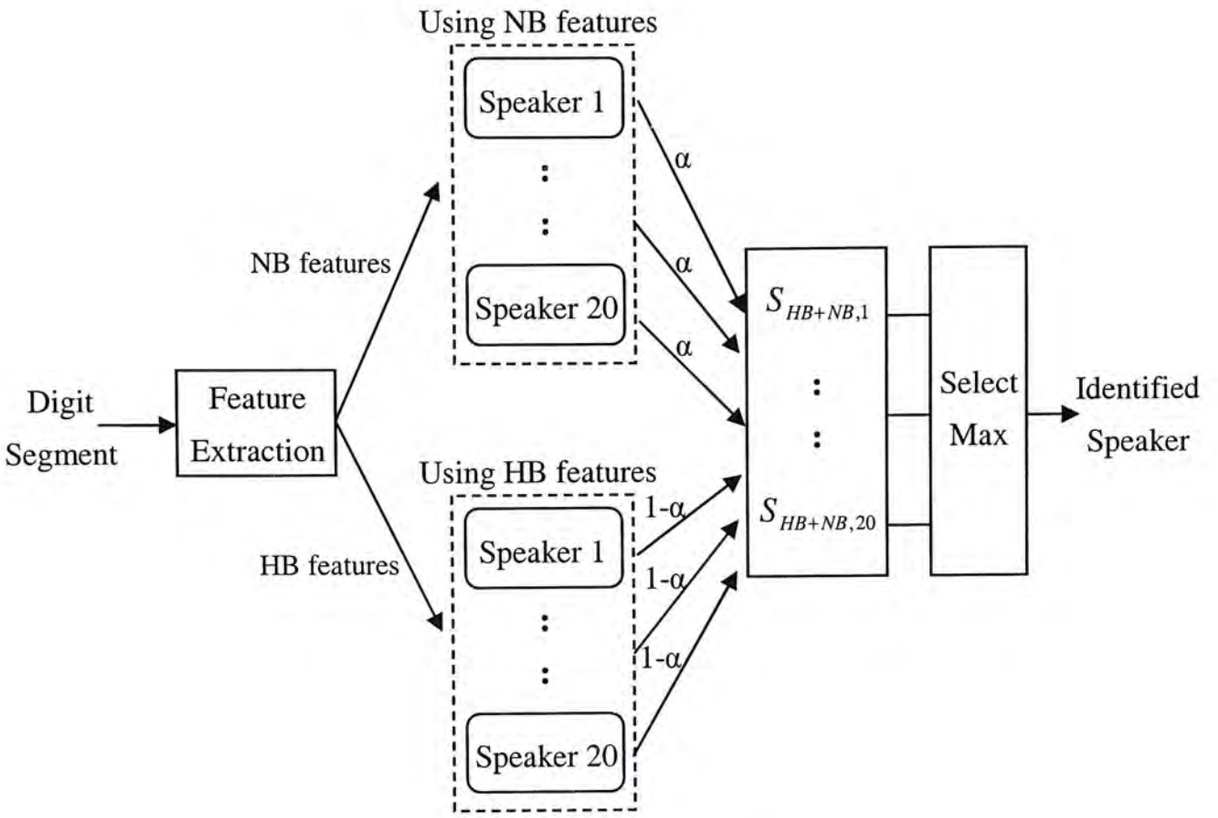


Figure 5-1 Steps of fusing NB and HB features at model level

5.1.2. Results and Analysis

Table 5-1 shows the SID rate by fusing features from NB and HB at model level with different values of α. A summary of the results is given in Table 5-2, which lists the maximum recognition rate given by model level fusion for the corresponding value of α.

	Overall SID rate (%)									
		Values of α								
Digit	WB	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
0	97.72	96.66	97	96.66	96.02	94.74	92.77	89.82	85.65	79.75
1	94.99	93.21	94.16	94.58	94.44	93.66	91.68	88.79	85.09	79.92
2	95.61	90.71	93.1	94.3	94.49	94.33	92.82	90.76	87.34	82.28
3	98.22	97.36	97.94	98.55	98.78	98.47	97.72	95.94	94.33	90.88
4	98.89	97.05	97.92	98.22	98.33	98	97.47	96.28	94.58	91.91
5	87.92	88.26	89.09	88.26	86.28	83.97	80.33	75.51	70.34	64.47
6	92.64	88.69	89.48	88.6	87.33	84.74	81.86	78.56	73.9	69.37
7	97.15	95.23	96.49	96.57	96.39	95.62	94.17	92.46	90.11	87.08
8	94.83	93.52	94.41	95.22	94.83	93.58	91.52	88.38	84.29	78.62
9	96.91	95.91	96.66	96.69	96.47	95.44	93.8	91.07	87.79	82.45
Average	95.49	93.66	94.63	94.77	94.34	93.26	91.41	88.76	85.34	80.67

Table 5-1 Overall SID rate (%) of fusing NB and HB features at model level with different values of α

Digit	Value of α	Overall SID rate (%)	
		Max. SID rate given by model level fusion	Digit Baseline 2
0	0.8	97	97.72
1	0.7	94.58	94.99
2	0.6	94.49	95.61
3	0.6	98.78	98.22
4	0.6	98.33	98.89
5	0.8	89.09	87.92
6	0.8	89.48	92.64
7	0.7	96.57	97.15
8	0.7	95.22	94.83
9	0.7	96.69	96.91
	Average	95.02	95.49

Table 5-2 Summarize the result of fusing features at model level from Table 5-1

Likelihood scores that are computed by speaker models in NB and HB systems have been normalized to the same range. Larger value of α implies that NB features

tend to play a more important role than the HB ones in determining the speaker's identity. We find that the best performance is given when α ranges between 0.6 and 0.8. Therefore, it shows that NB features are generally more important than that from HB.

The results of Digit Baseline 2 are also given in Table 5-2 for the ease of comparison. Only in the cases of digit '3', '5' and '8', using model level fusion outperform Digit Baseline 2 with improvement of 0.56%, 1.17% and 0.39% respectively. But on average, Digit Baseline 2 still gives the best performance. Fusing features at model level cannot perform as good as that using WB features. The possible reasons will be discussed later.

Speaker of the input speech is called true speaker. If likelihood scores from the true speaker model in the NB and HB systems are not the maximum, the overall score after linear combination may not be the maximum too. If this is the case, a recognition error is caused. Fusing features at model level mainly deals with these cases. It works by combining the scores from the HB and NB systems with proper weights such that the final score from the true speaker model becomes the largest. However, the result showed that this method does not perform as good as expected.

The likelihood scores of input speech computed from the 20 speaker models in the NB and HB systems are sorted and analyzed. The likelihood scores computed by model level fusion with the corresponding values of α that gave the highest identification rate are also analyzed.

The number of test utterances that do not have the maximum likelihood scores from the true speaker model in the NB and HB systems is counted ((1) in Table 5-3). On average, it accounts for 30% of the total test utterances. They are the targets that we hope to apply model level fusion to improve recognition performance. The number of test utterances that have the maximum likelihood scores from the true

speaker model after model level fusion is counted ((2) in Table 5-3). There are 85% of those counted in (1). On the other hand, (3) and (4) in Table 5-3 are the number of utterances with maximum likelihood scores from the true speaker model only after model level fusion, and that only in wideband system respectively. We can see that except digit '3', '5' and '8', the value listed in (3) is less than the corresponding value in (4) for other digits. That explains why only digit '3', '5' and '8' have performance gain by using model level fusion, compared with the result of Digit Baseline 2.

Digit	Total number of test utterances	(1)	%	(2)	% of (2) in (1)	(3)	% of (3) in (2)	(4)	% of (4) in (1)
0	3595	1033	28.73	925	89.55	11	1.19	39	3.78
1	3595	1140	31.71	945	82.89	35	3.70	56	4.91
2	3595	1182	32.88	984	83.25	41	4.17	87	7.36
3	3596	569	15.82	525	92.27	34	6.48	14	2.46
4	3598	577	16.04	517	89.60	9	1.74	29	5.03
5	3594	1714	47.69	1322	77.13	95	7.19	63	3.68
6	3395	1406	41.41	1049	74.61	31	2.96	144	10.24
7	3793	845	22.28	715	84.62	34	4.76	59	6.98
8	3596	1185	32.95	1013	85.49	54	5.33	48	4.05
9	3596	953	26.50	834	87.51	19	2.28	30	3.15

Table 5-3 Analyze the performance of fusing features at model level

Two reasons are suggested to explain why model level fusion cannot perform as good as that in WB system.

It is difficult to find a set of text-dependent weights such that the combined scores from the true speaker model become the largest in all cases. Therefore, the effectiveness of this approach is limited.

Similar approach of model level fusion was studied for large vocabulary word

recognition [2]. To get the maximum benefit, [2] suggested that the features used should be independent and they should provide similar recognition performance when they are used independently. In this experiment, we assumed that the NB and HB features are independent to each other and they are modeled individually. When α is equal to 0.5, the scores from NB and HB systems are combined in equal weight and the average SID rate is equal to 93.26%. The experimental result showed that Digit Baseline 2 gave better performance (average SID rate = 95.49%). It indicated that there is correlation between NB and HB features and the correlation is also useful in speaker recognition. Similar conclusions about the correlation between features from different subbands have been obtained in [3] [4]. Therefore, fusing features at model level cannot give much improvement. The correlation between NB and HB features should be retained to use for speaker recognition. It comes with the idea of fusing features at feature level that will be investigated in the next section.

5.2. Feature Level Fusion

In the computation of the likelihood score from the Gaussian mixture component, each element of the feature vector is attributed to the same weight. Recently, an approach of weighting feature components was investigated for robust speech recognition ([5] – [7]). It comes with the idea that some feature components are less corrupted by noise. A heavier weight should be given to those components to show their relative importance in speech recognition. Similarly, this approach can be applied to speaker recognition if some of the feature components are more discriminative than the others. For example, [8] studied the method of adjusting individual weight for each feature component to achieve minimum error rate.

In Chapter 4, we find that the importance of features from NB and HB is

text-dependent in speaker recognition. It motivates us to fuse features from these two bands with text-dependent feature weights.

5.2.1. Experimental Set-up

In this experiment, spectral envelope features extracted from WB are used. The experimental procedures are the same as the baseline system (Digit Baseline 1) in Chapter 3. However, the computation of likelihood score is different.

Traditionally the log-likelihood score for an input feature vector f^{NH} is computed by equation (5.3),

$$D(\theta^{NH}, f^{NH}) = c(\Sigma^{NH}) - \frac{1}{2}(f^{NH} - \mu^{NH})^T \Sigma^{NH^{-1}}(f^{NH} - \mu^{NH}) \qquad (5.3)$$

where $\theta^{NH} = \{\mu^{NH}, \Sigma^{NH}\}$ denote the parameters of the Gaussian mixture component. μ^{NH} and Σ^{NH} represent the mean vector and diagonal covariance matrix of the Gaussian distribution respectively. $c(\Sigma^{NH})$ is a constant. The superscript *NH* denotes WB features. f^{NH} is a 28-dimension vector and consists of two parts as,

$$f^{NH} = [f^N \quad f^H]$$

where f^N and f^H are a 20-dimension and a 8-dimension vectors respectively. The superscript *N* and *H* denote features from NB and HB.

To incorporate different weights for NB and HB, the likelihood score would be computed as,

$$D(\theta^{NH}, f^{NH}) = c(\Sigma^{NH}) - \frac{1}{2}(f^{NH} - \mu^{NH})^T W \Sigma^{NH^{-1}}(f^{NH} - \mu^{NH}) \qquad (5.4)$$

where W is a weight matrix and it is a diagonal matrix with the following structure:

$$W = diag[\underbrace{\alpha,...,\alpha}_{20}, \underbrace{\beta,...,\beta}_{8}]$$

α and β are adjusted to reflect the relative importance of features from NB and HB respectively in speaker recognition with the following constraint:

$$20\alpha + 8\beta = 28 \tag{5.5}$$

Different sets of α and β that satisfy equation (5.5) have been tried in the experiment. Notice that when both α and β equal to 1, equation (5.4) will become equation (5.3). It will give the result of Digit Baseline 2.

5.2.2. Results and Analysis

Table 5-4 gives the results of feature level fusion by using different values of α and β. The highlighted values are the best SID rates attained for different digits. Based on these results, α and β are adjusted in a finer step for higher SID rate.

Overall SID rate (%)

α	β	Digit									
		0	1	2	3	4	5	6	7	8	9
1.3	0.25	96.38	92.35	90.99	97.44	97.8	85.92	89.01	95.04	92.55	95.91
1.2	0.5	97.3	94.16	93.55	97.8	98.5	87.67	90.93	96.39	93.91	96.64
1.1	0.75	97.69	94.6	94.8	98.08	98.86	88.26	92.05	96.78	94.52	96.89
0.9	1.25	97.69	94.63	95.66	98.39	98.78	87.76	92.72	96.89	94.63	96.77
0.8	1.5	97.52	94.24	95.47	98.3	98.67	87.12	92.16	96.63	94.02	96.41
0.7	1.75	97.13	93.8	94.94	98.22	98.17	86.31	91.34	96.31	93.3	95.8
0.6	2	96.44	92.32	94.33	97.78	97.83	85.25	89.96	95.62	92.16	94.88
Digit Baseline 2		97.72	94.99	95.61	98.22	98.89	87.92	92.64	97.15	94.83	96.91

Table 5-4 Overall SID rate (%) of feature level fusion by using different sets of α and β

α and β are chosen by the following rules. If the SID rate given by feature level fusion is larger than that in Digit Baseline 2, the corresponding values of α and β are chosen. If there is more than one set of α and β that gives maximum SID rate, these values will be further adjusted in a finer step. On the other hand, if the SID rate given

by feature level fusion is less than or equal to that in Digit Baseline 2, both α and β are chosen to equal to 1. The chosen values of α and β and the corresponding SID rate for each digit are summarized in Table 5-5.

Digit	α	β	Overall SID rate (%)	
			Max. SID rate given by feature level fusion	Digit Baseline 2
0	1	1	97.72	97.72
1	0.98	1.05	95.02	94.99
2	0.95	1.125	95.74	95.61
3	0.883	1.2925	98.41	98.22
4	1	1	98.89	98.89
5	1.1	0.75	88.26	87.92
6	0.98	1.05	92.81	92.64
7	1	1	97.15	97.15
8	1	1	94.83	94.83
9	0.96	1.1	96.94	96.91
	Average		95.58	95.49

Table 5-5 List of chosen values of α and β and the corresponding SID rate given by feature level fusion

From Table 5-5, we can find that the average improvement given by feature level fusion is equal to 0.09%, compared with that of Digit Baseline 2. By examining the SID result for individual digits, only digit '1', '2', '3', '5', '6' and '9' have improvement in performance. Also, we can see that the chosen values of α and β are approximately equal to 1 for all digits.

To investigate if α and β depend on the number of feature components used in each subband, another experiment is performed. The experimental set-up is same as before but the number of feature components used in NB and HB are different (named as Feature Set 2). In NB, the 0^{th} to 3^{rd} cepstral coefficients from each subband are

used. In HB, the 0^{th} to 11^{th} cepstral coefficients are used. The result is summarized in Table 5-6.

Digit	α	β	Overall SID rate (%)	
			Max. SID rate given by feature level fusion	Given by using equal feature weights
0	1.09	0.88	97.91	97.8
1	1.06	0.92	94.94	94.91
2	1	1	95.49	95.49
3	0.85	1.2	98.22	98.03
4	1.09	0.88	99.11	99.08
5	1	1	87.95	87.95
6	1.06	0.92	92.19	92.11
7	1	1	96.84	96.84
8	1.03	0.96	95.08	95.02
9	1.015	0.98	97.08	97.05
	Average		95.48	95.43

Table 5-6 List of chosen values of α and β and the corresponding SID rate given by feature level fusion using Feature Set 2

From Table 5-6, we can find that the average improvement given by feature level fusion using Feature Set 2 is equal to 0.05%. The chosen values of α and β are approximately equal to 1 for all digits. These two sets of results imply that the relative importance of NB and HB features cannot be varied by adjusting α and β. Therefore, the approach of applying text-dependent feature weights cannot give much improvement.

5.3. Discussion

Results in Chapter 4 showed that the importance of NB and HB features in speaker recognition are text-dependent. Fusing features from these two bands at model level

and feature level have been investigated. They differ at which level the information of importance of these two bands is applied. When features are fused at model level, the importance of NB and HB features is incorporated during linear combination of the likelihood scores from these two bands. In feature level fusion, this information is applied to the feature components with subband feature weights during calculation of likelihood score.

For fusing features at model level, only some digits have improvement in performance. The overall performance is worse than that in Digit Baseline 2. On the other hand, feature level fusion can give performance as good as that in equal weight (i.e. Digit Baseline 2) while for some digits, they can have further improvement. The average performance gain given by feature level fusion is equal to 0.09%.

One of the direct ways to utilize different features is to fuse them at model level. To maximize the benefit of model level fusion, data streams should be statistically independent [2]. However, it is not the case for features extracted from frequency bands and there is correlation between them [3]. Therefore, it is not suitable to perform model level fusion. For the approach of feature level fusion, the statistical dependencies between the two data streams can be retained. However, the improvement given by this approach is small. It is suggested that adjusting feature weights α and β cannot affect the contribution of NB and HB features in speaker recognition.

It is believed that fusing features from NB and HB properly to reflect their relative importance provides a possible way to improve recognition performance. Therefore, it is required to investigate other approaches of fusing features from these two bands in future.

References

[1] R. Kober, U. Harz and J. Schiffers, "Fusion of visual and acoustic signals for command-word recognition", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 1997, vol. 2, pp. 1495 – 1497.

[2] V. N. Gupta, M. Lennig and P. Mermelstein, "Integration of acoustic information in a large vocabulary word recognizer", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing,* 1987, pp. 697 – 700.

[3] L. Besacier, J.F. Bonastre, "Subband approach for automatic speaker recognition: optimal division of the frequency domain", in *Proceedings of the Audio and Video based Biometric Person Authentication*, 1997, pp. 195 – 202.

[4] L. Besacier, J.F. Bonastre, "Subband architecture for automatic speaker recognition", *Signal Processing*, vol. 80, pp. 1245 – 1259, 2000.

[5] K. C. Huang and Y. T. Juang, "Feature weighting in noisy speech recognition", *Electronic Letters*, vol. 39, issue 12, pp. 938 – 939, 2003.

[6] T. Xu and Z. Cao, "Combination of feature weight and speech enhancement for robust ASR at low SNRs", in *Proceedings of the 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, 2002, vol. 1, pp. 441 – 444.

[7] J. Hernando, "Maximum likelihood weighting of dynamic speech features for CDHMM speech recognition", in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, 1997, vol. 2, pp. 1267 – 1270.

[8] H. Ney and R. Gierloff, "Speaker recognition using a feature weighting techniques", in *Proceedings of International Conference on Acoustics, Speech,*

and Signal Processing, 1982, vol. 7, pp. 1645 – 1648.

Chapter 6

Utterance-Level SID with Text-Dependent Weights

It was shown that, for different words, the contributions from subband features might be different. This chapter describes the use of subband spectral envelope features for SID decision on a complete utterance which consists of a sequence of digits.

6.1. Motivation

In previous chapters, subband spectral envelope features for text-dependent SID has been studied. Based on the features extracted from a digit segment, the speaker's identity is decided. However, in practical applications, identification decision is made for the entire utterance. An utterance may consist of one or more digits. SID at utterance level means that the classification results from each digit segment are combined to determine a speaker's identity.

From the results in Chapter 4, we find that some words can distinguish speakers better than the others for all types of features, i.e. MFCC, or spectral envelope features. It motivates us to apply text-dependent weights in utterance-level SID. Heavier weight will be assigned to those words that have higher text-dependent SID rates.

In Chapter 4, the use of features from NB and HB in text-dependent SID has been studied. It was found that their importance is text-dependent. Based on this result,

fusing features from these two bands with text-dependent weights has been investigated in Chapter 5. Recognition performance can be improved. Therefore, text-dependent subband feature weights used in utterance-level SID will be studied.

6.2. Utterance-Level SID

In utterance-level SID, the likelihood scores for individual digits spoken in the utterance are linearly combined. Decision is made based on the overall score of the utterance. Figure 6-1 shows the steps of utterance-level SID.

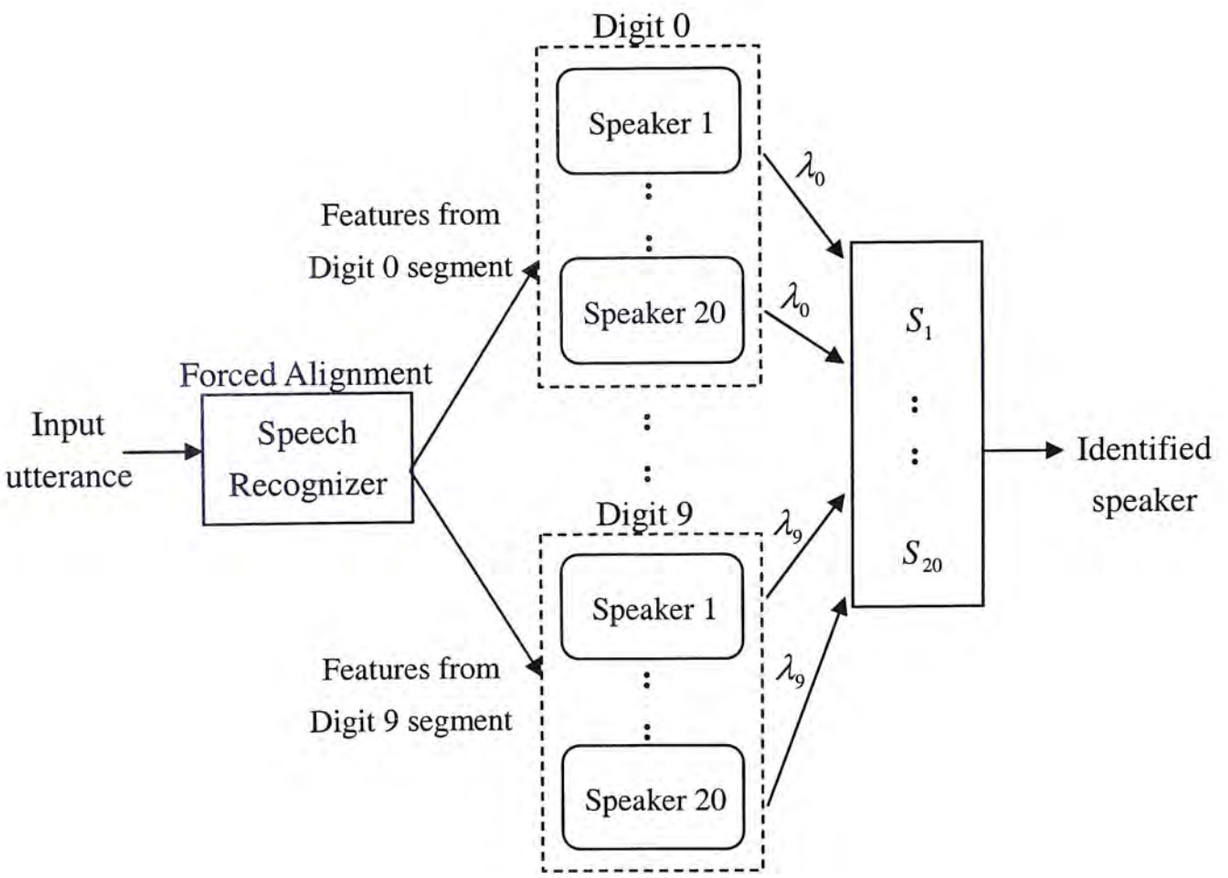


Figure 6-1 Steps of utterance-level SID

It is assumed that the content of the input utterance is known. A speech recognizer is used to find out the duration of each digit spoken in the utterance. After

that, features extracted from a particular digit segment are used to compute the likelihood scores from the digit-specific speaker models. The overall score of the utterance is given by linear combination of scores of all individual digits. For example, if the input utterance is '06', the overall score of the utterance is computed as follow:

$$S_j = \lambda_0 S_{0,j} + \lambda_6 S_{6,j} \text{ , j=1,..., 20} \quad (6.1)$$

$S_{0,j}$ and $S_{6,j}$ are the likelihood scores computed from the j^{th} speaker model for digit '0' and digit '6' respectively. λ_0 denotes the text-dependent weight for digit '0'. S_j is the overall likelihood score from the j^{th} speaker model. The model that gives the highest likelihood score will be identified as the recognized speaker.

6.3. Baseline System

In the baseline system, the digit-level likelihood score is computed with equal weights for features from different subbands and the scores from different digits are also equally weighted to generate the utterance-level decision.

6.3.1. Implementation Details

The baseline system follows the steps shown in Figure 6-1. The speech recognizer used in the experiment is the same as the one used in Chapter 3. Text-dependent weights λ are equal to 1 for all digits.

Two baseline systems have been established and evaluated. Utterance Baseline 1 uses MFCC features as described in Chapter 2 and Utterance Baseline 2 uses spectral envelope features extracted from WB as described in Chapter 4.

Among the 12 sessions of recordings, the first six sessions (S01 – S06) are used

to train digit-specific speaker models. The last four sessions (S09 – S12) are used to evaluate the system (see Table 6 –1).

Data	Use
S01 – S06	Train digit-specific speaker models
S07 – S08	Use in other experiments in this chapter
S09 – S12	Evaluate the system

Table 6-1 Use of data in Chapter 6

6.3.2. Results and Analysis

By examining the identification errors, it is found that over 50 % of data in one particular recording session from a speaker (M10) caused errors in the baseline system. The voice of speaker M10 in that session is significantly different from the other sessions. It might be due to sickness. The intra-speaker change is exceptionally large for this speaker. If data from this speaker is used for evaluation, it would affect the effectiveness of the analysis of identification results. Therefore, this speaker is excluded from the remaining experiments in this chapter.

The results of the baseline systems are listed in Table 6-2. The identification errors have been analyzed. For single-digit and multi-digit utterances, the number of identification errors are counted separately (Table 6-2).

Feature	Utterance level SID rate (%)	Number of identification errors	
		Single-digit utterances	Multi-digit utterances
MFCC (Utterance Baseline 1)	98.66	89	13
Spectral envelope features (Utterance Baseline 2)	98.58	96	12

Table 6-2 Results of baseline systems for utterance-level SID using MFCC features (Utterance Baseline 1) and spectral envelope features (Utterance Baseline 2)

6.4. Text-Dependent Weights

The baseline results on text-dependent SID (Digit Baseline 1 in Table 3-5 and Digit Baseline 2 in Table 4-1) show that some of the words are more reliable in discriminating speakers than the others. For example, digit '3' and '4' give relatively high SID rates, compared with that using digit '5' and '6'. It indicates that the ten digits are not equally effective in discriminating between speakers. Based on the text-dependent SID rate, we can assign text-dependent weights used in linear combination of likelihood scores from individual digit to give overall score for the utterance.

6.4.1. Implementation Details

The steps are the same as that in utterance-level baseline system. However, text-dependent weights λ are used. The weights are chosen to reflect the effectiveness of using a particular digit in discriminating between speakers. They are determined based on the text-dependent SID result. The weights will be higher for those digits that have higher SID rate so that they can contribute more in determining speaker's

identity.

Sessions S07 – S08 are used to perform text-dependent SID and the results are listed in Table 6-3. The SID rates for the 10 digits have been sorted in ascending order. The text-dependent weights λ will be determined based on this.

	Text-dependent SID rate (%)			
Digit	Using MFCC	SID rate sorted in ascending order	Using spectral envelope features	SID rate sorted in ascending order
0	98.58	8	98.25	6
1	97.83	7	96.75	4
2	96.66	4	98.08	5
3	99.5	10	99.5	10
4	97.5	6	99	8
5	93.73	1	94.07	2
6	95.32	2	92.73	1
7	97.32	5	98.49	7
8	95.57	3	96.16	3
9	99.33	9	99.08	9

Table 6-3 Text-dependent SID rate (%) using MFCC features and spectral envelope features for the use of adjusting text-dependent weights λ

Different sets of text-dependent weights λ have been tried in the experiment. The higher ranking the digit attains, the heavier weight it is assigned. One scheme of designing weights is to assign values ranged from 0.1 to 1 to each digit according to the ranking of the SID rate. Similar method is also tried but with values in smaller range. Another scheme is to divide the digits into two groups based on their SID rates. The first five digits that attained the highest SID rate are grouped together while the remaining digits form another group. Digits in the same group will be assigned with same weight. The weight equals to 1 for the former group while it is equal to 0.5 for

the latter group.

6.4.2. Results and Analysis

The best performance given by using text-dependent weights are listed in Table 6-4 and Table 6-5. The corresponding values of text-dependent weights are also given.

Using MFCC features

Digit	λ	Digit	λ
0	0.8	5	0.1
1	0.7	6	0.2
2	0.4	7	0.5
3	1	8	0.3
4	0.6	9	0.9

(a)

	Utterance level SID rate (%)	Number of errors for multi-digit utterances
Using text-dependent weights	98.72	8
Utterance Baseline 1	98.66	13

(b)

Table 6-4 Using text-dependent weights in utterance-level SID with MFCC features (a) Text-dependent weights for the 10 digits; (b) Result of utterance-level SID

Using spectral envelope features

Digit	λ	Digit	λ
0	0.9	5	0.8
1	0.85	6	0.8
2	0.9	7	0.95
3	1	8	0.85
4	0.95	9	1

(a)

	Utterance level SID rate (%)	Number of errors for multi-digit utterances
Using text-dependent weights	98.59	11
Utterance Baseline 2	98.58	12

(b)

Table 6-5 Using text-dependent weights in utterance-level SID with spectral envelope features (a) Text-dependent weights for the 10 digits; (b) Result of utterance-level SID

Applying text-dependent weights λ in utterance-level SID can deal with identification errors caused by utterances containing digit string. Those multi-digit utterances that caused identification errors in the utterance-level baseline are compared with that by using text-dependent weights (Table 6-6). We find that using text-dependent weights can correct some identification errors, but at the same time some other recognition errors are created.

	Using MFCC features	Using spectral envelope features
Number of identification errors solved	6	1
Number of identification errors created	1	0

Table 6-6 Identification errors that belong to multi-digit utterances and solved/created by using text-dependent weights in utterance-level SID are counted, in comparison with the identification errors in the utterance-level baseline system

This approach of applying text-dependent weights in utterance-level SID is probably effective if the likelihood scores from the true speaker model are the maximum for those digits with relatively high text-dependent SID rate. However, it

cannot help to solve recognition errors in some cases.

Digit with low text-dependent SID rate implies that it is not reliable in discriminating speakers. If all digits spoken in the utterance have relatively low text-dependent SID rate, applying text-dependent weights may not help. One example is '56'. From the result, we find that over 50% of identification errors are caused by utterances containing this digit combination.

In some cases, applying text-dependent weights in utterance-level SID may create more identification errors. The current scheme of applying text-dependent weights puts less emphasis on the contribution from digits with low text-dependent SID rate. However, if the likelihood scores from the true speaker model are the maximum for digits that have relatively low text-dependent SID rate, recognition error may then be created by this approach (see Example 1 below).

Example 1: An identification error is created by applying text-dependent weights

Digit string '56' is spoken by speaker M18 and MFCC features are extracted. It is wrongly identified as speaker M01 after applying text-dependent weights.

The five highest scores with the corresponding speakers are listed.

	Score for Digit 5	Speaker	Score for Digit 6	Speaker	Combine in equal weights	Speaker	Combine in text-dependent weights	Speaker
	-70.772	M12	-67.423	M15	-137.977	M07	-20.6622	M19
	-67.534	M19	-65.917	M07	-137.078	M19	-20.5727	M15
	-67.308	M01	-64.86	M18	-137.047	M07	-20.2964	M08
	-66.503	M06	-64.815	M02	-131.015	M01	-19.5014	M18
MAX	-65.294	M18	-63.707	M01	-130.154	M18	-19.4722	M01

After analyzing the identification errors, we find that in some cases, the likelihood scores from the true speaker model are the maximum for digits that are considered to be relatively unreliable in discriminating speakers. This may suggest

that the effectiveness of different digits in identifying speakers is speaker-dependent. The idea of designing speaker-dependent text-dependent weights can be explored in future.

From Table 4-1 in Chapter 4, we find that text-dependent SID using NB features also shows text-dependent performance. Such text-dependent performance is not coherent between NB and WB systems. Some digits exhibit higher reliability in discriminating speakers under WB system than that in NB system (e.g. digit '2'). It is due to different importance of subband features for different digits. If text-dependent weights in NB system are also found, we can use different sets of weights based on what kind of speech data (NB or WB) we have to do utterance-level SID.

6.5. Text-Dependent Feature Weights

It was shown that the importance of NB and HB features are text-dependent in speaker recognition. In Chapter 5, fusing features from NB and HB with feature weights for text-dependent SID has been studied. Recognition performance can be improved by this approach. Therefore, integrating text-dependent weights for subband features in utterance-level SID is investigated.

6.5.1. Implementation Details

The steps of this experiment are the same as the utterance-level baseline system. Spectral envelope features extracted from WB are used. Text-dependent weights λ are equal to 1 for all digits.

Calculation of likelihood score from individual digit is different in this experiment. It is computed by equation (5.4) with text-dependent weights for subband

features.

The feature weights (α and β) for NB and HB are found following the steps described in Chapter 5. Sessions S07 – S08 are used for this purpose. The chosen values of α and β are listed in Table 6-7.

Digit	α	β	Text-dependent SID rate (%)	
			Using text-dependent weights for subband features	Using equal weights for subband features
0	0.78	1.55	98.75	98.25
1	1	1	96.75	96.75
2	1	1	98.08	98.08
3	1.02	0.95	99.58	99.5
4	1.11	0.725	99.17	99
5	1.1	0.75	94.49	94.07
6	0.9	1.25	93.07	92.73
7	1	1	98.49	98.49
8	0.9	1.25	96.24	96.16
9	0.85	1.375	99.25	99.08

Table 6-7 Chosen values of text-dependent feature weights (α and β) used in utterance-level SID

6.5.2. Results and Analysis

In Chapter 5, it was shown that applying subband feature weights in text-dependent SID can improve the performance. It is expected that this approach can help reducing the number of identification errors from both single-digit utterances and multi-digit utterances in utterance-level SID.

Table 6-8 lists the result of using text-dependent feature weights in utterance-level SID. The SID rate given by this method is lower than that in Utterance Baseline 2. We find that the number of identifications errors caused by single-digit

utterances is increased while the number of errors caused by multi-digit utterances remains unchanged.

	Utterance level SID rate (%)	Number of identification errors	
		Single-digit utterances	Multi-digit utterances
Using text-dependent feature weights	98.56	97	12
Utterance Baseline 2	98.58	96	12

Table 6-8 Result of using text-dependent feature weights in utterance-level SID with spectral envelope features

Applying text-dependent feature weights in utterance-level SID cannot give improvement in performance. The possible reason is that data used for adjusting the feature weights (α and β) and evaluating the system are come from different sessions, i.e. there is data mismatch. In Chapter 5, the same set of speech data is used to adjust the feature weights and test the system. Improvement can be achieved by applying subband feature weights in this way. However, in this experiment, the data mismatch caused that the found values of α and β do not fully suit with the testing data. Hence, recognition errors may be created.

6.6. Text-Dependent Weights Applied in Score Combination and Subband Features

Text-dependent weights for utterance-level score combination and text-dependent subband feature weights have been studied separately in the previous sections. Using both approaches together to perform utterance-level SID is then investigated.

6.6.1. Implementation Details

Spectral envelope features extracted from WB are used. The experimental steps are the same as the utterance-level baseline system, except the following two processes.

Likelihood score for each individual digit is computed by equation (5.4) using the subband feature weights listed in Table 6-7. The set of text-dependent weights λ that gave the highest utterance-level SID rate in Section 6.4 is used.

6.6.2. Results and Analysis

Table 6-9 lists the result of using text-dependent weights in score combination and subband feature weights to perform utterance-level SID.

Using text-dependent weights in score combination and subband feature weights

Utterance level SID rate (%)	98.55
Number of identification errors from single-digit utterances	97
Number of identification errors from multi-digit utterances	13

Table 6-9 Result of using text-dependent weights in score combination and text-dependent feature weights to perform utterance-level SID with spectral envelope features

Using both approaches of subband feature weighting and text-dependent weighting in score combination cannot give further improvement than by using either one of them. As we have mentioned before, each approach has its own limitation. The experimental result indicates that both approaches are not complementary to each other. Therefore, no further improvement can be achieved by applying both approaches at the same time.

6.7. Discussion

Table 6-10 and Table 6-11 summarize the results of utterance-level SID by applying different approaches of text-dependent weights with MFCC and spectral envelope features respectively. Among the three approaches we studied in this chapter, the approach of using text-dependent weights in score combination gives the best performance.

Using MFCC features

	Utterance level SID rate (%)
Utterance Baseline 1	98.66
Using text-dependent weights in score combination	98.72

Table 6-10 Summarize the results of utterance-level SID with MFCC features

Using spectral envelope features extracted from WB

	Utterance level SID rate (%)
Utterance Baseline 2	98.58
Using text-dependent weights in score combination	98.59
Using text-dependent feature weights	98.56
Using text-dependent weights in score combination and text-dependent feature weights	98.55

Table 6-11 Summarize the results of utterance-level SID with spectral envelope features

The results show that applying text-dependent weights in score combination can help to reduce identification errors, where the performance gain for using MFCC and spectral envelope features are equal to 0.06% and 0.01% respectively. The approach of using text-dependent subband feature weights in utterance-level SID cannot

perform as good as it did in text-dependent SID. It may be due to the data mismatch between training feature weights and evaluating the system. We suggest that the above two approaches cannot complement each other. Therefore, using both approaches together cannot give further improvement.

In this chapter, applying text-dependent weights in score combination has been studied using multi-digit utterances. Actually, we can extend the use of this approach to other text-dependent applications (i.e. not only multi-digit utterances).

In our study, text-dependent weights are adjusted intuitively. In future, more study can be done on how to adjust the weights so as to better reflect the reliability for different words in discriminating speakers. We believe it can further improve the recognition performance.

Chapter 7

Conclusions and Suggested Future Work

7.1. Conclusions

Voice is one of the most natural and the least obtrusive biometric measures for the identification of a person. It is an attractive field of research nowadays. In real applications, many factors can degrade the recognition performance. This includes speaker-generated variability and variability induced by recording channels and conditions. In this thesis, we focus on the speaker-dependent information in different frequency subbands.

When speaker recognition is performed in machine, the bandwidth of speech signal is limited. For example, the telephone bandwidth is 0 – 4 kHz. Speaker recognition using telephone speech essentially assumes that useful speaker-dependent information can be found mostly at the frequency below 4 kHz. However, many studies showed that frequency band above 4 KHz does contain important features of a speaker's voice. In this thesis, we have studied the contributions of features from NB (0 – 4kHz) and HB (4 – 8kHz) in speaker recognition. Instead of lumping NB and HB features directly for speaker recognition, suggestions can be made on how to fuse features from these two bands based on their importance. We believe that such understanding can provide a possible means to improve recognition performance.

In this thesis, we focus on studying spectral envelope features extracted from NB and HB for text-dependent speaker identification (SID). The experimental results

confirm that HB contains useful speaker-dependent features. More precisely, it has been observed that these features come from some fricatives or other phonemes that have more energy concentrated in HB. Therefore, the contributions of NB and HB features in discriminating speakers are different for different speech sounds.

Two approaches of fusing features from NB and HB have been studied. They are model level fusion and feature level fusion. To maximize the benefit of model level fusion, data streams should be statistically independent. But it is found that there is correlation between features extracted from NB and HB and this correlation is also useful in speaker recognition. Therefore, it is not suitable to perform model level fusion. Feature fusion with text-dependent subband feature weights has been investigated. It can improve the identification accuracy by 0.09% on average. The improvement is small and it is suggested that adjusting feature weights α and β cannot affect the contribution of NB and HB features in speaker recognition. We believe that fusing features from NB and HB based on their importance in SID provides a possible means to improve recognition performance. Therefore, it is required to investigate other methods of fusing features from these two bands in future.

We extend our study on subband spectral envelope features to utterance-level SID. The result of text-dependent SID shows that some words are more reliable in discriminating speakers. For example, digit '3' and '4' give higher SID rates than the others. Based on this result, we can assign text-dependent weights used in linear combination of likelihood scores from individual digit so that those words with higher reliability contribute more in SID. The experimental results show that this method can achieve 0.06 % and 0.01 % improvement with MFCC and spectral envelope features respectively.

By summarizing the above results, some suggestions for speaker recognition are given as follows:

1. Fusing features from NB and HB with subband feature weights based on their importance is preferred. It is also suitable to apply this approach in other text-dependent applications (i.e. not only digit-dependent).
2. By studying text-dependent SID result, those words that are more reliable to discriminate speakers can be found. Based on their reliability in discriminating speakers, text-dependent weights used in score combination for utterance-level SID can be determined.

Although NB features still play the main role in determining speaker's identity, features from HB also give its contribution. With a better understanding on the contribution of features from these two bands, we can fuse them together in an appropriate way so as to maximize the benefits from these two bands. Similarly, if the word reliability in discriminating speakers can be known, we can design weighting to put more emphasis on those words in performing speaker recognition.

7.2. Suggested Future Work

1. Study fine details features in different frequency bands

In this thesis, we focus on the vocal tract characteristics. As we have discussed before, vocal source characteristics also contain speaker-dependent information. The fine details in the speech spectrum carry features of vocal source. By examining the speech spectrum carefully, we observe that the fine details show less harmonic in HB. So we suggest studying fine details features in different frequency bands in future. It is possible to apply the contribution of fine details features and spectral envelope features in NB and HB for speaker recognition. We believe it can further improve the performance of a speaker recognition system.

2. Study contribution of features in different frequency bands on phoneme basis

From the results of text-dependent SID experiment, it shows that the contribution of NB and HB features is text-dependent due to different phonetic composition. Therefore, it is suggested to study the contribution of features from these two bands on phoneme basis. It can enable us to have a clearer picture on how to use NB and HB features thoroughly.

Appendix

Appendix 1 Speech Content for Data Collection

Use in Session 1 and 2	Use in Session 3 and 4	Use in Session 5 and 6
0	0	0
5	5	5
7	7	7
9	9	9
2	2	2
3	3	3
6	6	6
8	8	8
1	1	1
4	4	4
17	39	04
68	80	32
93	21	91
79	57	67
26	46	85
3230 4104	8691 0473	0315 6842
0784 0331	7051 4260	3579 1096
2136 0587	1527 6938	8367 9524
9734 2954	2493 1857	4731 8270
7739 2608	3048 6952	1902 5846

Use in Session 7 and 8	Use in Session 9 and 10	Use in Session 11 and 12
0	0	0
5	5	5
7	7	7
9	9	9
2	2	2
3	3	3
6	6	6
8	8	8
1	1	1
4	4	4
01	24	09
74	51	18
59	06	34
36	37	56
82	89	27
4209 3618	3450 9428	8623 2570
8903 2567	7935 4261	7891 6745
4513 9408	8910 5637	0192 5384
2671 3095	2908 7136	3906 1947
5782 7461	8024 1675	2834 0517