Exploitation of Phase and Vocal Excitation Modulation Features for Robust Speaker Recognition

WANG, Ning

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

The Chinese University of Hong Kong
June 2011

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

To my parents.

Abstract of thesis entitled:

Exploitation of Phase and Vocal Excitation Modulation Features for Robust Speaker Recognition

Submitted by **WANG, Ning**

for the degree of **Doctor of Philosophy**

in **Electronic Engineering**

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

June 2011.

Speaker recognition (SR) refers to the process of automatically determining or verifying the identity of a person based on his or her voice characteristics. In practical applications, a voice can be used as one of the modalities in a multimodal biometric system, or be the sole medium for identity authentication. The general area of speaker recognition encompasses two fundamental tasks: speaker identification and speaker verification.

Mel-frequency cepstral coefficients (MFCCs) are widely adopted in speech recognition as well as speaker recognition applications. They are extracted to primarily characterize the spectral envelope of a quasi-stationary speech segment. It was shown that cepstral features are closely related to the linguistic content of speech. Besides the magnitude-based cepstral features, there are resources in speech, e.g, the phase and excitation source, are believed to contain useful properties for speaker discrimination. Moreover, in real situations, there are large variations exist between the development and application scenarios for a speaker recognition system. These include channel mismatch, recording apparatus mismatch, environmental variation, or even change of emotional/healthy state of speakers. As a consequence, the magnitude-based features are insufficient to provide satisfactory and robust speaker recognition accuracy. Therefore, the exploitation of complementary features with MFCCs may provide one solution to alleviate the deficiency, from a feature-based perspective.

Acknowledgements

First of all, I would like to thank my supervisor, Professor CHING Pak-Chung, for his advices and guidance in many ways through these years. He always gave me insightful suggestions in pursuing research and imparted to me his knowledge and philosophy in research. I am of sincere gratitude to my supervisor for his consideration, patience and encouragements in the course of this study. I would also like to express my thanks to Professor LEE Tan, for his enlightening instructions on this research, and his many help in improving my technical presentation and writing skills. My appreciation goes to Professor MENG Mei-Ling and Professor CHIEN Jen-Tzung in my thesis committee, for their kind help and advice in my thesis. I am deeply grateful to Professor WANG Shi-Yuan for his illuminating comments on my work and encouragements to me. My thanks also go to Professor MA Wing-Kin, for his brilliant course on optimization and his suggestions to me during the lab seminars from which I benefit greatly.

To my fellows in the DSPST lab, I feel lucky to work with all of you in these years. Particular and honest thanks are to my seniors, Dr. ZHENG Nengheng, who generously provided me very much help and valuable materials in the start of my research on speaker recognition, Dr. LEE Siu Wa, for her precious time spent on discussing with me. Technical support from Mr. LUK Kin On is also appreciated.

To my boyfriend Dr. YANG Chenguang, for his support, encouragements and sharing in belief and life. To my dearest parents, for their unconditional love, their always being there for me, through the good times and the bad. Finally, I am very grateful to the Chinese University of Hong Kong for providing me with this opportunity to undertake the PhD study.

摘 要

說話人識別 (Speaker Recognition) 是通過說話人的聲音特徵來自動識別其身份的。它可單獨用於身份驗證，或作爲多模態生物識別系統中的一個模態。說話人識別一般包含兩种工作模式：說話人辨認 (Speaker Identification) 及說話人確認/認證 (Speaker Verification)。

MFCCs 是一種廣泛應用於語音識別及說話人識別的特徵參數。該參數表徵了具有准平穩特性的語音信號的頻譜包絡，研究發現它與信號所傳遞的語言信息有密切關係。除了這類與幅度相關的倒譜特徵向量以外，語音信號中還包含了其他關乎說話人身份的重要信息，例如，相位和聲源信息。而且，一個說話人識別系統往往需要應用在與其訓練條件存在很大差別的環境下，這些差別通常來自信號的傳輸媒介，錄音設備，背景環境，甚或說話人本身的情緒及健康狀況。因此，單單倚靠與幅度相關的語音特徵難以提供令人滿意及穩健的說話人識別準確度。這使得借助於提取能夠補足 MFCCs 的特徵參數來提高識別性能成爲改善這一現狀的重要途徑。

AM-FM 信號建模是一種近年來應用於描述及分析語音信號的技術。根據其模型，一個多成分信號首先被分解，然後通過信號解調，每個單一成分的即時包絡和頻率分量實現分離。我們由此生成與相位或聲源相關的調製特徵參數來彌補 MFCCs 在這些方面的不足。借助這種多頻帶解調與分析的思路，我們提出了一個從語音及聲源信號中提取說話人特徵參數的新方法：將構成一個語音信號的主要頻率分量抽取出來並從中取得可用於辨別說話人的特徵參數，稱之爲 averaged instantaneous frequency of speech (SAIF)；另外，將作爲聲源信號中主要成分的有關幅度分量和頻率分量抽取並生成特徵向量 averaged instantaneous envelope/frequency of residue (RAIE/RAIF)。這些參數描述聲音特徵的能力經由特定實驗以及說話人識別系統驗證，他們與 MFCCs 之間的互補性也通過系統混合時的得分得到肯定。

對於實際應用中由於缺乏準確的特徵參數而造成的說話人認證準確率降低，我們提出了一種 feature mapping 的方法來提高特徵參數的穩健性，進而解決這個問題。通過對受到背景噪聲及傳輸媒介影響的特徵向量進行深入研究，我們發現這些干擾因素導致向量中參數的短時分佈發生畸變。試驗表明我們的方法可以改善這種趨勢。經由在麥克風及電話語音數據庫上開展的不匹配環境下的說話人認證實驗證明，本文所提出的穩健特徵提取方法有效地提高了系統辨識成功率。

AM-FM signal modeling is a technique that has been used in characterizing and analyzing speech properties recently. In the framework of AM-FM modeling, the multi-component signal is first decomposed, each single-component signal is then described by the instantaneous envelope and instantaneous frequency quantities. It is motivated in our work to capture the relevant phase and vocal excitation related modulation features in complementing with MFCCs. In the context of multi-band demodulation analysis, we present a novel parameterization of speech and vocal excitation signal. A pertinent representation for most dominant primary frequencies present in the speech signal is first built. It is then applied to frames of the speech signal to derive effective speaker-discriminative features, namely the averaged instantaneous frequency of speech (SAIF). The source-related amplitude and phase quantities are also parameterized into feature vectors, which are referred to as the averaged instantaneous envelope/frequency of residue (RAIE/RAIF). The speaker-distinguishable information conveyed by the proposed features is evaluated through a set of specifically designed experiments. The application of the features is also assessed in the context of a standard speaker identification and verification system. Complementary correlation between MFCCs and the modulation features is revealed by system fusion on score level.

In order to alleviate the problem of severe degradation of speaker authentication performance under mismatched conditions due to inadequate and inaccurate speaker-discriminative information, a method of feature mapping that can build a more robust representation for each stream of parameters in the feature vector is proposed. Through extensive observations on features when experiencing additive noise or unexpected communication environments, it is found that these adverse effects can distort the short-term distributions of the speaker parameters. It is also noted that by mapping each feature stream to a target distribution over a specific time interval, their robustness to environmental or channel mismatch can be enhanced. Through speaker verification experiments on microphone and telephone data, it has been proven that the proposed robust feature extraction front-end can consistently reduce the equal error rate.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Being capable of distinguishing among different people from their voices in whatever environments is a highly desired ability for machines operated for biometric authentication. This kind of application that uses a machine to recognize person from the spoken phrases is called automatic speaker recognition. It is an important biometric issue, as well as a fundamental research problem in speech processing. Speaker recognition systems work in two modes: identification and verification/authentication.

Studies indicate that good performance can be achieved if a well-trained speaker recognition system operates in an environment similar with that under which it is developed. However, various kinds of difference between the training and test scenarios, which as a result cause the so-called *mismatched condition*, exist ubiquitously in actual applications. The poorly performed systems to a concern degree suffer from this mismatch. A typical example for it is when a system is trained from clean data and tested by noise- or channel-corrupted speech. Robust speech processing techniques that emerge therefore attempt to minimize the latent mismatch and maintain the performance of speech processing systems such that they can operate under various types of environment.

This chapter reviews the basic knowledge in speaker recognition research first, and then introduces the difficulties and challenges a speaker authentication system faces in practical applications. Motivation, objectives and the organization of this thesis is presented in the end.

1.1 Speaker Recognition

Whenever people speak an utterance, they deliver not only a message that carries meaning, but also information about themselves as an individual. The same utterance spoken by any two persons will sound different, this is because the process of speaking involves extensively the neural, physiological, anatomical and physical systems of a specific individual in a concern circumstance. The speaker-specific characteristics in the speech signal can provide information relevant to the speaker's anatomy, physiology, linguistic background and emotional condition, etc. This information can be captured and processed by listeners and human-computer interfaces to describe and characterize speakers.

An important application of speaker recognition technology is forensic, to identify the persons involved in the voice record for criminal purposes, since it is a basic and essential way to exchange information through telephone conversations for the two parties. Ordinary persons will benefit from speaker recognition technology as well. Nowadays, biometric authentication systems are required in more and more areas, e.g., in remote access to database, telephony banking, *e*-commerce, etc. Usually, biometric systems recognize a person by using distinguishing traits. As a performance biometric, human voice cannot be forgotten or misplaced, like other physiological characteristics. Experts have predicted that in the future, telephone-based services with integrated speech recognition, speaker recognition, and language recognition will pose a potential supplement or even replacement to the human-operated service. In fact, the focus of speaker recognition research over the years has been tending towards such telephony-based applications. In addition, speaker recognition techniques also involve in indexing broadcast programs or annotating recordings, which is therein termed as speaker diarization. It is an extension of the speaker recognition techniques in multiple speakers case. Besides, research on socially assistant robotics nowadays began to introduce human-robot communication in virtue of speech interface to the robotic systems [1].

Speaker recognition systems, at another point, can be divided into *text-dependent* and *text-independent* ones. In text-dependent systems, the test ut-

terances are fixed, or known beforehand. In text-independent systems, there is little constraint on the words allowed to use by a speaker. Thus, in training and test stages, the utterances could be completely different in content, thus the recognition system should take into account this phonetic mismatch. It is seen that the text-independent recognition is more challenging than the text-dependent task.

In practical applications of speaker recognition, variations in the acoustic environment and technical factors, such as, recording, transmission, as well as intra-speaker variation, such as health condition, mood and aging, bring undesirable effects to the speaker recognition performance. In general, any variation between two recordings of the same speaker is known as *session variability* [2].

All classes of speaker recognition task, identification or verification, text-dependent or text-independent, own specific advantages and disadvantages. In practice, a proper choice among them depends on the application. Two modules are generally included in all systems, they are *feature extraction* and *feature matching*. Feature extraction is the process of parameterizing speech waveforms into vectors of specific types of coefficients. The extracted vectors are of much smaller dimension compare with the original data samples. The resultant feature vectors are used to represent relevant speaker. Feature matching refers to the procedure of comparing the extracted features with the ones that stored beforehand.

Speaker recognition has been studied as early as 1960s, the first study was carried out to learn how human recognize speakers and the reliability of human's recognition performance [3], [4], [5]. With the development of computer technology, 1970s saw the invent and arising interests in automatic speaker recognition by computer systems. In this stage, Fourier transform, linear prediction and cepstral analysis techniques have been employed in generating speech features. Long-time average of these parameters were used as speaker representative models.

Around the 1980s, dynamic time warping (DTW) [6], vector quantization (VQ) [7] techniques have been proposed, which greatly push forward the de-

velopment of current speaker recognition system. Cepstral features, i.e., the mel-frequency cepstral coefficients (MFCCs) [8], linear predictive cepstral coefficients (LPCCs) [9], and also line spectrum pairs (LSPs) [10], [11] were proposed and have obtained the most widely use until today. Also at this time, the dynamic features were suggested for use [12].

In the 1990s, sophisticated statistical techniques, such as, Gaussian mixture model (GMM) [13], support vector machine (SVM) [14], [15], has been used to train speaker models. Background score normalization techniques, i.e., cohort background model [13] and universal background model (UBM) [16] were developed for robust verification of speaker identity.

Through the years, many efforts have been made to explore new and alternative speaker representatives, however, up to date, most of the state of the art systems still adopt MFCC to produce their benchmarks.

1.2 Difficulties and Challenges of Speaker Authentication

Provided that the state-of-the-art speaker authentication techniques perform well in laboratory simulations, they have to face many difficulties when applying in actual applications. Generally speaking, nowadays, the primary challenges ahead are arisen from the followings.

- **Sparse data source and unstable acquisition environments**

Speaker authentication that involved in biometrics, access control and forensic applications usually based upon various sorts of data source, among them, those using mobile devices is becoming a convenient and important way to security control for remote services such as telephony banking and *e*-commerce. Such applications constitute quite challenging pattern recognition problem, basically owing to sparse data samples and unstable acquisition environments. The sparsity of available data set will lead to poorly-trained model for a speaker. Meanwhile, inconsistence of data quality that exists among acquired speech samples affect the ultimate performance as well. Similar rules also apply to the testing stage, where in some places the testing duration is kept as short as possible.

Given plenty of developing data source, we can have reliable and satisfactory accuracy achieved. Take the NIST Speaker Recognition Evaluation (SRE) for instance, recently, more and more sites tend to employ corpus-based approaches such as joint factor analysis (JFA), etc, to completely model all sorts of session variability [2], without identifying case by case the influential factors involved. However, in practice, the corpus-based approaches might not be suitable solutions in realistic cases, where conditions may very well fall outside the applying scope of the approaches.

- **Unexpected operation scenarios**

Most of the times, a speaker authentication system is operated under unexpected environments other than where it was developed. This may include the

ambient noise, reverberation and echo induced by the room acoustics, linear and non-linear distortions introduced by the acquisition process, bandwidth filtering and distortion involved by transmission channel, and others. These mismatches between the development and operation scenarios lead to inconsistence and annoyance to the acoustic patterns of individual speakers. Under these rigorous circumstances, concern degree of robustness to the undesirable effects that surround is considered indispensable.

1.3 Scope of Research and Thesis Outline

In this thesis, we concentrate on exploiting novel and efficient speaker-distinctive parameters for robust speaker recognition purpose. Therefore, we will focus more on the feature extraction module of a speaker recognition system, but less on the subsequent matching process.

It is known that Mel-frequency cepstral coefficients (MFCCs) have been most commonly used in both speech recognition and speaker recognition systems. They are extracted to primarily characterize the spectral envelope of a quasi-stationary speech segment. It was shown that cepstral features are closely related to the linguistic content of speech. Besides the magnitude-based cepstral features, there are resources in speech, e.g, the phase and excitation source, are believed contain useful properties for speaker discrimination. Moreover, in real situations, there are large variations exist between the development and application scenarios for a speaker recognition system. As a consequence, the magnitude-based features are insufficient to provide satisfactory and robust speaker recognition accuracy. Therefore, the exploitation of complementary features with MFCCs may provide one solution to alleviate the deficiency, from a feature-based perspective.

AM-FM signal modeling is a technique that has been used in characterizing and analyzing speech properties recently. In the framework of AM-FM modeling, the multi-component signal is first decomposed, each single-component signal is then described by the instantaneous envelope and instantaneous frequency quantities. It is motivated in our work to capture the relevant phase and vocal excitation related modulation features in complementing with MFCCs. In the context of multi-band demodulation analysis, we present a novel parameterization of speech and vocal excitation signal. The speaker-distinguishable information conveyed by the proposed features is evaluated through a set of specifically designed experiments. The application of the features is also assessed in the context of a standard speaker identification and verification system. Complementary correlation between MFCCs and the modulation features is revealed by system fusion on score level.

As far as the challenges existing in practical applications are concerned, a postprocessing step of feature enhancement is thought essential in order to provide robust speaker-distinctive features. Via observing and identifying the adverse effects endured by the concerned speaker representatives in mismatched-noise and -channel conditions, we decide to condition the short-term distributions of the parameter streams to a more robust representation, in order to maintain their characteristics in various kinds of environments.

This thesis is organized as follows:

- We in Chapter 1 first introduce the scope and methodologies of speaker recognition research. The problem to solve, objectives to achieve, as well as difficulties to tackle in this research are stated thereafter. Our opinion on this subject, the perspectives adopted and solutions submitted are also presented.

- Chapter 2 delivers the technical issues involved in speaker recognition research. Besides a review on speaker modeling and matching techniques, we give a particular close look at the feature extraction front-end, in which the derivation, physical meaning, applying fields of the primary types of speech parameters developed have been introduced systematically.

- Chapter 3 focuses on an important issue encountered when putting a speaker recognition system to use in actual applications, that is, its robustness to the potential mismatch with where it was developed. We are especially concerned for the variations caused by ambient noises and transmission channels in this study.

- In Chapter 4, a study is made on characterizing a speaker's vocal excitation pattern through modeling the corresponding voicing signals by amplitude-frequency modulation parameters. A novel parametrization method is raised to capture the essential source-related features from concerned sequences of instantaneous parameters. The extracted features are

then evaluated by theoretically designed experiments as well as examined by speaker recognition simulations on real database.

- Chapter 5 is dedicated to derive phase-related distinctive features for speaker recognition purpose. After looking into the amplitude-frequency components that present in speech signals, a method is proposed to identify and quantify these primary components. A pertinent representation for these most dominant primary frequencies present in the speech signal is then built and applied to frames of speech signal to derive effective phase-related features. Extracted parameters are passed to systematic inspection as well as evaluation in recognition tasks.

- To provide a valid solution to the robustness problem we raised at the beginning of this thesis, in Chapter 6, a feature enhancement approach is attached as a post-processing step to the feature extraction front-end of a speaker authentication system. Through speaker verification experiments under mismatched noise and channel conditions, it is observed that the proposed robust feature extraction front-end consistently reduces the equal error rate.

- Chapter 7 concludes the findings and output of this research first, and then give suggestions as future directions in proceeding this work.

Chapter 2

Fundamentals in Speaker Recognition

Automatic recognition of speaker identity has been a goal of research for many years. One of its objectives is to decide which particular voice model from a known set of voice models best characterizes a speaker; this task is referred to as Speaker Identification (SID). In the different task of Speaker Verification (SV), the goal is to decide whether a speaker corresponds to a particular known voice or to some other unknown voice. Speaker recognition systems have already been employed in applications where a sole medium for identity authentication, e.g., telephony banking, is needed; or as one modality in a multi-modal biometric system.

This chapter introduces the fundamentals of speaker recognition, with an emphasis on text-independent recognition. In Section 2.1, an overview of the state-of-the-art speaker recognition system is given. The basic physiology of speech production, human auditory system as well as digital speech model will be given in Section 2.2. Section 2.3 focuses on attributes of human voice which avail ourselves of distinguishing among different speakers, the common pre-processing of speech signals are indicated thereafter. Section 2.4 categorizes the primary parameter sets employed in speaker recognition systems into five classes and makes a review on them. Speaker modeling, as an essential component in this classification problem is introduced in Section 2.5. Different

recognition tasks that are being dealt with in this research are described in Section 2.6 together with their corresponding performance evaluation. Section 2.7 summarizes the salient points covered in this chapter.

2.1 Speaker Recognition System Overview

Speaker recognition involves identifying the person talking rather than what is being said, the speech signal must be processed to extract measures of speaker variability instead of being analyzed by segments corresponding to phonemes or pieces of text one after the other. For speaker recognition, only the identity of individual speaker will be classified based on an input of test utterance.

Both automatic speaker verification and speaker identification use a stored database of reference patterns (templates) for N known speakers. Both involve similar analysis and decision techniques. Verification is simpler because it only requires comparing the test pattern against one reference pattern and it involves a binary decision: Is there a good enough match against the template of the claimed speaker? The error rate for speaker identification can be much greater because it requires choosing which of the N voices known to the system best matches the test voice or "no match" if the test voice differs sufficiently from all the reference templates. Figure 2.1 depicts a systematic block diagram of a speaker recognition system using pattern recognition approach. The three basic steps in a pattern recognition model are (1) parameter measurement (in which a test pattern is created), (2) pattern comparison, and (3) decision making.

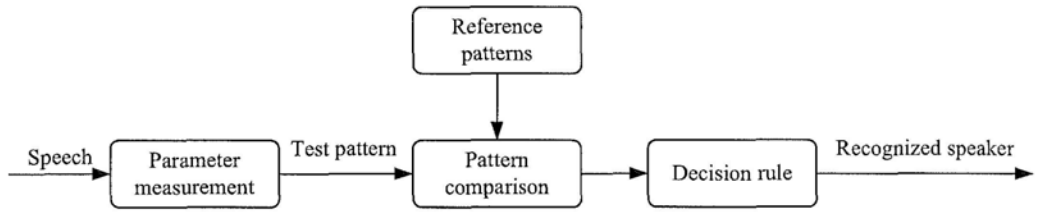
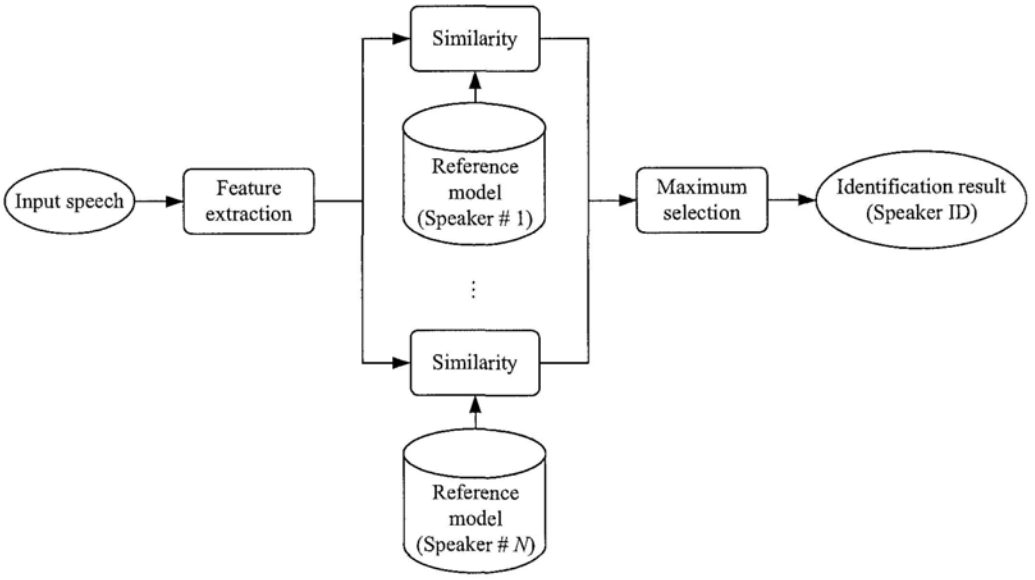
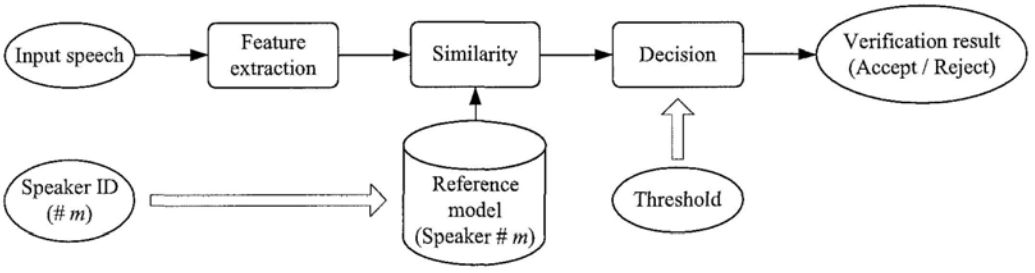


Figure 2.1: *Pattern recognition approach to speaker recognition*

The first step in a speaker recognition system, whether for identification or verification, is to build a *model* of the voice of each target speaker, as well as a model of a collection of background speakers, using speaker-dependent *features* extracted from the speech waveform. For example, the oral and nasal tract length and cross-section during different sounds, the vocal fold mass and shape,



(a). Speaker identification system.



(b). Speaker verification system.

Figure 2.2: *An overview of a speaker recognition system.*

and the location and size of the false vocal folds, if accurately measured from the speech waveform, could be used as features in an anatomical speaker model. We call this the *training* stage of the recognition process, and the associated speech samples used in building a speaker model is called the *training data.* During the recognition or *testing* stage, we attempt to match (in some sense) the features measured from the waveform of a test utterance, i.e., the *test data* of a speaker, against speaker models obtained during training. The particular speaker models we match against, i.e., from target and background, depends on the recognition task. An overview of these components of a speaker recognition system for the verification and identification tasks are given in Figure 2.2 [17].

2.2 Physiology of Speech and Hearing

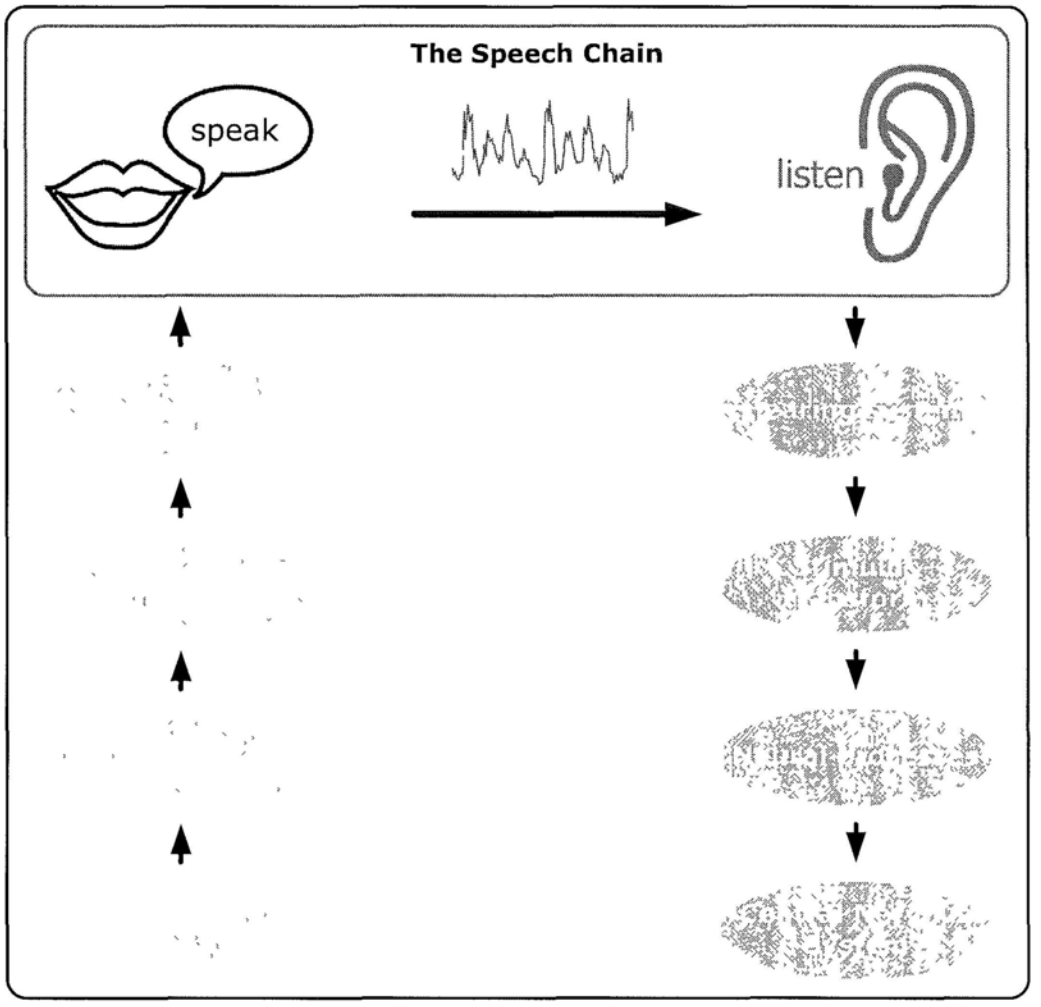


Figure 2.3: *The speech chain.*

Figure 2.3 illustrates the speech chain - a spoken message travels from the speaker to the listener [18]. The speech chain consists of three events: the production of speech sounds through the vocal apparatus of the speaker, the travelling of the acoustic signal through the air and, finally, its reception by the ear of the listener. The brain of the speaker controls the production of speech sounds and the brain of the listener analyzes the signal and converts it into meaning.

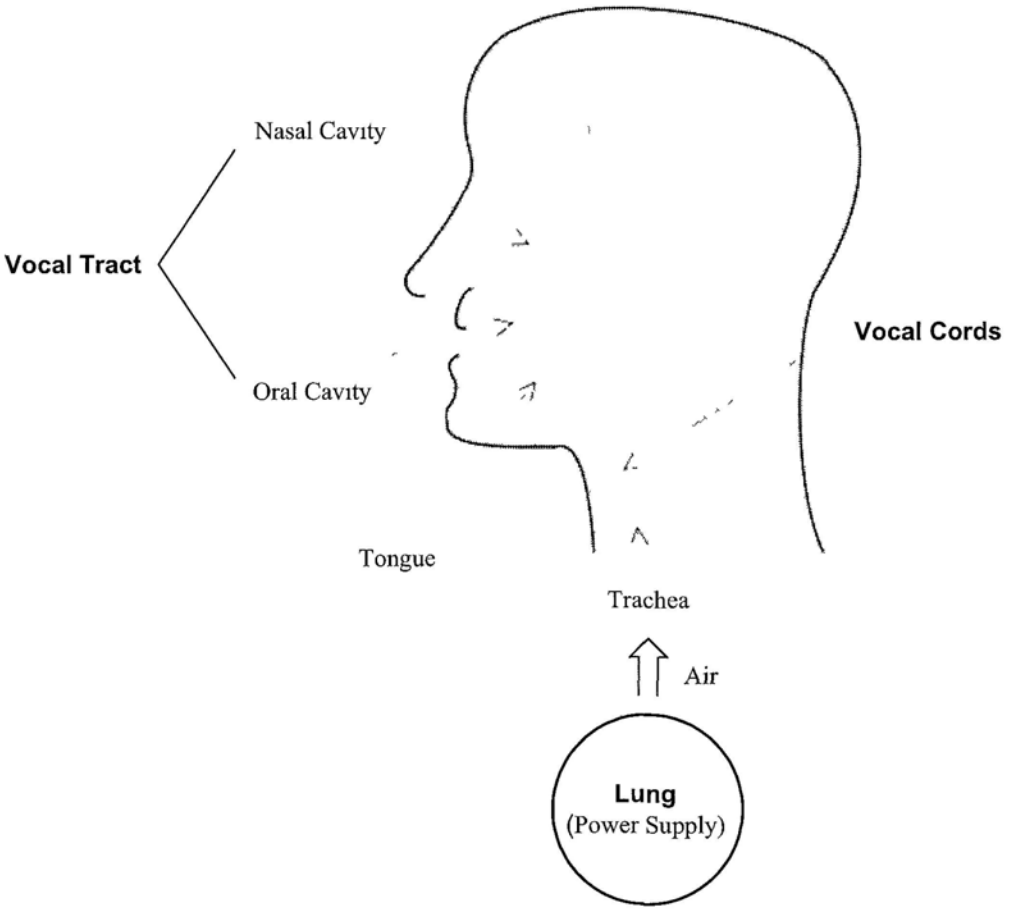


Figure 2.4: *The human vocal system.*

2.2.1 Speech production mechanism

One of the important links in the speech chain is speech production, the specialized movements of our vocal organs that generate speech sound waves. A simplified view of speech production is given in Figure 2.4, where the speech organs are divided into three main groups: the lungs, vocal cords, and vocal tract. Airflow is necessary for sound to be generated. The lungs act as a power supply and provide airflow to the vocal cords stage of the speech production mechanism. The vocal cords consist of ligament and muscle, and are adjustable under muscle control. The cartilage surrounding the vocal cords provides support. The opening that allows air to pass through the vocal cords from the trachea to the larynx is called the glottis. Depending on the absence and pres-

ence of vocal cord vibration, the airflow from the lungs is modulated into either a periodic puff-like or a noisy airflow source to the third organ group, the vocal tract. The vocal tract gives the modulated airflow its "color" by spectrally shaping the source. Sound sources can also be generated by constrictions and boundaries, not shown in the figure, that are made within the vocal tract itself, yielding in addition to noisy and periodic sources, an impulsive airflow source. Following the spectral coloring of the source by the vocal tract, the variation of air pressure at the lips results in a travelling sound wave that the listener perceives as speech.

Principal components in speech production

Speech differs from breathing in that at some point in the path you set the air in rapid motion or vibration. There are two principal components in speech production:

- Excitation - "creates" a sound by setting the air in rapid motion.
- Vocal tract - "shapes" the sound.

The excitation of speech has three forms:

- Phonation: vibration of vocal cords. The result of the vibrating excitation is a quasi-periodic release of air, the fundamental frequency of the vocal cord opening/closing cycle becomes the *fundamental frequency* (F0), or equivalently, *pitch* of the resulting sound.
- Frication: turbulent air flow. The excitation is set up by forcing air past a constriction at some point in the vocal tract.
- Plosive: closure at some point in the vocal tract, followed by a release of air.

There are then three general categories of the source for speech sounds: *periodic*, *noisy*, and *impulsive*, although combinations of three sources are often

present [17]. Examples of speech sounds generated with each of these source categories are seen in the word "shop", where the "sh", "o", and "p" are generated from a noisy, periodic, and impulsive source, respectively. Such distinguishable speech sounds are determined not only by the source, but by different vocal tract configurations, and how these shapes combine with periodic, noisy, and impulsive sources. Also defined in this production process are the speech properties that characterizes a speaker.

A widely used model for speech production is based upon the assumption that the vocal tract can be represented as a concatenation of lossless acoustic tubes [19]. The tube is closed at the glottis end and open at the mouth end. The modes of vibration of the vocal tract is resonances. In speech, the resonant frequencies are called *formant* frequencies.

Relation between excitation and the vocal tract

It is shown that the vocal tract modeled as a uniform tube can be represented with an *all-pole* transfer function. In terms of the LTI system model, the excitation is the input function $x(t)$, the vocal tract acts as the system function $H(e^{j\omega})$, and the speech is the output $s(t)$. The shape of the resulting spectrum is given by $H(e^{j\omega})$. In a simplified model for phonated sounds, the glottal pulses of $x(t)$ form an impulse train, with interval T between pulses. This appears in the speech spectrum as pulses at frequency intervals $1/T$, shaped by the $H(e^{j\omega})$ envelope. This system is not time-invariant, but for a short time interval of 10-30 *msec*, it can be viewed as a "piecewise" LTI system, where $h(t)$ is the impulse response. Thus, for a specific segment of speech, its excitation and vocal tract is related in the following manner:

$$s(t) = x(t) \otimes h(t). \tag{2.1}$$

Speech analysis systems typically assume all-pole filters are identical to tube models for all speech sounds.

2.2.2 Auditory system: hearing and perception

The other important link in the speech chain is speech perception, which processes the received speech sound. The sound waves are first collected by our outer ear, and then amplified in some frequencies. In this way, the vibrations of air are translated to vibrations of the tympanic membrane. This kind of vibration will be translated to oscillations of liquid in the inner ear by the middle ear. In the inner ear, the cochlea transforms mechanical vibrations into nerve impulses, and then to the brain. The processing in the inner ear obeying the "Place Theory": basilar membrane vibrates in response to sound; point of maximum vibration (i.e., displacement) depends on the frequency of the sound. Movement of basilar membrane translates mechanical signal to electrical signal.

2.2.3 Digital speech model

Section 2.2.1 introduces the physiology of speech production. It is possible to relate these physiological features with the speech signal model by deriving rather detailed mathematical representations for the acoustics that involved in the speech production process. It is seen that sound is generated in three ways, and that each mode results in a distinctive type of output. The vocal tract imposes its resonances upon the excitation so as to produce the different sounds of speech. This is the essence of modeling speech waveform in a digital manner.

Vocal tract

The resonances (formants) of speech correspond to the poles of the transfer function $V(z)$. An all-pole model is a very good representation of vocal tract effects for a majority of speech sounds; however, the acoustic theory tells us that nasals and fricatives require both resonances and anti-resonances (poles and zeros). In these cases, we may include zeros in the transfer function or we may reason with Atal [20] that effect of a zero of the transfer function can be achieved by including more poles. In most cases, this approach is to be

preferred.

$$V(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}, \tag{2.2}$$

where $\{\alpha_k, k = 1, 2, \ldots, p\}$, p and G are the order and gain of the filter, respectively. These are time-varying parameters that determine the speech model.

Since the coefficients of the denominator of $V(z)$ in Equation 2.2 are real, the roots of the denominator polynomial will be either real or occur in complex conjugate pairs. A typical complex resonant frequency of the vocal tract is

$$s_k, s_k^* = -\sigma_k \pm j2\pi F_k. \tag{2.3}$$

The corresponding complex conjugate poles in the discrete-time representation would be

$$\begin{aligned} z_k, z_k^* &= e^{-\alpha_k T} e^{\pm j2\pi F_k T} \\ &= e^{-\alpha_k T} cos(2\pi F_k T) \pm j e^{-\alpha_k T} sin(2\pi F_k T). \end{aligned} \tag{2.4}$$

The bandwidth of the vocal tract resonance is approximately $2\sigma_k$ and the center frequency is $2\pi F_k$.

Radiation

In obtaining a discrete-time representation of the speech signal, the radiation effects brought by the pressure at the lips is usually considered. A reasonable approximation to the radiation effects is obtained with a first order backward difference,

$$R(z) = R_0(1 - z^{-1}). \tag{2.5}$$

This radiation "load" can be cascaded with the vocal tract model $V(z)$ as depicted in Equation 2.2.

Excitation

As we have known that the majority of speech sounds can be classed as either voiced or voiceless, to generate an appropriate input to the vocal tract radiation

system, the excitation generally can produce either a quasi-periodic pulse waveform or a random noise waveform. In the case of voiced speech, the impulse train generator produces a sequence of unit impulses which are spaced by the desired fundamental period. This signal in turn excites a linear system whose impulse response $g(n)$ has the desired glottal wave shape. A gain control, A_v, controls the intensity of the voiced excitation. Concerning the choice of the form of $g(n)$, Rosenberg [21] found in a study of the effect of glottal pulse shape on speech quality that the natural glottal pulse waveform could be replaced by a synthetic pulse waveform of the form

$$\begin{aligned} g(n) &= \frac{1}{2}\big[1 - cos(\pi n/N_1)\big] && 0 \leq n \leq N_1 \\ &= cos\big[\pi(n - N_1)/2N_2\big] && N_1 \leq n \leq N_1 + N_2 \\ &= 0 && otherwise. \end{aligned} \tag{2.6}$$

Source-filter model

In practice, the speech sound waves are always sampled to a digit format in storage and transmission. In the digital model of speech signals, the glottal source and formants are often interpreted as excitation and resonance properties of the linear system; and the essence of the model is that the vocal tract imposes its resonances upon the excitation so as to produce the different sounds of speech. It is simply that a valid approach to representation of speech signals is in terms of a "source-filter" model [19] such as depicted by Figure 2.5.

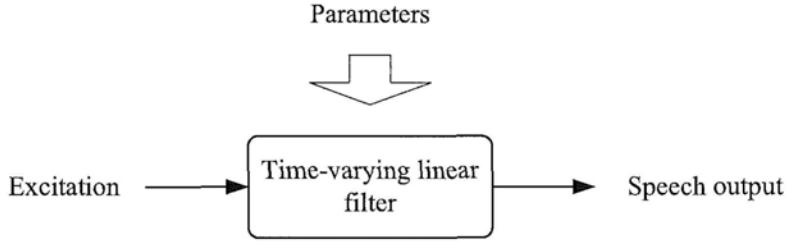


Figure 2.5: *The source-filter model of speech production.*

The important features in the acoustic theory of speech production, as the

sound generation, propagation, and radiation can in principle be solved with suitable values of the excitation and vocal tract parameters to compute an output speech waveform. Figure 2.5 shows a general block diagram which is a typical representative of numerous models that have been used as the basis for speech processing. These models all have in common that the excitation features are separated from the vocal tract and radiation features. The vocal tract and radiation effects are accounted for by the time-varying linear system. The linear filter that characterizes the model is to deliver the resonance effects that we have discussed in Section 2.2.1. The excitation in the system is a signal that is either a train of (glottal) pulses, or random noise [22]. The parameters of the source and filter are chosen so that the resulting output has the desired speech-like properties.

To produce a speech-like signal, the mode of excitation and the resonance properties of the linear filter must change with time, yet, the properties of the speech signal change relatively slowly with time. The nature of this time variation can be seen from the short time stationary characteristics of speech signal [19]. For many speech sounds, it is reasonable to assume that the general properties of the excitation and vocal tract remain fixed for periods of 10-30 *msec.* Thus, the source-filter model involves a slowly time-varying linear system excited by an excitation signal whose basic nature changes from quasi-periodic pulses for voiced speech to random noise for unvoiced speech [19].

2.3 Speaker-distinctive Characteristics

A speaker recognition system, at its most elementary level, comprises a collection of algorithms drawn from a wide variety of disciplines, including signal processing, statistical pattern recognition, among others. Although variations resulting from different recognizers exist, the greatest common denominator for all recognition systems is the signal processing front-end, which converts the speech waveform to some type of parametric representation for further analysis and processing.

There are a variety of voice attributes that characterize a speaker. Speech is a complicated signal produced as a result of several transformations occurring at several different levels: semantic, linguistic, articulatory, and acoustic [17]. Differences in these transformations appear as differences in the acoustic properties of the speech signal. Speaker-related differences are a result of a combination of anatomical differences inherent in the vocal tract and the learned speaking habits of different individuals. In speaker recognition, all these differences can be used to discriminate between speakers.

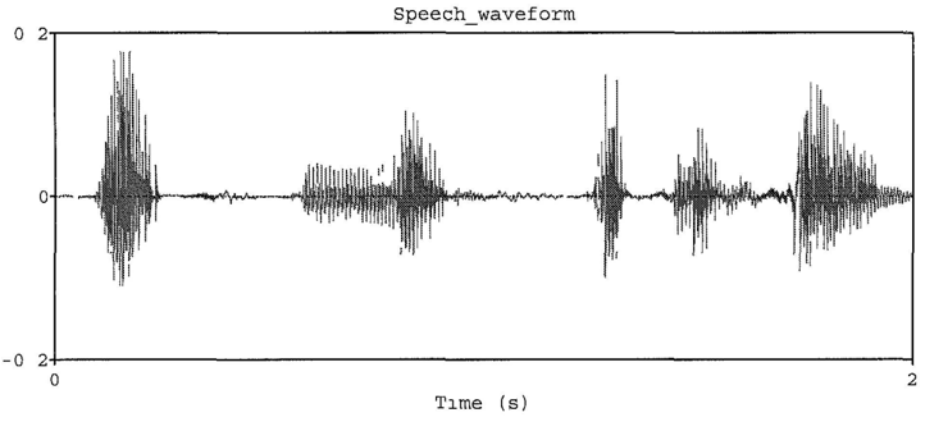
Human perception of speech sounds is a complex process, to measure the speech waveform into representative features, the consideration of the human auditory properties is essential, and the preprocessing techniques used will largely affect the feature extraction output. In this section, we will first outline the fundamentals of human-being in perceiving speaker identity. Some inevitable analysis and preprocessing approaches are introduced thereafter.

2.3.1 Human vocal attributes

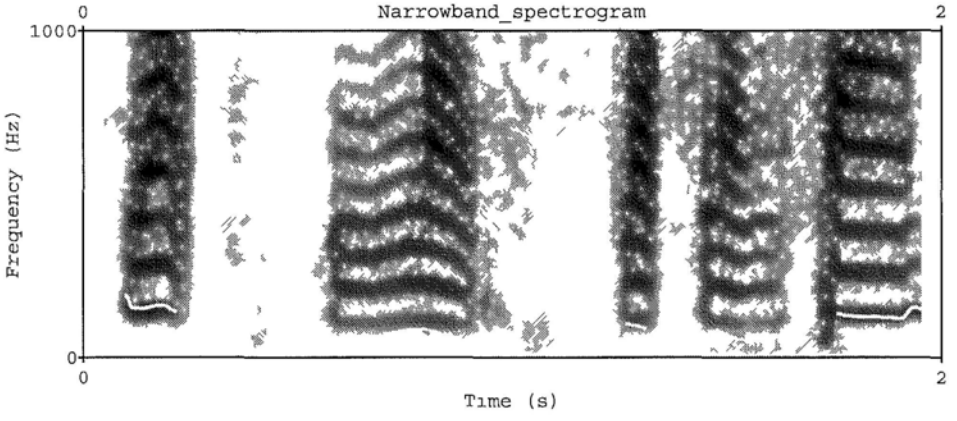
The voice has various attributes, these are chiefly *frequency*, *harmonic structure*, and *intensity*.

Fundamental frequency

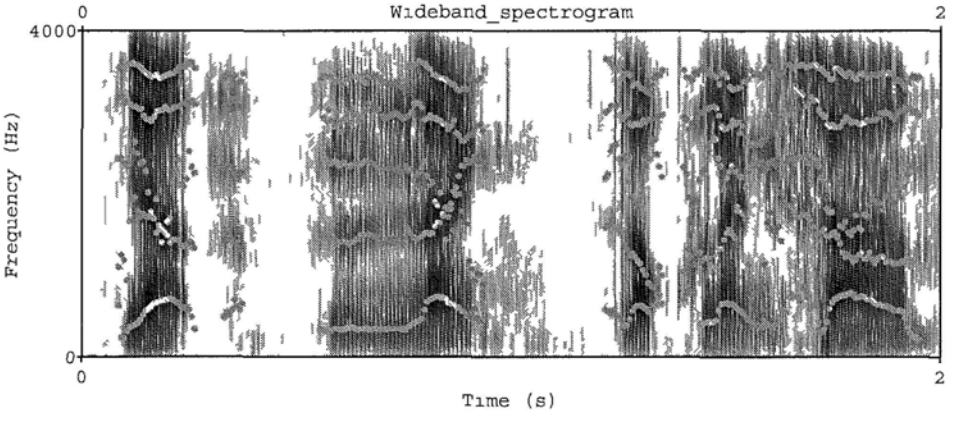
The immediate result of vocal cord vibration is the fundamental tone of the voice, which determines its pitch. In physical terms, the frequency of vibra-



(a) Speech waveform



(b) Narrow-band spectrogram Principal F0 harmonics and pitch contour manifested



(c) Wide-band spectrogram Dominant formants exhibited

Figure 2 6 *Waveform and spectrograms of a speech signal*

tion as the foremost vocal attribute corresponds to the number of air puffs per second, counted as cycles per second (cps or Hertz).

This frequency is controlled by a combination of effects, both stable and variable factors. The stable determinants of the individual voice range depend on the laryngeal dimensions as related to gender, age, and body type. The smaller a larynx, the higher its pitch range. Within this individually fixed range, variables that influence the pitch of a given phonation include: tension of the cord, force of glottal closure indicated by the glottal resistance, and expiratory air pressure. During speech, we continually alter the tension and length of the vocal cords, and the air pressure from the lungs, until we get the desired frequency. The range of vocal cord frequencies used in normal speech extends from about 60 to 350 cps, or more than two octaves [18].

As a measurement of frequency, F0 is recorded at several locations along an utterance, considering its change over time. Such measurements are undoubtedly somewhat correlated, but in addition to the average value, they also contain information about the pitch contour, which has been used for speaker recognition by Atal [23]. An illustration of pitch contour is given in Figure 2.6(b).

Harmonic structure

A second attribute of vocal sound, harmonic structure, depends on the waveform produced by the vibrating vocal cords. Like any musical instrument, the human voice is not a pure tone (as produced by a tuning fork); rather, it is composed of a fundamental tone (or frequency of vibration) and a series of higher frequencies called upper harmonics. As long as the harmonics are precise multiples of the fundamental, the vocal will sound clear and pleasant. If nonharmonic components are added, increasing degrees of roughness, harshness, or hoarseness will be perceived in relation to the intensity of the noise components in the frequency spectrum.

The primary harmonic structure is radiated into the vocal tract then. The shaping of the vocal tract determines the modulation of the voice through res-

onance and damping. As a general rule, a long and wide vocal tract enhances the lower harmonics, producing a full, dark, and resonant voice. Conversely, shortening and narrowing of the vocal tract leads to higher resonances with lightening of the voice and the perceptual attributes ranging from shrill and strident to constricted and guttural.

Intensity

Vocal intensity, the third major vocal attribute, depends primarily on the amplitude of vocal cord vibrations and thus on the pressure of the subglottic airstream. The greater the expiratory effort, the greater the vocal volume. Another component of vocal intensity is the radiating efficiency of the sound generator and its superimposed resonator. The larynx has been compared to the physical shape of a horn. This construction is most efficient in acoustical practice, as seen in the shape of wind instruments, car horns, sirens, loudspeakers, etc. A well-shaped, wide, and flexible vocal tract enhances the projective potential of the voice. Conversely, a morphologically narrow, pathologically constricted, or emotionally tightened throat produces a muffled, constricted sound with poor carrying power. The inborn automatic reflexes of laughing and yawning illustrate the resonator action of the vocal organ. Together with a widely opened mouth, flat tongue, elevated palate, and maximally widened pharynx, the larynx assumes a lowered position with maximally elevated epiglottis. This configuration is ideal for the unimpeded radiation of the vocal cord vibrations so that the resulting sound is loud and bright, with a gaily ringing quality; it is the sound of happy laughter. The opposite is present with the painfully tight-throated, choked sobbing of someone crying in despair. Generally, the resonant effects of the vocal tract on the harmonic structure is reflected by the spectral envelope of a speech segment. Individual formants usually present themselves with a peak in the spectrum. In Figure 2.6(c), the formants are reviewed through the spectro-temporal spectrogram display.

Voice quality

Apart from the variable influences of the vocal tract on the momentary vocal resonance according to training and intention, the resonator exerts a constant influence on the vocal quality by shaping its individual characteristics. The anatomical shape and the physiologic flexibility of the vocal tract serve to mold the individual vocal personality in at least two ways: by its inborn shape and by the learned behavior of using it for communication. Perceived characteristic "acoustic coloring" of voice that derived from a variety of physiological features forms clusters of identifiable voice types. Modal voice, breathy voice, pressed voice, creaky voice, tense voice, harsh voice, nasal voice are all examples of different voice types.

In English, apart from distinguishing voiced and voiceless sounds, voice quality does not make linguistic contrasts, but conveys information about the speaker. In some languages, differences in voice quality or pitch trajectory are used to convey linguistic meanings. Languages and dialects have characteristic voice qualities; personal voice quality enables a listener to recognize a particular individual. Furthermore, the quality of someone's voice also conveys emotions and attitudes.

Any individual's mother tongue shapes his articulatory behavior into certain patterns, which remain audible in all languages that he learns after puberty and constitute one aspect of the so-called foreign accent. The ability to recognize a given speaker solely by the quality and inflection of his voice is the basis of efforts to produce "voiceprint" that should be as unmistakably identifying as fingerprints are.

2.3.2 Signal processing front-end

In the process of feature extraction of a speaker recognition system, there are some common signal processing techniques that are essential as the preprocessing steps of speech signal [24]. While, some perceptual cues and auditory findings are useful for speech analysis, for example, the auditory frequency scales

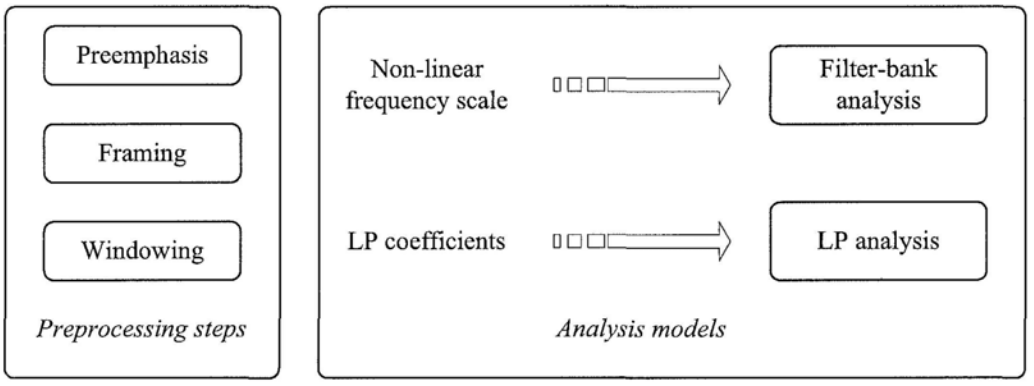


Figure 2.7: *Signal processing front-end for feature extraction - a glance.*

that employed in multi-band energy analysis. Besides, spectral analysis models, such as filter-bank model, LP model of short-term speech spectrum are viewed as important parametric representations of speaker's vocal tract-related characteristics, they are therefore considered as the core of this kind of feature extraction front-ends [25].

Preprocessing steps

1. Preemphasis

The digitized speech signal, $s(n)$, is put through a low-order digital system (typically a first-order FIR filter), to spectrally flatten the signal so as to reduce the dynamic range. A first-order filter is used in our work:

$$H(z) = 1 - 0.97z^{-1}. \tag{2.7}$$

Figure 2.8 shows the magnitude characteristics of $H(e^{j2\pi f})$. It can be seen that at $f = 1$ (half the sampling rate) there is a 36dB boost in the magnitude over that at $f = 0$.

2. Frame blocking

Speech signal is quasi-periodic in voiced segment, and it can be viewed as short-time stationary within $10 - 30$ *msec*. Hence, the preemphasized speech signal should be framed into short segment before further processing. In this

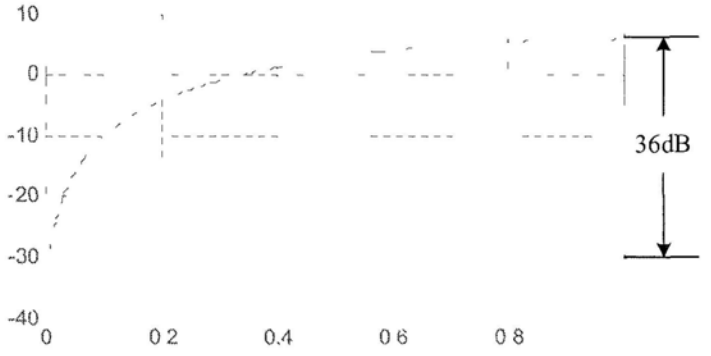


Figure 2.8: *Magnitude response of the 1st-order high-pass filter for pre-emphasis.*

step, the signal is blocked into frames of N samples, between two adjacent frames, there is an overlap of M samples. Figure 2.9 illustrates the blocking into frames for the case in which $M = (1/2)N$.

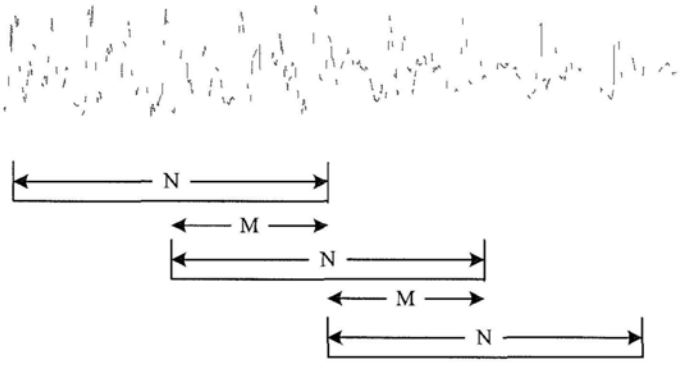


Figure 2.9: *Blocking of speech into overlapping frames ($N = 2M$).*

3. Windowing

To minimize the signal discontinuities at the beginning and end of each frame, usually there is a windowing process applied. A typical window used for the speech front-end parametrization is Hamming window, which has the following form,

$$w(n) = 0.54 - 0.46\cos(\frac{2\pi n}{N-1}), \quad 0 \leq n \leq N-1. \tag{2.8}$$

Non-linear frequency scale warping

Human ears resolve frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves recognition performance. There are many non-linear frequency scales that approximate the sensitivity of the human ear [26]. For example:

- The Mel frequency scale
- The Bark frequency scale
- The Equivalent Rectangular Bandwidth (ERB) scale

♦ **The Mel frequency scale**

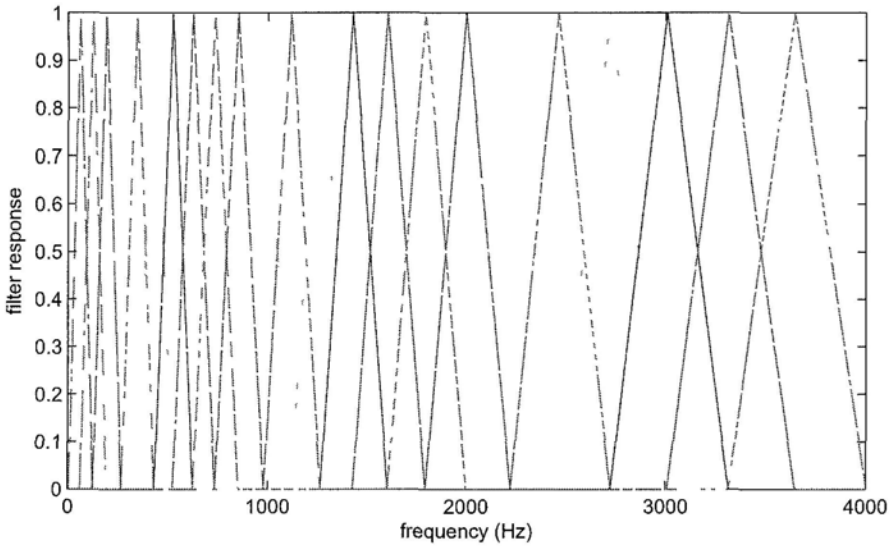


Figure 2.10: *A bank of triangular filters at the Mel frequency scale.*

In hearing sounds, human ears map the acoustic frequency, f, to a "perceptual" frequency scale. A most popular approximation to this type of mapping in speaker recognition is known as the *mel* scale:

$$mel(f) = 2595 log_{10}(1 + \frac{f}{700}) \tag{2.9}$$

The *mel* scale attempts to map the perceived frequency of a tone, or pitch, onto a linear scale. It is often approximated as a linear scale from 0 to 1000Hz, and then a logarithmic scale beyond 1000Hz. A bank of triangular bandpass filters that spaced along the *mel* scale is shown in Figure 2.10. This set of filter-bank is always employed in parameterizing speech signal into subband energy quantities.

♦ The Bark frequency scale

Another important perceptual approximation in human hearing is Bark frequency scale. The Bark scale ranges from 1 to 24 Barks, corresponding to the first 24 critical bands of hearing [27]. The center-frequencies and bandwidths of the Bark bands are to be interpreted as samplings of a continuous variation in the frequency response of the ear to a sinusoid or narrow-band noise process. That is, critical-band-shaped masking patterns should be seen as forming around specific stimuli in the ear rather than being associated with a specific fixed filter bank in the ear. The Bark scale is defined above in terms of frequency in Hertz versus Bark number.

♦ The Equivalent Rectangular Bandwidth (ERB) scale

Moore and Glasberg [28] have revised Zwicker's loudness model. The modification replaces the Bark scale by the equivalent rectangular bandwidth (ERB) scale. The *ERB* of the auditory filter is assumed to be closely related to the critical bandwidth and is defined analytically, thus, it is also more smoothly behaved than the Bark scale data.

At moderate sound levels, the *ERB* in Hertz is defined by

$$ERB(f) = 0.108f + 24.7, \tag{2.10}$$

where f is center-frequency, normally in the range 100 Hz to 10 kHz [28]. The *ERB* is generally narrower than the classical critical bandwidth.

The *ERB* scale is defined as the number of *ERBs* below each frequency [28], that is,

$$ERBS(f) = 21.4log_{10}(0.00437f + 1). \tag{2.11}$$

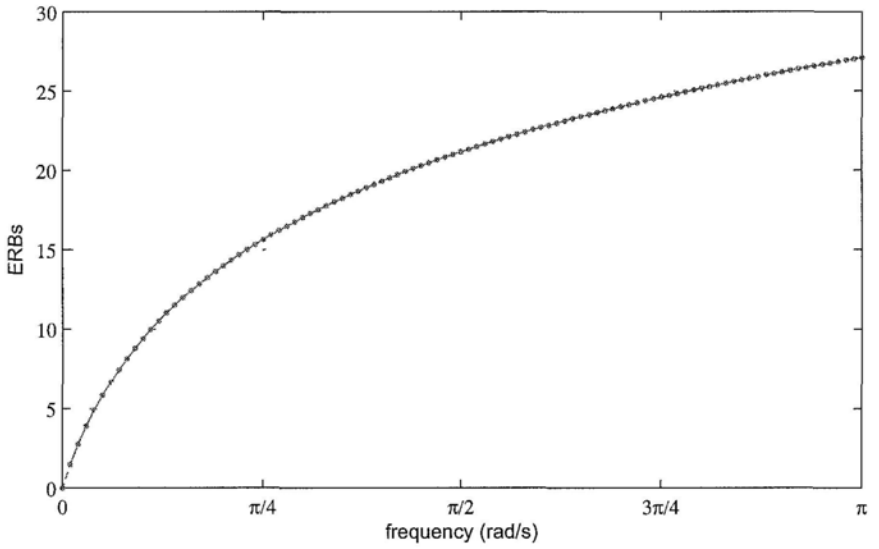


Figure 2.11: *An overlay of the normalized Bark and* ERB *frequency warping.*

An overlay of the normalized Bark and *ERB* frequency warpings is shown in Figure 2.11. The *ERB* warping is determined by scaling the inverse of Equation 2.11, evaluated along a uniform frequency grid from zero to the number of *ERBs* at half the sampling rate, so that DC maps to zero and half the sampling rate maps to π.

Auditory filter banks

Auditory frequency scale warping is closely related to the topic of auditory filter banks which are non-uniform bandpass filter banks designed to imitate the frequency resolution of human hearing [29]. Classical auditory filter banks include constant-Q filter banks such as the widely used third-octave filter bank. More recently, constant-Q filter banks for audio have been devised based on the wavelet transform, including the auditory wavelet filter bank [30]. Auditory filter banks have also been based more directly on psychoacoustic measurements, leading to approximations of the auditory filter frequency response in terms of a Gaussian function [31], a "rounded exponential" [32], and more recently the gamma-tone (or "Patterson-Holdsworth") filter bank [33]. The gamma-chirp

filter bank further adds a level-dependent asymmetric correction to the basic gamma-tone channel frequency response, thereby providing a yet more accurate approximation to the auditory frequency response [34]. Figure 2.12 illustrates the magnitude response of an 20-channeled Gamma-tone filter bank.

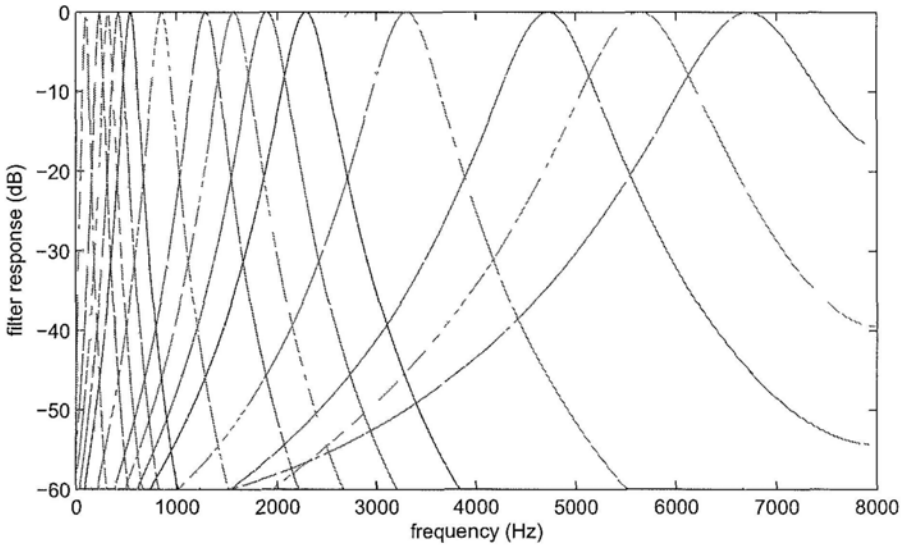


Figure 2.12: *20-channeled Gamma-tone filter banks.*

Spectral analysis models

♦ Bank-of-filters model

The overall structure of the bank-of-filters model is shown in Figure 2.13. The sampled speech signal, $s(n)$, is passed through a bank of Q bandpass filters, giving the signals

$$\begin{aligned} s_i(n) &= s(n) * h_i(n), \qquad 1 \leq i \leq Q \\ &= \sum_{m=0}^{M_i-1} h_i(m)s(n-m), \end{aligned} \tag{2.12}$$

where we have assumed that the impulse response of the i^{th} bandpass filter is $h_i(m)$ with a duration of M_i samples; hence, we use the convolution representation of the filtering operation to give an explicit expression for $s_i(n)$,

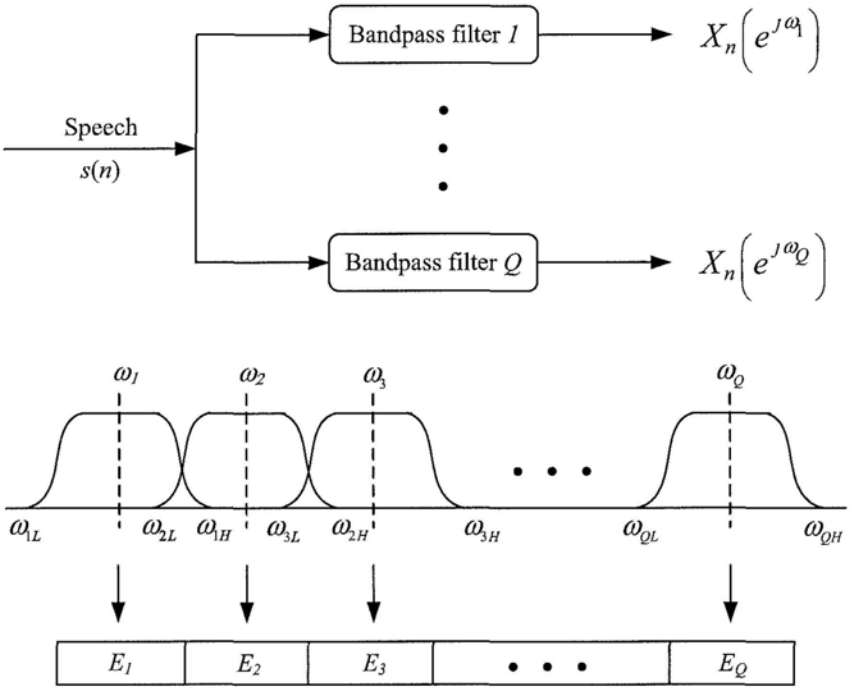


Figure 2 13 *Bank-of-filter spectral analysis model*

the bandpass-filtered speech signal The purpose of the filter-bank analyzer is to give a measurement of the energy of the speech signal in a given frequency band [22] And the model output E_i, $i = 1,$ $, Q$ will be useful in the feature parametrization process in later discussion

♦ LP analysis model

The basic ideal behind the LP model is that a given speech sample at time n, $s(n)$, can be approximated as a linear combination of the past p speech samples, such that

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + \quad + a_p s(n-p), \tag{2 13}$$

where the coefficients $a_1, a_2,$ $, a_p$ are assumed constant over the speech analysis frame We convert Equation 2 13 to an equality by including an excitation term, $Gu(n)$, giving

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + Gu(n), \tag{2 14}$$

where $u(n)$ is a normalized excitation and G is the gain of the excitation. By expressing Equation 2.14 in the z-domain we get the relation

$$S(z) = \sum_{k=1}^{p} a_k z^{-k} S(z) + GU(z) \tag{2.15}$$

leading to the transfer function

$$H(z) = \frac{S(z)}{GU(z)} = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} = \frac{1}{A(z)}. \tag{2.16}$$

The interpretation of Equation 2.16 is given in Figure 2.14, which shows the normalized excitation source, $u(n)$, being scaled by the gain, G, and acting as input to the all-pole system, $H(z) = \frac{1}{A(z)}$, to produce the speech signal, $s(n)$. $H(z)$ and $A(z)$ are named LP filter and LP inverse filter, respectively. Based on our knowledge that the actual excitation function for speech is essentially either a quasi-periodic pulse train (for voiced speech sounds) or a random noise source (for unvoiced sounds), the appropriate synthesis speech model, corresponding to the LP analysis, is as shown in Figure 2.15. Here the normalized excitation source is chosen by a switch whose position is controlled by the voiced/unvoiced character of the speech, which chooses either a quasi-periodic train of pulses as the excitation for voiced sounds, or a random noise sequence for unvoiced sounds. The appropriate gain, G, of the source is estimated from the speech signal, and the scaled source is used as input to a digital filter ($H(z)$), which is controlled by the vocal tract parameters characteristic of the speech being produced. Thus the parameters of this model are voiced/unvoiced classification,

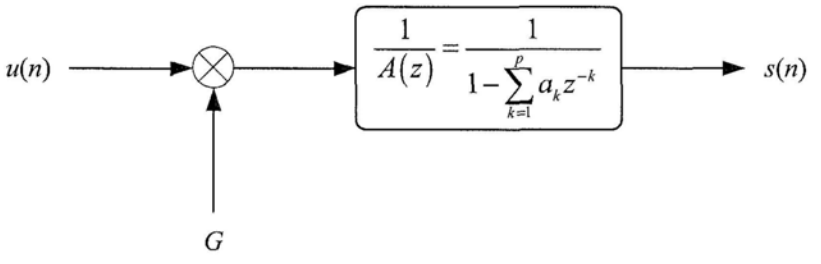


Figure 2.14: *All-pole LP speech model.*

pitch period for voiced sounds, the gain parameter, and the coefficients of the digital filter, a_k [20]. These parameters all vary slowly with time.

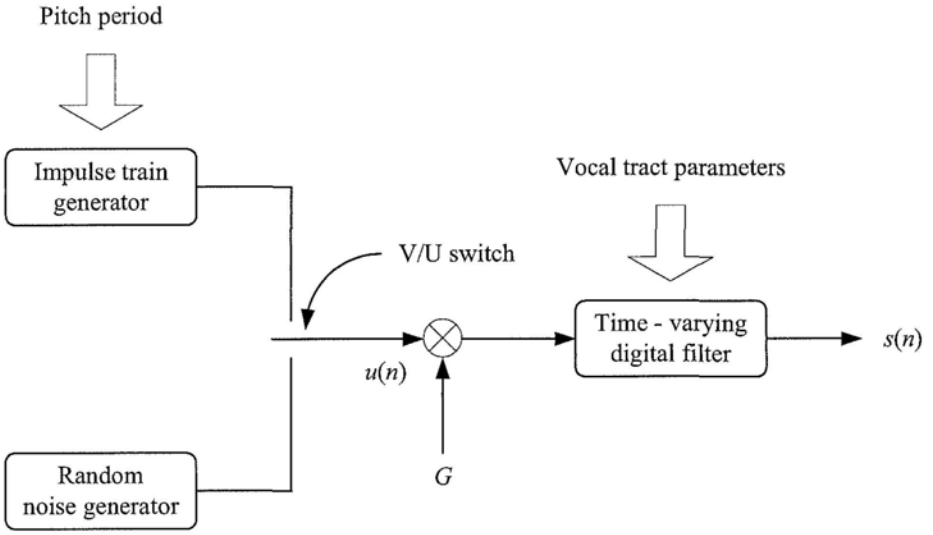


Figure 2.15: *LP-based speech synthesis model.*

As an inverse process of the LP synthesis, the LP analysis model is illustrated by Figure 2.16. It is due to the fact that the spectra of pulse trains and random noises are both flat, the all-pole filter $H(z)$ actually determines the spectral envelope of the speech signal. Thus, the purpose of the LP analyzer is to give a representation of the speech spectral envelope in a set of prediction coefficients, $a_k,\ k = 1, \ldots, p$.

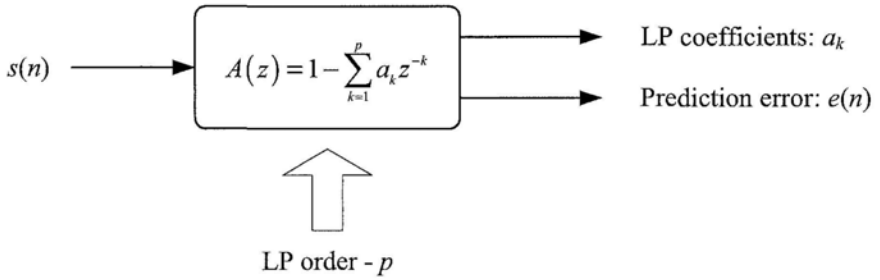


Figure 2.16: *The LP analysis model.*

Energy, delta and acceleration coefficients

Usually, to augment the spectral parameters derived from filter-bank or LP analysis model, an energy term can be affiliated to the feature vector. The energy of an analysis frame is computed as the log of the frame energy, that is, for speech frame of samples $s(n),\ n = 0, \ldots, N-1$,

$$E = \log \sum_{n=0}^{N-1} s^2(n). \tag{2.17}$$

The basic static parameters of the speech spectrum provides a good representation of the local spectral properties of the signal for the given analysis frame. However,an improved representation can be obtained by extending the analysis to include information about the temporal cepstral derivative (both first and second derivatives, which are also termed as delta and acceleration coefficients). The delta coefficients are computed using the following regression formula

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta(c_{t+\theta} - c_{t-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2}, \tag{2.18}$$

where d_t is a delta coefficient at time t computed in terms of the corresponding static coefficients $c_{t-\Theta}$ to $c_{t+\Theta}$. Θ is the delta window. The acceleration coefficients can also be obtained by applying the same equation. Since Equation 2.18 relies on past and future speech parameter values, some modification is needed at the beginning and end of the speech. Usually a replica of the first or the last vector is used to fill the regression window in this case.

2.4 Review of Feature Representations

The function of the measurement phase of a speaker recognition system is to perform a number of characterizing measurements on the voice pattern under test. An ideal speaker-discriminative feature representation is expected to [35][36]:

- have large inter-speaker variability and small intra-speaker variability,
- be reasonably robust to background noise and distortions,
- occur naturally and frequently in normal speech,
- be easily measurable,
- be stable over time or not be affected by the speaker's health/mood,
- be difficult to mimic.

Dimension of the feature vector should also be relatively low. For statistical speaker modeling method, such as the Gaussian mixture models, the number of required training samples for reliable density estimation grows exponentially with the dimension of feature. This problem is known as the *curse of dimensionality* [37]. The computational savings are also obvious with low-dimensional features.

Differences in voices stem from two broad bases: organic and learned differences. Organic differences are the result of variations in the sizes and shapes of the components of the vocal tract: larynx, pharynx, tongue, teeth, and the oral and nasal cavities. Since the resonances of the vocal tract and the characteristics of the sound energy sources depend upon just these anatomical factors, organic differences lead to differences in fundamental frequency, laryngeal source spectrum, and formant frequencies and bandwidths. Learned differences are the result of differences in the patterns of coordinated neural commands to the separate articulators learned by each individual. Such differences give rise to variations in the dynamics of the vocal tract such as the rate of formant transitions and coarticulation effects. Naturally, many speaker-dependent characteristics are affected by both of these factors.

There are much efforts that have been devoted to exploiting different measurement schemes for speaker-distinguishable characteristics delivered by speech utterances throughout the years. Wolf in [36] has categorized the measurement schemes of speaker-dependent parameters into three main classes. Kinnunen *et al.* [38] lately made a survey and summarized the primary speech features that have been employed for speaker recognition purposes. Besides the classical and leading features, some recently derived parameter sets have also been included. The features are generally labeled as five clusters from the viewpoint of their physical interpretation. Their physical meanings and the typical processing approaches applied are briefly described as follows:

- **Short-term spectral features**

  ◊ Describing the short-term spectral envelope, is an acoustic correlate of timbre, as well as the resonance properties of the vocal tract.

  ◊ Auditory frequency warping, bank-of-filters model, LP spectral analysis, dynamic coefficients appending, etc.

- **Voice source features**

  ◊ Characterizing the glottal flow.

  ◊ Pitch determination, pitch-synchronous analysis, pitch-epoch localization, etc.

- **Spectro-temporal features**

  ◊ Interpreting speaker properties in flexible time-frequency resolutions.

  ◊ Subband energy separation, multiple frequency band demodulation, etc.

- **Prosodic features**

  ◊ Including pitch, intonation, duration and rhythm, usually span over tens or hundreds of milliseconds.

  ◊ F0 tracking, dynamic coefficients appending, etc.

- **High-level features**

◊ Attempting to capture conversation-level characteristics of speakers.

◊ Speech recognizer, statistical language modeling, etc.

Generally speaking, short-term spectral and voice source features are relatively easy to extract, and there is no need for huge amount of data. Up until now, the short-term spectral features have always dominated the front-end of the leading speech, speaker, even language recognition systems. Besides the stable performance provided, their low demand on computational cost makes the real-time application feasible. However, the biggest challenge they are faced is the parameter degradation in presence of background noise and in channel mismatch conditions. Prosodic and high-level features are believed to be more robust, but less discriminative and easier to impersonate. High-level features, since connecting with the personalized lexicon and recording the idiolectal pattern of individual speakers, are less affected by the variation in noise or channel conditions. The high-level speaker-related characteristics are whereas difficult to extract, and there will be a lot of training data needed in the feature extraction process. Thus, it is hard to apply this genus of features into real-time recognition tasks considering their delay in making the decisions. To conclude, there does not yet exist globally "best" feature but the choice is a trade-off between speaker discrimination, robustness, and practicality.

2.4.1 Short-term spectral features

In early 1980s, Mel-frequency cepstral coefficients (MFCCs) [8] were introduced by Davis *et al.* for speech recognition and then adopted in speaker recognition. This set of parameters exploits auditory principles, as well as the decorrelating property of the cepstrum. In addition, the mel-cepstrum is amenable to compensation for convolutional channel distortion. As such, the MFCCs have proven to be one of the most successful feature representations in speech-related recognition tasks. The extraction of MFCC static coefficients is illustrated by Figure 2.17.

As another principal spectral analyzer for speech signal, the LP coefficients

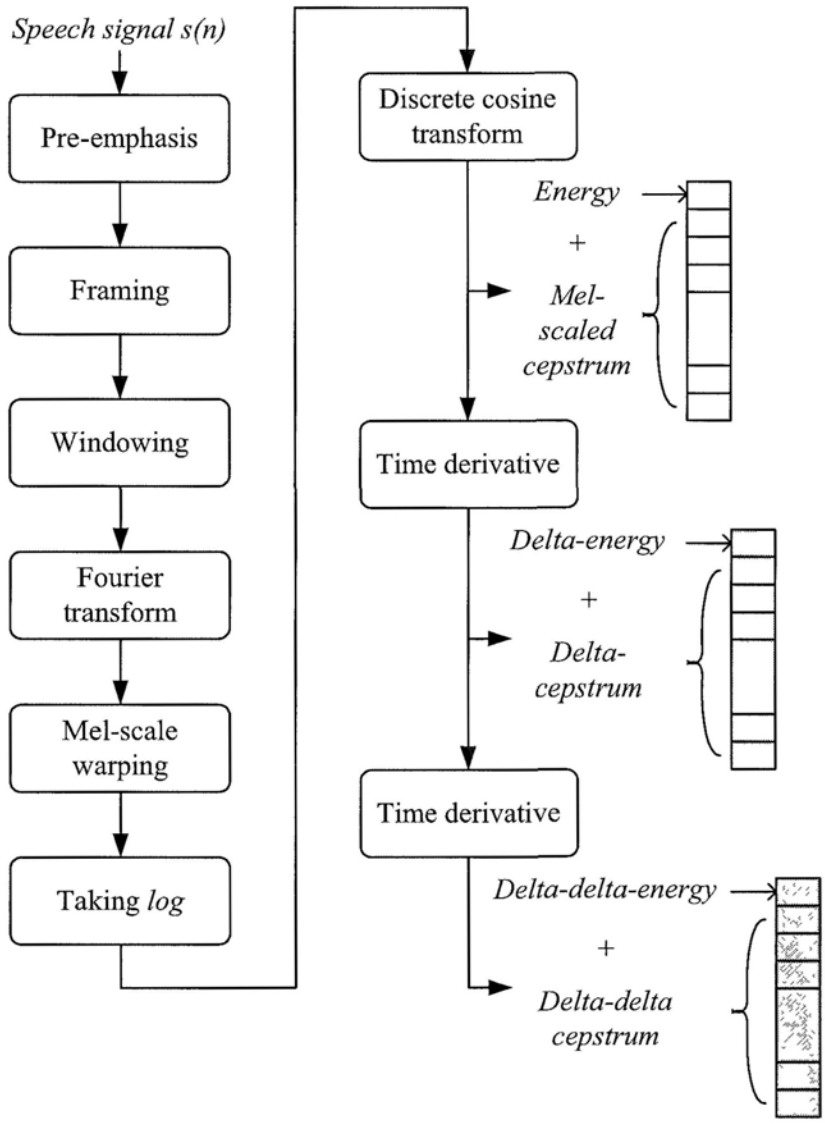


Figure 2.17: *Mel-frequency cepstrum and their delta, delta-delta coefficients.*

$\{a_k\}$ are rarely used as features directly, but they are transformed into robust and less correlated features such as linear predictive cepstral coefficients (LPCCs) [9]. Alternatively, line spectral pair (LSP) [10], [11], and perceptual linear prediction (PLP) coefficients [39] are all describing the characteristics of spectral envelope of individual speakers.

2.4.2 Voice source features

Voice source features characterize the glottal excitation signal of voiced sounds such as glottal pulse shape and F0. It is long believed that voice source features carry speaker-discriminative information. Besides F0, which delivers message about the rate of vocal cords vibration, the other vocal source-related parameters mainly focus on the glottal pulse shape. Glottal pulse shape solely is the most important aspect of speech in the way it affects our perception of voice quality. Mathematical representations, among which the Liljencrants-Fant (LF) glottal pulse model [40] is the most commonly employed, were built for derivative of glottal pulse airflow. Primary indexes in recording the pulse shape include pitch period, open quotient, glottal closure instant, etc.

The glottal features are not measurable due to the vocal tract modulating effects. LP inverse filtering is usually employed to distinguish the glottal source from those effects by the vocal tract, such as those reported in [41], [42], [43], [44]. Other methods used include the closed-phase covariance analysis on the vocal cords close portions [45]. In [42], the glottal flow parameters were employed as the speaker identification features, while, speaker-dependent parameter sets were usually parameterized by analysis methods like auto-associative neural network [46], pitch-synchronous wavelet transform [43], cepstral analysis [45], [44], Hilbert transform [47], etc.

As reported by many, speaker recognition systems seldom depend solely on the vocal source features, since they are not discriminative as vocal tract features. Nevertheless, their complementarity with the vocal tract parameters are broadly identified, thus, fusing these two information sources can improve accuracy [43], [47]. Another advantage of the vocal source features lie in their

less affiliation with the phonetic content of spoken utterances. Therefore, they require less on the phonetic balance and amount of training data set [46].

2.4.3 Spectro-temporal features

Spectro-temporal details that capture formant transitions, formant-harmonic interactions of a voiced interval of speech signal are useful speaker-specific information. Furui in 1981 [12] has come up with a way to incorporate some temporal information to the cepstral features, i.e., to add the dynamic and acceleration coefficients. This method has been found successful and widely employed in speech related recognition applications up until now. Apart from this and other similar approaches like regression line fitting [22] which all target on the Fourier analysis based features, time-frequency processing front-ends were brought in for extracting features with flexible spectro-temporal resolutions in recent years. To name a few representatives, there are time-frequency principal components [48], data-driven temporal filters [49], temporal discrete cosine transform [50], pitch-synchronous wavelet transform [43], modulation analysis [51], [52], [53], [54], [55], [56], [57], [58], [59], [60], etc. It is worth mentioning that the spectro-temporal feature extraction method has also extended their capability to exploit vocal source features, for instance, in works of Zheng *et al.* [43], and Wang *et al.* [61]. Like the vocal source features, fusion of these features with the cepstral coefficients are still necessary in recognition systems.

2.4.4 Prosodic features

Prosody refers to non-segmental aspects of speech, including for instance syllable stress, intonation patterns, speaking rate and rhythm. Prosody dislikes the short-term spectral features first in that it always span over long segments like syllables, words, and utterances. Then, prosody-related features usually convey speaker-specific information about the speaking style, language background, sentence type and emotions.

The first and foremost prosodic parameter is F0. It reflects the vibration condition of the speaker's vocal cords, which not necessarily depend on the spo-

ken context, but its range, variation pattern are speaker-dependent. Since F0 is a one-dimensional feature, it is non expected to be very discriminative mathematically. Therefore, different kinds of F0-related feature vector has been proposed throughout these years. Atal in [23] has used pitch contour for identifying speakers in early 1970s. F0 dynamics is another important form, which models both the local and long-term temporal variations of F0 [62], [63], [64].

2.4.5 High-level features

It is found that speakers differ not only in their voice timbre, speaking style, but also in their lexicon, i.e., idiolect. Doddington [65] in 2001 initiated the research on high-level conversional features, the idiolect, for speaker recognition. This kind of high-level modeling converts each utterance into a sequence of tokens where the co-occurrence patterns of tokens characterize speaker differences. To list a few, the tokens accounted include words [65], phones [66], prosodic gestures [63], [64], and articulatory tokens, e.g., articulation manner & place [67].

2.5 Speaker Modeling Techniques

In speaker recognition systems, speaker models are constructed from the extracted features. When enrolling a speaker into the system, a model of the voice, based on the extracted features, is generated and stored. Then, to identify or authenticate a speaker, the matching algorithm compares/scores the incoming speech signal with the model of the claimed speaker.

There are two types of models: template models and stochastic models.

♦ **Template models**

The simplest template model consists of a single template $\bar{x}$, which is the model for a frame of speech. The match score between the template $\bar{x}$ for the claimed speaker and an input feature vector $\boldsymbol{x}_i$ from the observation (a collection of feature vectors from the unknown speaker) is given by $d(\boldsymbol{x}_i, \bar{x})$. The model for the claimed speaker could be the centroid (mean) of a set of N training vectors

$$\bar{x} = \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i. \tag{2.19}$$

Many different distance measures between the vectors $\boldsymbol{x}_i$ and $\bar{x}$ can be expressed as

$$d(\boldsymbol{x}_i, \bar{x}) = (\boldsymbol{x}_i - \bar{x})^T \boldsymbol{W} (\boldsymbol{x}_i - \bar{x}) \tag{2.20}$$

where $\boldsymbol{W}$ is a weighting matrix. If $\boldsymbol{W}$ is an identity matrix, the distance is *Euclidean*; if $\boldsymbol{W}$ is the inverse covariance matrix corresponding to mean $\bar{x}$, then this is the *Mahalanobis distance*. The most popular method to compensate for speaking-rate variability in template-based systems is known as Dynamic Time Warping (DTW) [6]. Another form of template model uses multiple templates to represent frames of speech and is referred to as Vector Quantization (VQ) codebook modeling [7]. And a new method combining the strengths of the DTW and VQ methods is called Nearest Neighbors (NN) [68].

♦ **Stochastic models**

Using a stochastic model, the pattern-matching problem can be formulated as measuring the likelihood of an observation given the speaker model. One

way to represent the speaker is to model the distribution of feature vectors that extracted from the speaker's speech using a Gaussian mixture density, this is regarded as GMM-based speaker model [13]. In recent years, GMM-based speaker modeling have been applied widely, and it consistently produced state-of-the-art performance.

The basis for both the identification and verification systems is the GMM used to represent speakers. For a D-dimensional feature vector denoted as $\boldsymbol{x}$, the mixture density for speaker s is defined as

$$p(\boldsymbol{x}|\lambda_s) = \sum_{i=1}^{M} p_i^s b_i^s(\boldsymbol{x}). \tag{2.21}$$

The density is a weighted linear combination of M component uni-modal Gaussian densities, $b_i^s(\boldsymbol{x})$, each parameterized by a $D \times 1$ mean vector, $\boldsymbol{\mu}_i^s$, and a $D \times D$ covariance matrix, Σ_i^s;

$$b_i^s(\boldsymbol{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_i^s|^{1/2}} \times \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i^s)^T(\Sigma_i^s)^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i^s)\}. \tag{2.22}$$

The mixture weights, p_i^s, furthermore satisfy the constraint $\sum_{i=1}^{M} p_i^s = 1$. Collectively, the parameters of speaker s's density model are denoted as $\lambda_s = \{p_i^s, \boldsymbol{\mu}_i^s, \Sigma_i^s\}$, $i = 1, \ldots, M$.

Maximum likelihood speaker model parameters are estimated using the iterative Expectation-Maximization (EM) algorithm [69]. Generally five iterations are sufficient for parameter convergence.

The Gaussian components can be considered to be modeling the underlying broad phonetic sounds which characterize a person's voice. The following characteristics of GMM justify its effectiveness in modeling speakers:

1. The GMM can be viewed as a hybrid between two effective models for speaker recognition: a uni-modal Gaussian classifier and a vector quantizer codebook. The GMM combines the robustness and smoothness of the parametric Gaussian model with the arbitrary density modeling of the non-parametric VQ model.

2. The GMM can also be viewed as a single-state HMM with a Gaussian

mixture observation density or an ergodic Gaussian observation HMM with fixed, equal transition probabilities.

2.6 Performance Evaluation of Speaker Recognition System

This section first describes the different tasks that speaker recognition research nowadays are mainly involved in, and then the associative evaluation metrics are introduced.

2.6.1 Speaker recognition tasks

Identification system

The identification system is a straight-forward maximum-likelihood classifier [13]. For a reference group of S speakers $\Psi = \{1, 2, \ldots, S\}$ represented by models $\lambda_1, \lambda_2, \ldots, \lambda_S$, the objective is to find the speaker model which has the maximum posterior probability for the input feature vector sequence, $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$. The minimum error Bayes' decision rule for this problem is

$$\hat{s} = \arg \max_{1 \leq s \leq S} Pr(\lambda_s | X) = \arg \max_{1 \leq s \leq S} \frac{p(X|\lambda_s)}{p(X)} Pr(\lambda_s). \tag{2.23}$$

Assuming equal prior probabilities of speakers, the terms $Pr(\lambda_s)$ and $p(X)$ are constant for all speakers and can be ignored in the maximum. Using logarithms and the assumed independence between observations, the decision rule becomes

$$\hat{s} = \arg \max_{1 \leq s \leq S} \sum_{t=1}^{T} \log p(\boldsymbol{x}_t | \lambda_s), \tag{2.24}$$

in which $p(\boldsymbol{x}_t|\lambda_s)$ is given in Equation 2.21. A block diagram of the speaker identification system is shown in Figure 2.18.

GMM-UBM verification system

- **Log likelihood ratio detector**

Speaker verification problem requires a binary decision (detection) based on two hypothesis, i.e., the input voice came from the claimed speaker, hypothesis H_0, or *not* from the claimed speaker, hypothesis H_1. Cast in a hypothesis testing framework, for a given input utterance $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ and a claimed identity, the choice is between H_0 and H_1:

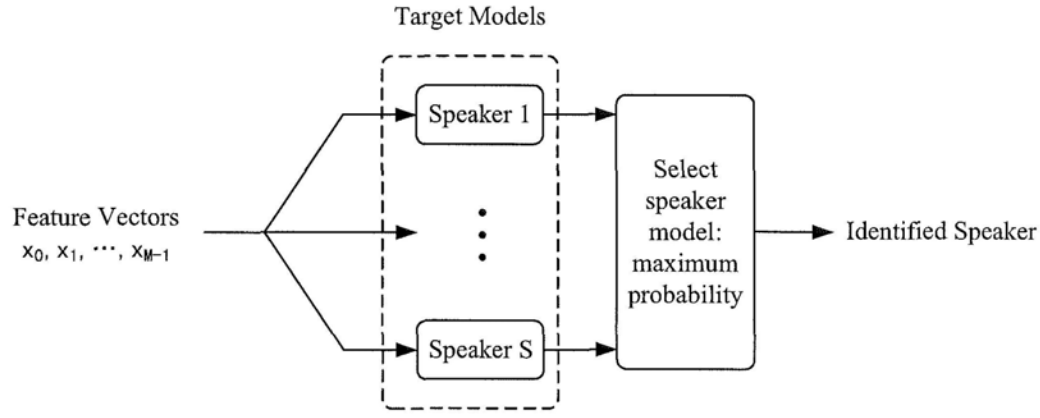


Figure 2.18: *GMM-based speaker identification system.*

H_0 : X is from the claimed speaker.

H_1 : X is *not* from the claimed speaker.

In this hypothesis test, an implicit assumption is that X contains speech from only one speaker. Thus, the task can be termed single-speaker detection. The optimum test to decide between the two hypotheses is a likelihood ratio test given by

$$\frac{p(X|H_0)}{p(X|H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{reject } H_0, \end{cases} \tag{2.25}$$

where $p(X|H_i)$, $i = 0, 1$, is referred to as the *likelihood* of the hypothesis H_i given the utterance X. The decision threshold for accepting or rejecting H_0 is θ. The basic goal of a speaker verification system is to compute the two likelihoods, and then determine a decision threshold to accept or reject the identity claim [70].

For an utterance $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ and a claimed speaker identity with corresponding model λ_C, the likelihood ratio is

$$\frac{Pr(\text{X is from the claimed speaker})}{Pr(X \text{ is } \textit{not} \text{ from the claimed speaker})} = \frac{Pr(\lambda_C|X)}{Pr(\lambda_{\bar{C}}|X)}. \tag{2.26}$$

Applying Bayes' rule and discarding the constant prior probabilities for claimant and imposter speakers, the likelihood ratio in the log domain becomes

$$\Lambda(X) = \log p(X|\lambda_C) - \log p(X|\lambda_{\bar{C}}). \tag{2.27}$$

The term $p(X|\lambda_C)$ is the likelihood of the utterance given it is from the claimed speaker and $p(X|\lambda_{\bar{C}})$ is the likelihood of the utterance given it is not from the claimed speaker. The likelihood ratio is compared to a threshold θ and the claimed speaker is accepted if $\Lambda(X) > \theta$ and rejected if $\Lambda(X) < \theta$ as described by Equation 2.25.

♦ Universal Background Model (UBM)

The GMM-UBM system use a single, speaker-independent background model to represent $p(X|\lambda_{\bar{C}})$, which is termed as the *Universal Background Model* (UBM) [16]. UBM is a large GMM trained to represent the speaker-independent distribution of features. There is no objective measure to determine the right speaker population or amount of speech to use in training a UBM. But empirically, it is preferable to select speech that is reflective of the expected alternative speech to be encountered during recognition. This applies to both the type and the quality of speech, as well as the composition of speakers. In training the UBM, the GMM mixture parameters are computed from the statistic estimates of the training data.

♦ Bayesian adaptation of speaker model

In the GMM-UBM system, we derive the target speaker model by adapting the parameters of the UBM using the speaker's training speech and some model adaptation methods. *Maximum a Posteriori* (MAP) [71], [72] and *Maximum Likelihood Linear Regression* (MLLR) [73] are the two model adaptation methods that perform the best. In this thesis, we adopt the Bayesian adaptation approach, i.e., the MAP estimation. Unlike the standard approach of maximum likelihood training of a model for the speaker independently of the UBM, the basic idea in the adaptation approach is to derive the speaker's model by updating the well-trained parameters in the UBM via adaptation. The updating is achieved by combining the old statistics from the UBM mixture parameters with the new statistic estimates that are extracted from the target speaker's training data.

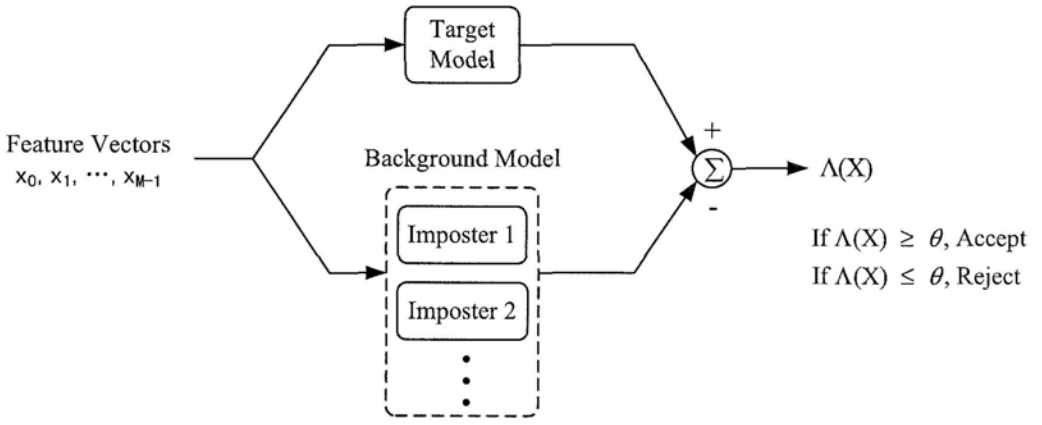


Figure 2.19: *GMM-based speaker verification system.*

A block diagram of the GMM-UBM speaker verification system is shown in Figure 2.19.

2.6.2 Performance evaluation metric for different tasks

To compare the performance of different speaker recognition system, standard evaluation metrics are indispensable. The evaluation processes are task-oriented, and are derived from the classification mechanism of the recognition task.

Speaker Identification (SID)

Speaker identification is the task of deciding, given a sample of speech, who among many candidate speakers said it. This is an N-class decision task, where N is the number of candidate speakers.

It is straightforward to use the identification error rate, $IDER$, as the performance measure of speaker identification task:

$$IDER = \frac{\text{number of misidentified trials}}{\text{total number of identification trials}} \times 100\% \qquad (2.28)$$

Speaker Verification (SV)

Speaker verification is a detection task, it can be viewed as involving a tradeoff between two error types: false acceptance (FA) and false rejection (FR). Generally, a decision threshold is selected such that the false acceptance rate equals to the false rejection rate, and this error rate is usually referred to as equal error rate (EER) [74].

2.7 Summary

As an introduction to speaker recognition system, in this chapter, we first introduce the motivation, formulation as well as the tasks under the framework of automatic speaker recognition. Then, we addressed the principal components needed when establishing a speaker recognition system, and an overview of the system setup was given. The speaker-discriminative feature extraction, as the core of our work, was discussed in detail, a literature review in the pertinent research area was made then. Subsequently, the pattern classification approaches that used in speaker identification and verification tests were presented systematically. Finally, performance evaluation metrics, as dispensable elements in speaker recognition tasks, are described respectively.

Chapter 3

Robustness of Speaker Recognition System

Reliable performance is expected from speaker recognition systems that operate in real-world applications. This requires their robustness against environmental noise and handset/channel mismatch. State-of-the-art speaker identification and verification systems can perform well in clean and perfect conditions, however, the severe degradations resulted from various mismatches or insufficiency of distinguishable acoustic attributes prevent them from performing as sole indicator in rigorous identity authentication applications.

This chapter focuses on the methods for enhancing the robustness of a speaker recognition system. The adverse effects that need to be dealt with include those resulted from background noises, transmission channel variations, training and testing data mismatch etc. Essential robust speech processing techniques required in the three main parts of a recognition system, viz, feature extraction front-end, discriminative speaker model training, and score normalization will be described in Section 3.2 through Section 3.4, respectively.

An overview of the interference sources for speech communication, and a general framework of the compensation processing approaches employed in the feature, model and match-score domains are given in Figure 3.1. In this section, we shall focus on the two adverse scenarios highlighted in gray as shown in Figure 3.1, namely speech signals with background additive noise and transmission

channel with convolutive noise. While, the three compensation methods colored in light green in the figure will also be discussed in greater detail.

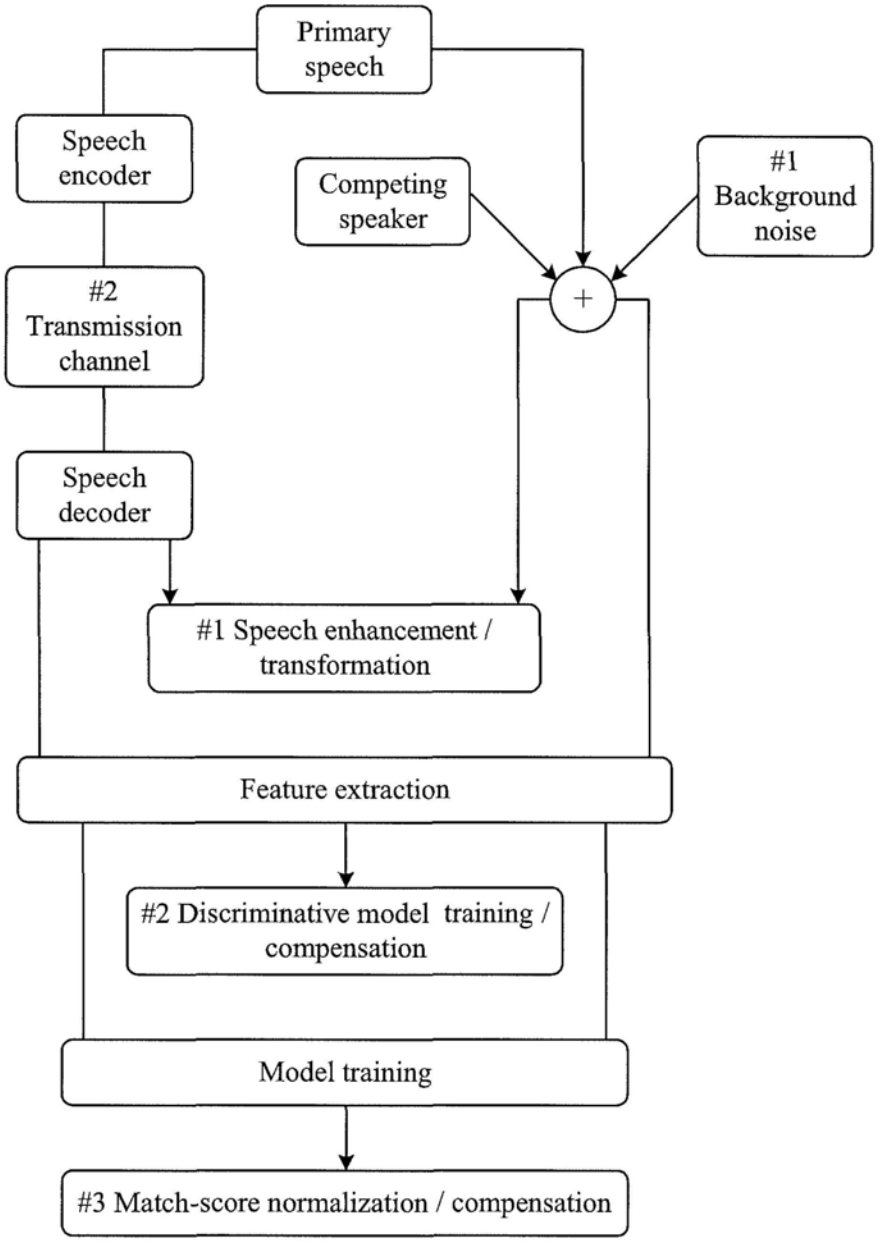


Figure 3.1: *Interference sources and compensation processing approaches in robust speaker recognition.*

The relationship between primary speech signal and interference source in

this two scenarios are formulated respectively as:

- Additive noise corruption: $y(n) = s(n) + d(n)$
- Convolutive channel distortion: $z(n) = s(n) \otimes c(n)$

3.1 Different Scenarios of Environmental-robust Speaker Recognition

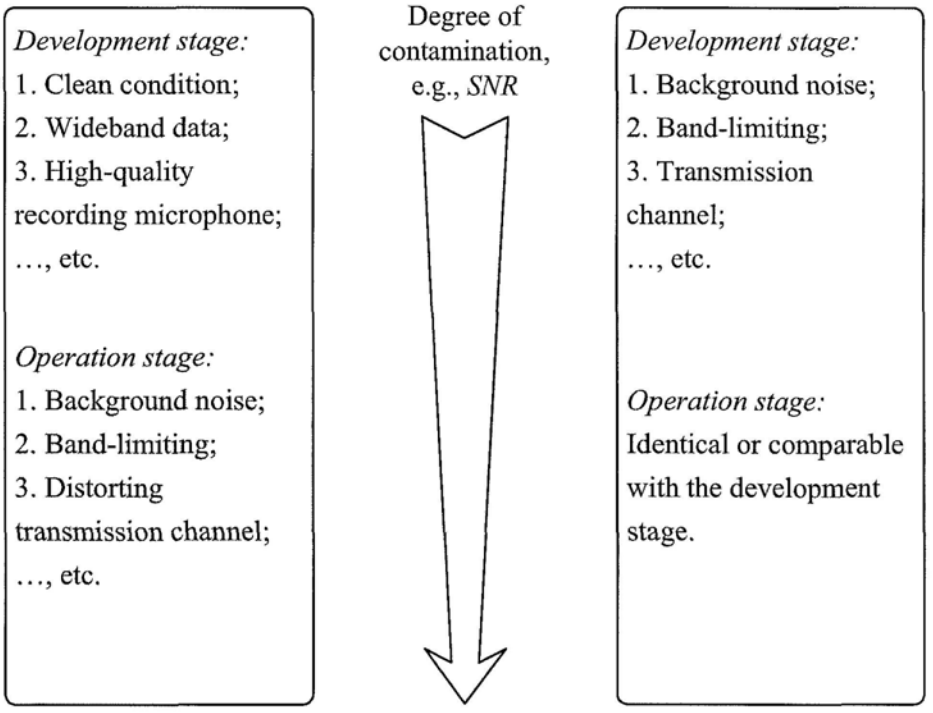


Figure 3.2: *Robustness scenarios faced in speaker recognition system realization.*

In order to deploy speech technologies for practical applications, environmental robustness is an important issue that needed to be addressed. Figure 3.2 outlines the two conditions in which robustness of a speaker recognition system is expected and deemed absolutely necessary. It is generally assumed that speaker models are built in ideal conditions as described in the left-hand side of the figure. However, it is well known that a speaker recognition system often fails to maintain reasonable performance as the acoustic conditions of its operating environment depart from the ones that were used for the training. This phenomenon is observed not only for applications in which the system is used in very noisy environments, but also for those with more subtle changes in the acoustic environment that cause no problems for human listeners. It

is observed that the huge the inconsistence between the acoustics of the test utterances and the training data, the severer degradation will occur.

The right-hand side of Figure 3.2 gives a slightly different story. It is found that the system is operating in a similar or even identical acoustic environment with that of the development phase, however, undesirable factors enter into the environment, in whatever way. With the deterioration of the environmental conditions, the system performance get impaired as well as that in the first case, although, they usually differ in degradation degrees. Comparison of the two scenarios reveals that mismatch in acoustic conditions inflicts the largest damage on a speaker recognition system, however, their consistence still can hardly maintain the original performance if environment degrades. This is therefore thought to be due to the decline of the discriminative power possessed by both the speaker models and the testing data and the resultant ambiguity in speaker boundaries.

In the following sections, we will focus on compensation algorithms developed to eliminate the effects induced by various environmental and transmission factors. There is no clear owner affiliation of them to either of the above mentioned scenarios, rather, they are flexibly and widely engaged in for robust speaker discrimination purposes. Another important issue that should be addressed on employing the compensation methods is the tradeoff exists between the quantity of unreliable information that can be removed from the speaker features versus the speaker specific information that can be preserved, to achieve optimal recognition. As it is noted in [75] that, for clean speech using the same microphone for recognitions, many of the enhancement techniques may actually reduce system performance.

3.2 Robust Feature Extraction

Speaker-specific parameters are the input for a speaker recognition system. When speech signals corrupted with background noise, or transmitted by an unknown handset/channel are taken as the source for the feature extraction front-end, undesirable factors inevitably penetrate into the feature vector output, which directly leads to performance degradation of the system. It is thus seen that in order to investigate and mitigate the adverse effects environmental noises or transmission channels lead to a speaker recognition system, studies about their effects on the pertinent speaker-discriminative parameters will play an indispensably significant role.

The topic of robust feature extraction pertains to at least two aspects, they are, to have the speech signal "cleaned" first, and then to perform feature extraction on the enhanced signal; besides, to transform speech signals into certain forms and identify the most speaker-specific components therein, then extract them exclusively out from contaminated speech utterances. The essence of the first strategy lies in the efficacy of speech enhancement algorithms employed for recovering or reconstructing the speech signals. The more speech properties essential for discriminating different speakers are retained in the processed speech, the more beneficial the subsequently extracted speaker characteristics are for the recognition accuracy.

3.2.1 Feature enhancement

In many speech communication settings, the presence of background interference causes the quality or intelligibility of speech to degrade. In a quiet environment, information exchange between a speaker and listener is easy and accurate even if the two persons are not allowed to see each other, however, a noisy environment always reduces the listener's ability to understand first what is said, and then who is speaking. When extending beyond interpersonal communication, speech can be transmitted across telephone channels, loudspeakers, or headphones, etc, which largely profits from the sophisticated modern communication network.

Inevitably in transmissions are some forms of data conversion, i.e., quantization, compression, amplification, etc. All these detrimentally affect the quality of speech signals. The quality of a speech signal can be perceptually judged by human-beings as in interpersonal dialogues. Besides, in many recognition-related applications where they are responsible for providing accurate phonetic or speaker-specific clues, all are determined on recognition performance.

Generally speaking, the purpose of enhancement algorithms is to reduce background noise, improve speech quality, or suppress channel or speaker interference. Other applications include suppression of distortion from voice coding algorithms, suppression of a competing speaker in a multi-speaker setting, enhancing speech as a result of a deficient speech production system, e.g., speakers with pathology or divers breathing helium-oxygen mixture, or enhancing speech for hearing-impaired listeners. It is seen that the possible applications of speech enhancement are really broad. The success of an enhancement algorithm depends on the goals and assumptions made in deriving the approach. Determined by the specific application, a system may be directed at one or more objectives, such as improving overall quality, increasing intelligibility, or reducing listener fatigue. Speaker recognition system operating in the presence of background noise mostly take speech enhancement algorithms for robust feature extraction by including them into the signal processing front-end. The enhancement sector is expected to remove the more the better noise-dominant components, while retain as much as possible speaker-specific speech properties in the utterances. Our selection for a speech enhancement algorithm is determined by the discrimination power reflected by the denoised features in speaker recognition applications.

Many approaches have been taken for robust speaker recognition purposes in the past years, each attempting to capitalize on specific characteristics or constraints, all with varying degrees of success. There are a number of ways in which speech enhancement systems can be classified. A broad grouping is concerned with the manner in which the speech is modeled. They can also be partitioned depending on whether a single-channel or dual-channel (or multi-

channel) approach is used. For single-channel applications, only a single microphone is available. Characterization of noise statistics must be performed during periods of silence between utterances, requiring a stationary assumption of the background noise. Usually only a single channel is available in situations such as telephone or radio communications. In dual-channel algorithms, the acoustic sound waves arrive at each sensor at slightly different times, because one is usually a delayed version of the other. In our discussion of speech enhancement algorithms, we shall concentrate on methods that assume

- noise distortion is additive;
- noise and speech signals are uncorrelated; and
- only one input channel is available.

Through the many years, linear and nonlinear compensation techniques have been proposed, with applications to feature, model and match-score domains, respectively. Some of the techniques were first developed in speech recognition research. Automatic speech and speaker recognition systems under noisy conditions were found more reliable if a speech enhancement scheme incorporated as a preprocessing stage. Examples of the feature compensation methods include the well-known spectral-subtractive type algorithms [76], [77], [78], [79], [80], [81], Wiener and Kalman filtering [78], [82], [83], [84], [85]. Boll initiated the spectral subtraction method in late 1970's [76], and pointed out its usage in speech and speaker recognition applications. In the following three decades, there are continuous efforts pushing forward and refining this method by overcoming its shortcomings and combining with other skills, to list a few representatives, for instance, spectral subtraction using oversubtraction was proposed by Berouti *et al.* [77], Mcaulay *et al.* in 1980 proposed a soft-decision noise suppression filter [78], nonlinear spectral subtraction [79], multi-band spectral subtraction [86], masking property of human auditory systems was taken into accounts by Virag in [80], MMSE-based algorithm [81], etc. Spectral-subtractive type algorithms, among others, are most widely applied for speech and speaker recognition purposes, especially for denoising spectral-based parameters [87]. This type of

algorithms generally take phase information less important for speech quality, thus only estimate the spectral magnitude quantity. It is therefore considered proper way to estimate the magnitude-based spectral features. Methods falling into this category usually share a similar implementation mechanism as shown in Figure 3.3. This sort of frequency-domain estimators measure the clean speech spectral magnitude for (1) spectral feature extraction front-end of automatic speech/speaker recognition, and (2) speech synthesis protocol where the noisy phase components are recombined with a standard overlap-add procedure.

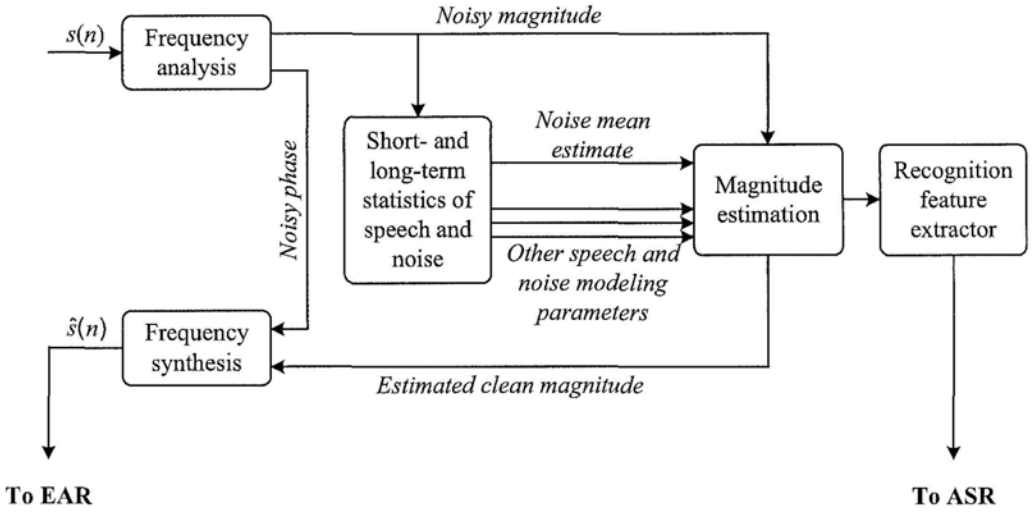


Figure 3.3: *Spectral magnitude estimator: A glance.*

Wiener filters are employed under the assumption that the signals analyzed are stationary. It generally assume a model of the clean spectrum and attempted to estimate the parameters of the model. The Wiener filters are considered to be linear estimators of the clean signal spectrum, and they are optimal in the mean-square sense. The Wiener filters can also be extended to handle nonstationary signals and noise with the use of Kalman filters. Kalman filters can be viewed as sequential mean-square estimators of a signal embedded in noise. Several speech enhancement methods based on Kalman filtering were proposed and made contribution in improving the robustness of speech recognition and

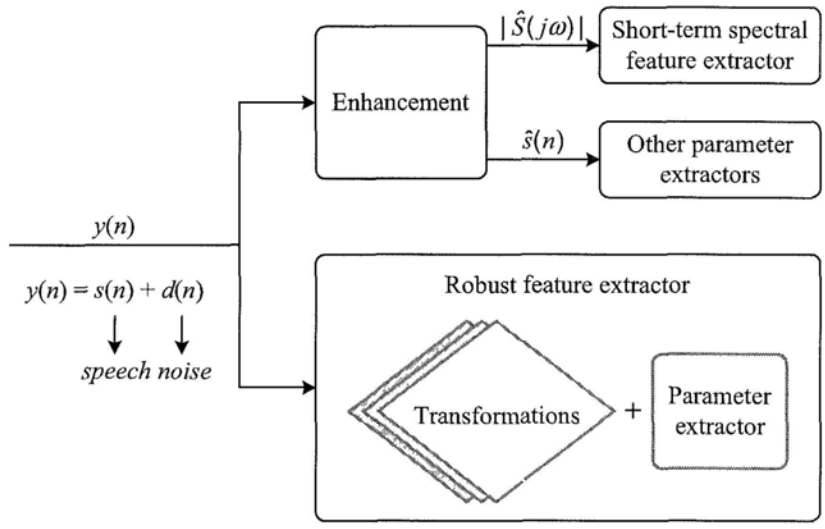


Figure 3.4: *Two approaches of robust feature extraction.*

speaker tracking systems, e.g., [84], [88].

As an essential component of robust speech processing, many works are devoted to estimating noise in various environmental conditions. The primary noise estimation algorithms include subband-based recursive method [89], histogram-based method [89], quantile-based method [90], and minimal-tracking algorithms [91], [92]. These noise estimators are in general applied prior to the feature extraction as a preprocessing procedure, and they are reported have reduced the error rates of various recognition tasks together with efficient enhancement method.

3.2.2 Discriminative feature design, transformation and normalization

Figure 3.4 shows two schemes taken for robust feature extraction. Other than the first strategy of feature enhancement which was introduced in Section 3.2.1, discussion in this part centers on another handling approach, that is, to transform the noisy speech signals into some domains, usually not the conventional Fourier frequency domain, and then to dig out those noise-free but speaker-specific properties for representing the specific speaker's identity. In this second

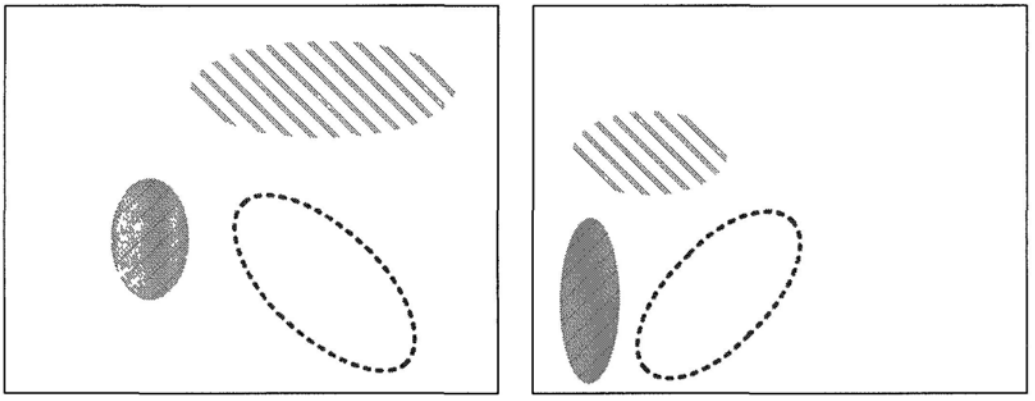


Figure 3.5: *2D pictorial illustration of noise/channel effects on feature vector space - scaling, rotation, and translation: (a) the spatial distribution of the clean vectors; (b) the spatial distribution of vectors of noise and/or contaminated speech.*

scheme, the involved transfromation/compensation methods either claim their ability of mitigating noise's effects in a way or can identify the speaker-related components easily. Successful examples in this category actually have long been applied in removing the effects caused by transmission channels, for example, the cepstral mean subtraction (CMS) [9], [93], [94], [95], cepstral attributes normalization [96], [97], [98], and relative spectral (RASTA) [95], [99].

When the speech signal is contaminated by a distorting channel or noise from the environment, the feature vectors are found to be rescaled, rotated, and translated. This can be seen in Figure 3.5.

In practice, CMS and other techniques previously mentioned have been found to offer enhanced robustness of speaker-recognition systems. Referring to notations used in Section 2.3.2, we note individual cepstral coefficients by c_i. After doing CMS to normalize the channel, and performing cepstral liftering to reduce the noise effect, the corrected cepstral coefficients are then denoted by $\hat{c}_i$. It can be represented as

$$\hat{c}_i = w_i c_i - \bar{c}_i. \tag{3.1}$$

The generalization of the relationship can be written as an affine transform

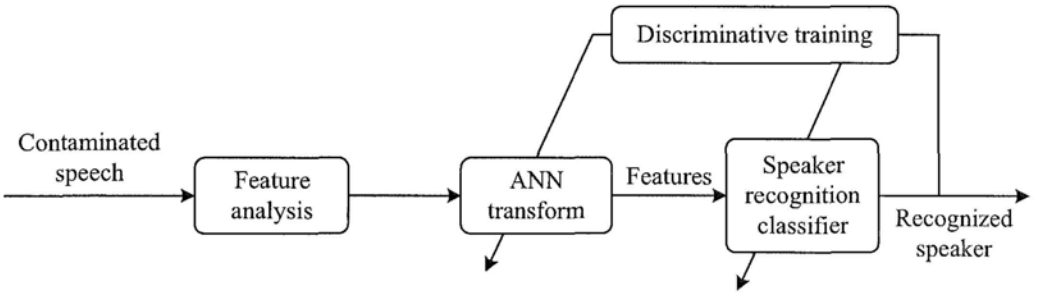


Figure 3.6: *Block diagram of discriminative feature and classifier design approach. The speech signal is corrupted by a number of environmental factors, which the approach attempts to compensate for by adapting the artificial neural network (ANN) feature transform and speaker recognition classifier based on an estimate of speaker recognition performance.*

given by

$$\boldsymbol{c}' = \boldsymbol{A}\boldsymbol{c} + \boldsymbol{b}, \tag{3.2}$$

where $\boldsymbol{c}'$ is the cepstrum of the degraded speech and $\boldsymbol{c}$ is the cepstrum of the original clean speech. This becomes a similarity transform when the matrix $\boldsymbol{A}$ is diagonal and the vector $\boldsymbol{b}$ is zero. It implies that a degraded, spectrally similar set of cepstral vectors would undergo the same transformation. The concept of using an affine transform to correct the distortions of the cepstral coefficients caused by the channel and noise interferences has been proposed in [93], [100]. Furthermore, nonlinear spectral magnitude normalization [101] is another representative among various feature transformation methods. From another point of view, transforms that are capable of decomposing speech signal with multiple time-frequency resolution also claim denoising capability, e.g., wavelet transform [43], [102], [103], they are sometimes combined with the parametrization schemes mentioned in Section 2.4 of Chapter 2 for alternative acoustic parameters.

With the advent of discriminative training techniques, model learning has become a task of maximizing class separability rather than a likelihood func-

tion. However, this progress is limited by the type of features used in the recognition design. Although cepstral-based features are widely used in the field, their design criterion is *not* consistent with the objective of maximizing speaker recognition rates. Integrated feature and model design under a single training objective clearly provides an additional benefit over conventional systems and remains a challenging problem in speaker recognition research. Integrated feature and model design through discriminative training has been the subject of several recent studies [104], [105]. Block diagram of discriminative feature and classifier design approach mentioned in [105] is shown in Figure 3.6 as a model for this type of methods. As it is shown, the feature extractor contains two parts: an initial feature analysis and a nonlinear feature transformation. The feature analysis is used to convert the speech signal into a collection of feature vectors. These features are then processed by the nonlinear feature transformation before being passed on to the speaker recognition classifier. The feature transformation is implemented as an artificial neural network. During the feature design phase, the speaker recognition classifier is also implemented as a neural network. Like the feature transformation sector, the classifier is trained to reduce the effects of nonlinear handset distortions on speaker discrimination. However, after the feature design phase, other classifier types can be used to carry out the speaker recognition task.

Other methods derived to normalize the feature vectors include feature warping [95], [106], short-time Guassianization [107], etc.

Compensating the features rather than the models has the advantage that the transformed parameters can be used with models of different nature and complexity, and also for different tasks.

3.3 Matching-score Normalization

It has been widely observed in the literature that handset variability that brought by different microphone handsets, causes significant performance degradation in speaker recognition systems. Channel compensation in the front-end processing addresses linear channel effects, but there is evidence that handset transducer effects are nonlinear in nature and are thus difficult to remove from the features prior to training and recognition. Because the handset effects remain in the features, the speaker's model will represent the speaker's acoustic characteristics coupled with the distortions caused by the handset from which the training speech was collected. The effect is that log-likelihood ratio scores produced from different speaker models can have handset-dependent biases and scales. This is especially problematic when trying to use speaker-independent thresholds in a system, as for most of the state of the art systems. A handset compensation technique H-norm in the score domain which normalize the distribution of the scores by removing the mean and scaling by the standard deviation was proposed [108].

Other prevalent score domain normalizations include Z-norm [109] and T-norm [110], which are concentrated on removing the biases or effects caused by speaker, linguistic content, etc. These two normalizations are sometimes used simultaneously, where the Z-norm is used to characterize the response of each speaker model to a variety of (impostor) test segments, followed by T-norm to compensate for the variations of the testing segments, such as duration and linguistic content.

Recently, a new compensation method use Mento-Carlo method to normalize matching score was described as D-norm [111]. The D-norm is based on the use of Kullback-Leibler (KL) distances in an speaker verification context. This new method has revealed itself with comparable effectiveness with the Z-norm, but excels at requiring no additional speech data nor external speaker population.

3.4 Speaker Model Compensation

In recent years, the efforts towards eliminating the mismatches in environmental noise, channel characteristics, as well as those induced by the speaker himself/herself have emerged from either enhancing acoustic features or normalizing the match-score.

For robust speech recognition tasks that under noisy conditions, the parallel model combination (PMC) method has been proposed. With the assumption of available statistics of environment, it targets to mitigate the mismatch between training and test conditions by first building an HMM model for the undesirable environmental factors, and then combining it with the original clean speech model for an environmental matched one in the testing. It was raised in the field of speech recognition [112], and was employed on text-dependent speaker recognition tasks [113] with some success. Similar techniques focus on noise compensation include Jacobian environmental adaptation [114].

Besides, there was a new perspective that tends to show out all these undesirable effects together with other known or unknown factors by a single term *session variability* [115]. New algorithms arising therefrom attempted to directly model session variability in the model space without discrete categories and with less restrictive data labeling requirements. It is proposed to incorporate session differences into the way a speaker is modeled within a speaker recognition system, in both the training and test phases of the system. In [115], the presented approach does not perform speaker adaptation in a subspace adopting a more traditional GMM-UBM structure and obviating the need to train a speaker subspace transform and significantly reducing training complexity. The necessity to constrain session variability modeling to a low-dimensional space is also emphasized. Prior to this work, there were a few model-based techniques, e.g., speaker model synthesis (SMS) [116], feature mapping [117], etc. Kenny *et al.* proposed a joint model of inter-speaker and inter-session variability lately [2]. Eigenchannel MAP was proposed in [118] as a model of inter-session variability, and in [119] eigenvoice MAP was taken as model of inter-speaker variability. The proposed analysis model was also applied on GMM [2]. Channel adaptation

and inter-speaker variability were studied in [120], [121] and [122], respectively. Concerning the cross-channel degradation, recently, a method named Nuisance Attribute Projection (NAP) was proposed [123], [124].

3.5 Summary

This chapter reviews the literature on robust speaker recognition fields spanning from the classical algorithms to the most up to date techniques. The methods surveyed are grouped in three categories, which cover the main processing stages of a speaker recognition system, i.e., the speaker-specific feature extraction/enhancement, the discriminative model training, and the matching-score normalization. Methods that pioneered and of crucial importance in each areas are first looked into with greater details, with their numerous followers cited thereafter. From another point of view, we focus on compensation methods that are essentially dealing with nonspeaker factors, such as the session variability caused by background noise, channel mismatch, handset variability, etc.

Chapter 4

Characterization of Individual Speakers with Distinctive Vocal Excitation Features

Although reckoned as one dimensional quantity, speech processing touches upon a number of areas, physiological, psychological, psychoacoustic, auditory, cognitive, linguistic, etc. A general model for speech signal can be extremely difficult if perfect solution is expected in all aspects. A few speech models took shape throughout the years, each leaning to a specific application area. As one application among others, speaker characterization for the sake of discrimination, involves a few signal processing techniques, e.g., Fourier spectral analysis, linear prediction modeling, multi-band signal decomposition, etc. These analysis models each offer useful speaker-specific parameters for distinguishing different speakers.

Vocal excitation contributes in personalizing a speaker's voice. The continuous vibration of the vocal folds produces voicing source for speech production, and personal patterns for characterizing the speaker as well. The quasi-periodic motion of relevant vocal organs brings periodicity to the excitation signal, while the pulse-like epoch shapes vary among people as well. Classical voicing models either fit the waveform point-by-point for an accurate approximation, or transform it into other domains, e.g., Fourier spectral domain, for picking out the

primary harmonic structure. The pitch harmonics of interest mostly locate in the lower frequency region of a voice signal, while the higher frequency part appears to be noise-like, making the included personal attributes less concerned. If view the problem at another angle, all components in the voicing signal are produced by the set of vibration activities, they as a whole should be accountable for delivering the inherent properties of a speaker. For the study of interior speaker-specific characteristics, a voicing model that considers systematically most primary elements is essential.

In this chapter, we make a study on characterizing a speaker's vocal excitation pattern through modeling the corresponding voicing signals. In Sections 4.1 and 4.2, different types of voice models are introduced. Section 4.3 discusses issues related to the modeling of source signals with AM, FM parameters. The excitation related modulation properties are studied with the help of multi-band demodulation method, and source-related amplitude and phase quantities are parameterized into feature vectors in Section 4.4. Evaluation of the proposed features is carried out first through a set of designed experiments on artificially generated inputs, and then by simulations on speech database in Section 4.5. Section 4.6 summarizes the work in this chapter.

4.1 Excitation Waveform Modeling

Analyzing speech signals by separating the vocal source- and vocal tract-related properties is a frequently used method. As we have mentioned in Chapter 2 that source-related parameters can characterize a person's vocal attributes, where the time domain parameters usually care about the glottal pulse-related properties, and the spectral characteristics concern the harmonic structure and F0-dependent features. Likewise, the establishment of source excitation models has been considered as one of the important issues in synthesizing natural sounding voices. Features that might be useful in designing a source excitation model for synthesizing natural sounding speech and identifying vocal disorders are selected and summarized in [125]. It is thought that if based upon an established voicing source model, with all the controlling parameters known, one can alter the vocal properties easily by tuning those parameters. In this process, the usefulness of the parameters in different applications can be evaluated according to relevant metrics. For instance, in characterizing a speaker's vocal attributes, where the ultimate objective is speaker identifiability, the modeling parameters that are able to affect speaker individuality are viewed predominant.

In this study, we investigate the aspects of voicing source modeling that related to vocal fold vibratory patterns. Our purpose is neither to develop new voice source models for synthesizing natural sounding speech, nor to perceptually study the vocal source-related characteristics. We target to exploit speaker-distinctive features from a pertinent voicing model, that is, parameterize the physically sounding model parameters into speaker representatives. The retaining of physical meaning for the features and their potential in reconstructing the voice source claim novelty for this method. Study on speaker-discriminative source feature derivation in this part considers following three steps: (1) select proper source excitation model to use; (2) parameterize model parameters into feature vectors; and (3) analyze whether the feature vector could identify a speaker's vocal characteristics through observing their expressiveness for a set of typical vocal properties.

4.1.1 Voicing source model

Quite a few source excitation models have been designed to function in synthesizing natural sounding voices, which include but are not restricted to the following: Fant's [40], [126], Klatt's [127], Childers's [128], etc. Considering we need to model the excitation signals first, and then measure the model parameters' expressiveness for relevant vocal properties owned by the signals, a model mathematically formulated in the time-frequency domain will be desirable.

Reviews in Chapter 2 deals with the subject of speech production, where it tells us that the quasi-periodic vibration of the vocal folds first produces excitation for a voice speech, and when passing it through the vocal tract. The excitation resonates with the vocal tract at several frequencies, thus alters the magnitude of speech components over frequency span of the sound. The speech signal is therefore known as an oscillatory signal, where the vocal tract is viewed as an oscillating system. Modulation theory is applied to speech signal successfully in dividing and demodulating these resonances. Meanwhile, properties of the excitation signal is looked into via observing either its periodic pattern shown in the waveform or the harmonic structure exhibited by its spectrum. Pitch periodicity, pitch epoch shape, and certain details embedded are the principal temporal patterns considered. Additionally, in the frequency domain, people concern the primary F0 and its harmonics. These are important and essential characteristics of an excitation signal indeed, however, regarding the generation of excitation signal, our views are as follows: first, there must be some time-frequency patterns exist in the signal; second, these patterns may be the primary components of the signal; third, these components will be the elements of an appropriate excitation model.

Significant observations was made on the amplitude and frequency modulations present in speech formants through speech signal analysis. If these modulations are removed, noticeable deterioration on speech quality and the specific speaker's vocal properties will occur. This phenomenon was reported in many research work on speech coding and synthesis [129]. It reveals the perceptual importance of modulations in speech formants, and motivates the introduction

of relevant modulation models into speech decomposition, especially for voiced sounds. To list a few of the associated models, there are sinusoidal speech model by Mcaulay *et al.* [130], harmonic plus noise model (HNM) by Stylianou [131], AM-FM modulation model in a series of works by Maragos, Kaiser, Quatieri, Potamianos *et al.* where [129] and [132] are the primary ones, and the recently proposed Quasi-harmonic model (QHM) [133], etc. Strictly speaking, AM-FM model and QHM model are both derived from the sinusoidal model.

4.1.2 Sinusoidal modeling of excitation signal

Sinusoidal excitation model represents the excitation signal waveform as a sum of sine waves. Ever since the classical source-filter speech production model, where the voiced excitation is generated as a periodic pulse train, the harmonicity of excitation signals are broadly taken in speech processing. It is therefore motivated to formulate the voiced excitation by means of Fourier series decomposition in which each harmonic component corresponds to a single sinusoid.

In sinusoidal model, the excitation waveform $e(n)$ of a segment of voiced speech signal $s(n)$ is represented as a sum of harmonically related complex exponentials, where these components assume unit amplitudes and zero initial phase. The model can be expressed mathematically as follows,

$$e(n) = \sum_{k=1}^{K(n)} e^{\jmath\Theta_k(n)}, \tag{4.1}$$

where $\Theta_k(n)$ is the instantaneous phase of the kth harmonic component in this excitation, and $K(n)$ is the number of harmonics present at time instant n. Both parameters vary with time.

Under the assumption that the excitation waveform is purely composed of harmonics, the instantaneous frequencies, $f_k(n)$, are interrelated in the following manner,

$$f_k(n) = kf_0(n), \tag{4.2}$$

where $f_0(n)$ is the estimate of fundamental frequency which also varies with time.

In applications like speech analysis, synthesis, transformation, etc, a complete speech model is sometimes founded on this representation of excitation signal. An additional time-varying linear filter is usually introduced to model the combined effect of the transmission characteristics of the vocal/nasal cavity, as well as the radiation at the mouth opening and also the glottal pulse shape.

4.1.3 Harmonic plus noise model for speech

Harmonic plus Noise Model (HNM) has attracted much attention after the pioneering works by Griffin *et al.* [134] and Abrantes *et al.* [135]. In HNM, it is assumed that a speech signal $s(n)$ is composed of a harmonic part $h(n)$ and a noise part $n(n)$. There exists a time-varying frequency delimiter $F_m(n)$ in the spectrum dividing the a voiced speech signal into two bands, where the lower band of the spectrum refers to the harmonic part, while the upper band represents the noise part. The harmonic part, $h(n)$, which accounts for the periodic structure of the voiced speech signal in HNM, is formulated by a sum of harmonically related sinusoidal components with discrete time-varying amplitude and phase quantities as that shown in Equation 4.3,

$$h(n) = \sum_{k=1}^{K(n)} A_k(n) cos\big[\Theta_k(n)\big], \tag{4.3}$$

where $A_k(n)$ and $\Theta_k(n)$ are the amplitude and phase quantities of the kth harmonic component, respectively. $K(n)$ is the number of harmonics present in the harmonic part at time instant n. Similar with the sinusoidal model, the instantaneous frequency in the HNM, i.e., $f_k(n)$, also equals to k times of the fundamental frequency $f_0(n)$.

In addition to the unvoiced frame which requires a noise-like formulation, Stylianou in [131] accounted the friction noise and the period-to-period fluctuations produced by the turbulence of the glottal airflow as the causes for the noise part $n(n)$ of a voiced sound. He meanwhile pointed out that, the noise part in

a voiced signal usually behaves like a high-pass signal which exhibits certain energy distribution of the related frequency content, as well as a time-domain structure of the signal.

Thus, to describe the noise part $n(n)$ in speech $s(n)$, a time-varying AR envelope is used to shape the frequency content, while its time-domain structure is determined by a piecewise linear energy-envelope function. In Equation 4.4, $n(n)$ is produced by first filtering a white Gaussian noise $u(n)$ by a time-varying, normalized all-pole filter $h(n,\tau)$, and then shaping the result by an energy envelope function $e(n)$,

$$n(n) = e(n)\big[h(n,\tau) \otimes u(n)\big]. \tag{4.4}$$

Finally, the speech signal in the form of HNM is expressed as follows,

$$s(n) = h(n) + n(n). \tag{4.5}$$

4.1.4 Parameters estimation

In estimating the parameters of sinusoidal excitation model, the stationary hypothesis applies within a frame. For the case that the frequencies are strictly harmonically-related, i.e., $f_k(n) = kf_0(n)$, a high-performance pitch detector is necessary. The amplitude and phase values can be obtained by linear interpolation along time within a specific frame [136]. Provided accurate fundamental frequency estimation, this approach leads to good frequency estimates, thus the mean square error between the original signal and the sum of harmonics can be small.

In conditions where the sinusoids are not multiples of the fundamental frequency, a peak-picking algorithm [130] is usually taken to estimate the frequency of each underlying sine wave. This algorithm is operated in the frequency domain, where the peaks in the periodogram are first located, and then the amplitudes and phases at the frequencies of these peaks are obtained by evaluating the short-time Fourier transform (STFT). The STFT-based method prefers long

frames for sufficient spectral resolution, this limits the use of this class of methods when the pitch-period, amplitude and phase quantities vary rapidly.

4.2 AM-FM Representation for Excitation Signal

In this section, the rationales and details of representing vocal excitation signals of voiced speech sound with AM-FM model are introduced. Various examples are used in delivering this idea.

4.2.1 Fundamentals of modulation

Strictly speaking, a band-limited signal does not exist in reality, however, most of the signals closely approximate the band-limited signals. Natural signals, since most of their energy is carried by components lying within a certain frequency interval, are usually considered to be band-limited for practical reason.

For a band-limited signal $f(n)$, in the frequency domain where $f(n) \leftrightarrow F(e^{j2\pi f})$, we have

$$F(e^{j2\pi f}) = 0, \quad |f| > f_m. \tag{4.6}$$

where f_m is the maximal frequency of $f(n)$.

A narrow-band signal, whose bandwidth is sufficiently small, can be viewed as a *monocomponent* amplitude and frequency modulating (AM-FM) signal, among the frequencies spanning over the signal spectrum, there is one frequency assuming a majority of the signal energy. Specifically, this frequency component in the time domain corresponds to an exponential signal. The two determining parameters in an exponential signal is the amplitude and frequency. In a mechanical system, where the vibration activity is usually described by a sinusoid, the extent of the activity is decided by the amplitude, while the rate of vibration is indicated by the instantaneous frequency quantity. In a communication system, the amplitude modulation (AM) essentially transmits the information signal by multiplying it with a sinusoidal signal at the carrier frequency. When the carrier frequency varies, its variation can be used to transmit another message, this is called frequency modulation (FM). The frequency modulation refers to narrow-band FM in this thesis.

4.2.2 Description of AM-FM modeling

It is known that in producing a voiced sound, the vibration of vocal folds is essential. The successive vibrations cause the generation of a quasi-periodic signal, which excites the vocal tract system to shape the frequency elements contained in the excitation signal. The whole process is articulation. An analogy between the vocal folds' oscillation in phonation and a spring's movement in mechanical activities indicates some likeness in their behaviors. The voice range of a person is largely determined by the structure of his/her vocal organs, while in mechanics, a spring's scope of action is restricted within a bound that depends on its material, texture and other specifications.

The existence of inter-person distinction in phonation- and articulation-related organs forms the primary and crucial evidences for speaker recognition research, which is by nature a biometric pattern matching task. Individual speakers are discerned by the joint work of vocal folds as well as other vocal organs, like vocal/nasal tract. The excitation signal which is yielded from vocal folds' oscillation before the resonation by vocal tract, conveys primary speaker-specific traits about the oscillations occurring in the vocal folds.

Simple harmonic motion: an example

We can get a thorough understanding of the fundamentals of vibration theory by studying the simple mass spring damper model, which is an example of simple harmonic oscillator. In fact, a mechanical structure even complex can still be modeled by summing a set of simple mass spring damper representations. Similar oscillators can be found in many places, for example, in an RLC circuit.

The second-order differential equation in Equation 4.7 describes the motion of a mass m suspended by a spring of force constant k, as shown in Figure 4.1.

$$\ddot{x} + \frac{k}{m}x = 0. \tag{4.7}$$

where $\ddot{x} = d^2x/dt^2$.

The displacement $x(t)$ produced by the mass-spring is an undamped linear oscillating signal formulated as follows,

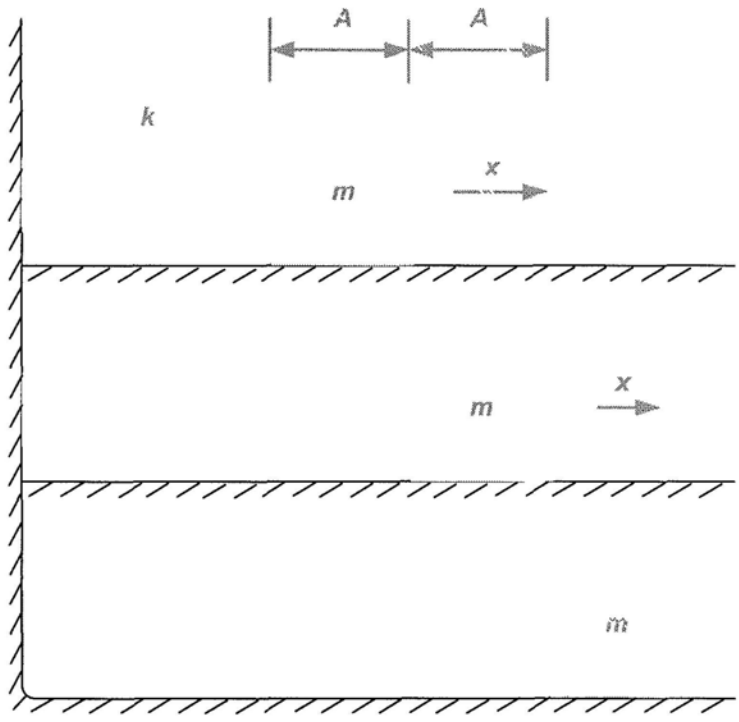


Figure 4.1: *Simple harmonic motion of a mass spring oscillator.*

$$x(t) = Acos(\omega t + \phi), \tag{4.8}$$

where the parameters A and ω are the amplitude and frequency of the oscillation, respectively. ϕ is the arbitrary initial phase.

The total energy of this vibration consists of two parts, i.e., the kinetic and potential energy, they together equal to $(m\dot{x}^2 + kx^2)/2 = (m/2)A^2\omega^2$. It is therefore found to be proportional to the squared product of amplitude and frequency.

Representing excitation signal in AM-FM model

Conventional studies on extracting vocal source-related characteristics for speaker recognition purposes, as we have reviewed in Section 2.4 of Chapter 2, were in a degree connected with either pitch-periodicity, or the relevant harmonic structure properties, but consider less about the inter-correlation among the multicomponents existing in an excitation signal. Excitation waveform mod-

eling approaches, on the other hand, consider more about the components contained in an excitation signal, as described earlier in this section. A common approach of getting these inclusive components are through nonlinear signal decomposition. After some kind of modifications or other processing steps, these components together can re-synthesize the excitation signal under the framework of, for example, the sinusoidal modeling. This is popular in various applications like speech modification and speaker transformation.

A *monocomponent* AM-FM signal is described by Equation 4.9 [132],

$$x(n) = A(n)cos\big[\Theta(n)\big], \tag{4.9}$$

where $A(n)$ denotes the instantaneous amplitude of the mono-component signal and $\Theta(n)$ denotes its instantaneous phase.

In practice, *multicomponent* AM-FM signals are present everywhere in natural sounds, including human speech. As we have mentioned about the pitch-periodicity property of the vocal excitation signal for a voiced speech, it is proper to interpret it as a multi-component AM-FM signal. The advantage of this approach lies in: first, it is a complete description of the signal rather than only noting some parameters, like the pitch-period or F0; second, it is convenient to represent the inclusive components in terms of their time-varying amplitude and frequency quantities, rather than fitting the waveform or spectrum; third, it is easy to investigate the individual components for capturing useful information. It is known that in many applications, there is great interest in analyzing the input signal by decomposing them into time-varying amplitude and frequency components. Depending on the fields applied, the estimation of amplitude-frequency modulation parameters plays essential but different roles. For instance, in speech coding, the accurate recovering of the speech signal is the ultimate goal; while in voice transformation, either in altering the speech properties or changing the speaker identity, before re-synthesizing the speech signals through the inverse model, the voices are converted by tuning their amplitude and frequency parameters.

4.2.3 Estimation of modulation parameters

Before introducing how to estimate the AM-FM model for the excitation signals, let us get familiar with its physical meaning by making analogy with mechanic vibration. In the initial work by Kaiser [137], an alternative "energy" calculation method, other than in traditional signal processing literature where the energy of a signal refers to the mean of magnitude squares of the exponentials present in the signal, was mentioned. With the traditional method, two acoustic signals with the same amplitude but different frequencies, say, $2cos(10\pi n)$ and $2cos(1000\pi n)$, assume equal energy. However, it was revealed in mechanical experiments that the energy required to generate the acoustic signal $2cos(1000\pi n)$ is much greater than that for the other signal which is of a lower frequency.

In the case of harmonic motion, as in the production of an excitation for a voiced speech segment, the vocal folds vibrate to produce a fundamental sinusoidal oscillation. As revealed earlier through the example in Section 4.2.2, the energy required to generate an oscillating signal is given by the square of the product of the signal's amplitude and frequency [137]. This type of energy calculation method is very useful in analyzing single component signals, like linear oscillators. In a quite complicated process, given the activity function can be expressed as products of simpler functions, its energy function can also be determined in this way. This is the case for most classes of signal in the world, including speech signals. Modulation-type process which is usually found in speech production, underwater acoustic signal generation and many other places typically belongs to this class of process. Therefore, this method is applicable to the quantitative analysis of energy distribution among different amplitude-frequency components that indwell in the production process of voiced speech.

AM-FM structure of oscillatory signals were taken used extensively in communication systems for transmitting information. Teagers through a series of pioneering works in, e.g., [138], [139], initiated the research on exploring relevant amplitude-frequency modulation patterns in speech resonances. Other researchers in this area, like Maragos *et al.* [132], also made effort to induce this method an appropriate tool for speech analysis. Owing to evidences for

the existence of modulation phenomena in speech production process, Teager's energy operator is found to be a useful tool for analyzing and estimating the characteristics of the existing amplitude and frequency modulation patterns in a vocal excitation signal

Teager's energy separation algorithm

♦ Algorithm description

Teager's energy separation algorithm takes use of a nonlinear differential operator to detect modulations in AM-FM signals Through the example AM-FM signal defined in Equation 4 9, we look into and identify the information transmitted by the AM and FM part of the signal as in Equation 4 10,

$$\begin{aligned} x(n) &= A(n)cos\big[\Theta(n)\big] \\ &= A(n)cos\left[\Omega_c n + \Omega_m \sum_{r=1}^{n} q(r) + \phi\right], \end{aligned} \tag{4 10}$$

and we have the instantaneous phase defined by

$$\Theta(n) = \Omega_c n + \Omega_m \sum_{r=1}^{n} q(r) + \phi \tag{4 11}$$

The concerned AM-FM signal $x(n)$ is transmitted at the carrier frequency Ω_c with time-varying amplitude signal $A(n)$ and angle signal $\Theta(n)$ To identify the instantaneous frequency components contributed to $\Theta(n)$, we further obtain a time-varying instantaneous angular frequency (IF) signal $\Omega(n)$ by taking backward difference on $\Theta(n)$ between two consecutive instants as in Equation 4 12

$$\begin{aligned} \Omega(n) &= \Theta(n) - \Theta(n-1) \\ &= \Omega_c + \Omega_m q(n) \end{aligned} \tag{4 12}$$

It is clearly seen that at each specific moment n, there is a frequency deviation $\Omega_m q(n)$ from the carrier Ω_c assumed by the IF signal $\Omega(n)$ In communication systems, this quantity is used to carry the message when transmitting

through FM scheme. Since it is always true that $|q(n)| \leq 1$, the maximum frequency deviation actually depends on a constant Ω_m. Without loss of generality, Ω_m is normalized to be one in our analysis.

The Teager energy separation algorithm can efficiently estimate the amplitude and frequency modulating signals based on an "energy-tracking" operator, which is named Teager Energy Operator (TEO). The TEO takes nonlinear processing for a discrete-time signal as shown in Equation 4.13,

$$\Psi_d\big[x(n)\big] \triangleq \big[x^2(n) - x(n-1)x(n+1)\big]/T^2, \tag{4.13}$$

where the subscript d in the operator $\Psi_d[\cdot]$ implies the discrete-time domain. T is the sampling period, and in the remainder of this thesis, $T = 1$ is assumed, thus we can discard it from the expression of $\Psi_d[\cdot]$. This operator was first introduced systematically by Kaiser to track the energy of simple harmonic oscillators [137]. If apply a continuous version of this operator, i.e., $\Psi_c\big[x(t)\big] \triangleq \big[\dot{x}(t)\big]^2 - x(t)\ddot{x}(t)$, on the simple harmonic motion mentioned in Equation 4.8, we can get

$$\begin{aligned}
\Psi_c\big[Acos(\omega t + \phi)\big] &= \big[-A\omega sin(\omega t + \phi)\big]^2 - \big[Acos(\omega t + \phi)\big]\big[-A\omega^2 cos(\omega t + \phi)\big] \\
&= (A\omega)^2 sin^2(\omega t + \phi) + (A\omega)^2 cos^2(\omega t + \phi) \\
&= (A\omega)^2,
\end{aligned} \tag{4.14}$$

which is identical with the total energy of the oscillator we obtained over there.

For time-varying amplitude and frequency modulation signal, like that described by Equation 4.10, this operator is also revealed very useful, as shown below in Equation 4.15,

$$\Psi_d\Big[A(n)cos\big[\Theta(n)\big]\Big] = \Psi_d\Big[A(n)cos\Big(\Omega_c n + \Omega_m \sum_{r=1}^{n} q(r) + \phi\Big)\Big] \approx \big[A(n)\Omega(n)\big]^2. \tag{4.15}$$

The next goal in the parameter estimation process is to separate the instantaneous envelope $|A(n)|$ from the instantaneous frequency $\Omega(n)$ in the output of the TEO. This step of computation is called discrete energy separation algorithm (DESA).

For this bandlimited AM-FM signal, we have

$$\Psi_d\big[x(n)\big] = A^2(n)\Psi_d\Big[cos\big[\Theta(n)\big]\Big] + \Psi_d\big[A(n)\big]\Big\{cos^2\big[\Theta(n)\big] - \Psi_d\Big[cos\big[\Theta(n)\big]\Big]\Big\}. \tag{4.16}$$

The unsolved part in this equation is $\Psi_d\big(cos\big[\Theta(n)\big]\big)$. By reformulating the signal representation as $x(n) = A(n)cos\big(\sum_{r=1}^{n}\Omega(n) + \phi\big)$ first, and then go through some derivation steps omitted here, we get an approximate solution for it, that is,

$$\Psi_d\Big[cos\big[\Theta(n)\big]\Big] = \Psi_d\Big[cos\Big(\sum_{r=1}^{n}\Omega(n) + \phi\Big)\Big] \approx sin^2\big[\Omega(n)\big]. \tag{4.17}$$

Then, we have

$$\begin{aligned}\Psi_d\big[x(n)\big] &\approx A^2(n)sin^2\big[\Omega(n)\big] + \Psi_d\big[A(n)\big]\Big[cos^2\big[\Theta(n)\big] - sin^2\big[\Omega(n)\big]\Big] \\ &\approx A^2(n)sin^2\big[\Omega(n)\big],\end{aligned} \tag{4.18}$$

where there is an assumption about $A(n)$, that is,

$$\Psi_d\big[A(n)\big]_{max} \ll \Big\{\big[A(n)\big]_{max}\Big[sin\big[\Omega(n)\big]\Big]_{max}\Big\}^2. \tag{4.19}$$

Thereafter, we get

$$\Psi_d\big[y(n)\big] \approx 4A^2(n)sin^2\Big[\frac{\Omega(n-\frac{1}{2})}{2}\Big]sin^2\Big[\Omega(n-\frac{1}{2})\Big]. \tag{4.20}$$

By combining Equations 4.18 and 4.20, we eventually get the solutions for the instantaneous envelope $|A(n)|$ and angular frequency $\Omega(n)$ as follows,

$$|A(n)| \approx \sqrt{\frac{\Psi_d[x(n)]}{1-\Big(1-\frac{\Psi_d\big[x(n)-x(n-1)\big]}{2\Psi_d\big[x(n)\big]}\Big)^2}}, \tag{4.21}$$

and

$$\Omega(n) \approx arccos\Big(1-\frac{\Psi_d[x(n)-x(n-1)]}{2\Psi_d[x(n)]}\Big). \tag{4.22}$$

This algorithm applies to band-limited signals. It can estimate instantaneous frequencies up to 1/2 the sampling frequency, i.e., $0 < \Omega(n) < \pi$. For multicomponent AM-FM signals that encountered in practical applications, a band-pass filtering process is needed before demodulation.

♦ Example: signal demodulation

In order to have a direct-viewing experience on the separation algorithm described above, we would like to make some observations on the separation results through an example.

The signal for demonstration is a mono-component signal which has a standard AM-FM representation $x(n) = A(n)cos[\Theta(n)]$. Instantaneous frequency of the signal varies linearly to time, the amplitude function contains a sine wave and DC component, they are defined respectively as follows,

$$A(n) = 0.75 + 0.25cos(\pi n/50), \tag{4.23}$$

$$\Theta(n) = \begin{cases} 0.15\pi n + \pi(n-100)^2/4000, & n = 0,\ldots,200, \\ 0.20\pi n - \pi(n-200)^2/4000 + \pi/2, & n = 201,\ldots,400. \end{cases} \tag{4.24}$$

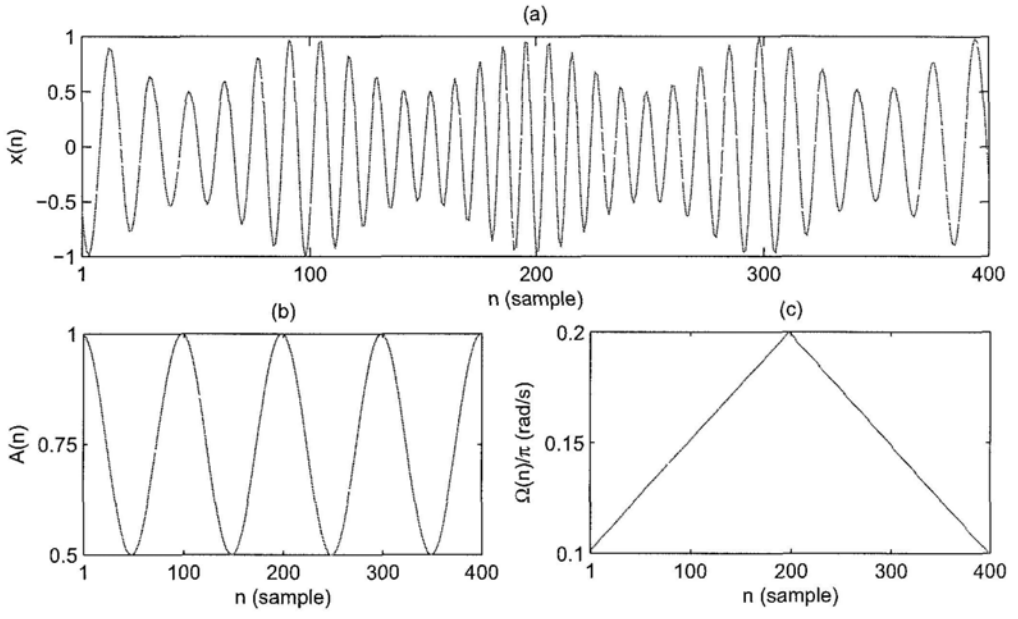


Figure 4.2: *An example AM-FM signal, its time-varying amplitude and frequency quantities: (a) signal $x(n)$, (b) amplitude $A(n)$, and (c) instantaneous frequency $\Omega(n)/\pi$.*

Figure 4.2 visualizes these three time-varying sequences, where the upper

layer shows the AM-FM signal $x(n)$, the lower two columns record the instantaneous amplitude $A(n)$ and angular frequency quantity $\Omega(n)/\pi$, respectively.

In the process of separating the amplitude and frequency components for signal $x(n)$, we put it to the Teager's energy operator in Equation 4.13, and then get the amplitude and frequency parameters by Equations 4.21 and 4.22, respectively. As a result, the operation output in these steps are illustrated in Figure 4.3.

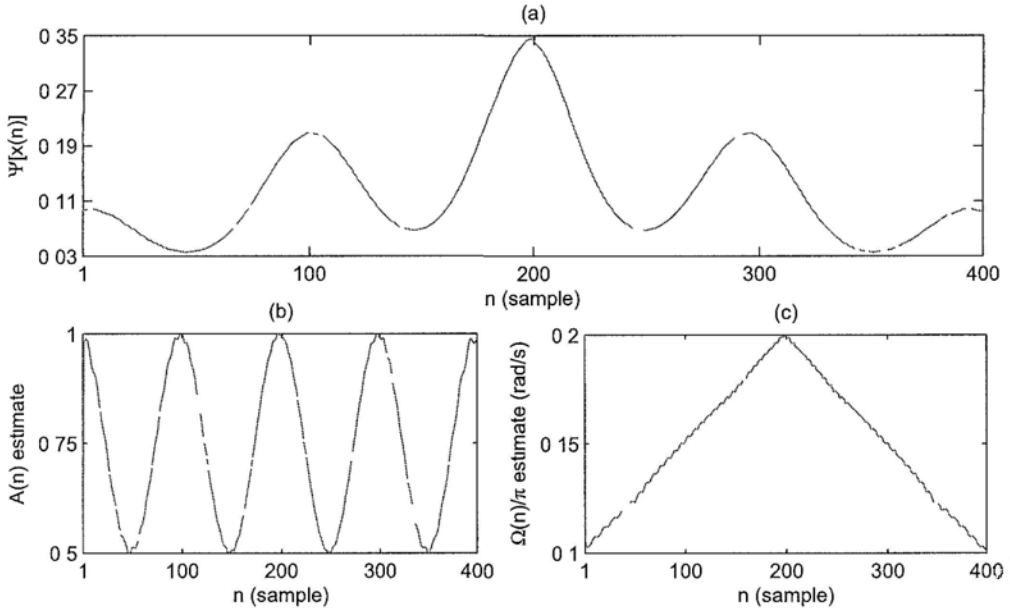


Figure 4.3: *Teager energy operator output for the AM-FM signal and the estimated amplitude, frequency parameters: (a) $\Psi_d[x(n)]$, (b) estimated amplitude, and (c) estimated instantaneous frequency.*

Figure 4.2 and Figure 4.3 display the courses of modulation and demodulation one after another. Results achieved by the DESA method can be revealed by comparing the corresponding parameters between them as well.

Other approach: Hilbert transform

Hilbert transform separation algorithm (HTSA) is another main tool applied to signal demodulation. Unlike DESA which uses a nonlinear differential operator,

the HTSA involves a linear integral transform in the computation. In detecting the envelope and frequency elements of a real-valued modulation signal, HTSA as a first step constitutes an analytical signal from it, and then get the corresponding modulus and phase derivative as the estimates of the amplitude and frequency part.

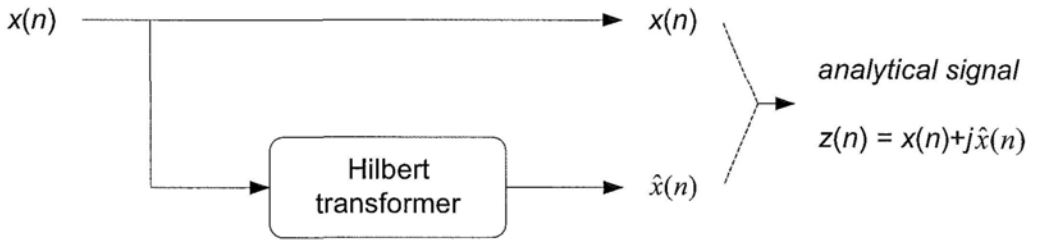


Figure 4.4: *Creation of an analytical signal from a real-valued signal.*

For a causal and real-valued AM-FM wave $x(n) = A(n)cos[\Theta(n)]$, its Hilbert transform is noted to be $\hat{x}(n)$, then an analytical signal $z(n)$ is derived from them by following the process shown in Figure 4.4. The analytical signal $z(n)$ can be alternatively represented by its modulus $r(n)$ and phase $\phi(n)$ as $z(n) = x(n) + j\hat{x}(n) = r(n)e^{j\phi(n)}$. These two variables can be worked out by Equations 4.25 and 4.26 as

$$r(n) = \sqrt{[x^2(n) + \hat{x}^2(n)]}, \tag{4.25}$$

$$\phi(n) = arctan\left[\frac{\hat{x}(n)}{x(n)}\right]. \tag{4.26}$$

The envelope and instantaneous frequency of the signal $x(n)$ in the HTSA are found closely related with these two sequences in the following manner, i.e., $|A(n)| \approx r(n)$ and $\Omega(n) = \phi(n) - \phi(n-1)$.

In [140], Potamianos and Maragos who proposed the energy-tracking based method, made a systematical comparison on the above two types of signal demodulation method, for general synthetic AM-FM signals as well as genuine speech resonances. The time-varying envelope and instantaneous frequency estimates were evaluated in terms of estimation error. Strong evidences were

provided to conclude that, the smoothed energy operator approach is comparable with the Hilbert transform approach for speech applications, in terms of estimation error, but the energy operator approach wins in computational complexity and faster adaptation due to its instantaneous nature. In addition, an advantageous feature is found for DESA: it produces the energy required to generate each mono-component AM-FM signal. This is favorable for us to investigate the vibrations yielded by the vocal folds for possible speaker-distinctive patterns.

4.3 Characterizing Voicing Source by Modulation Parameters

It is hard to exactly separate the real voicing source from a speech segment and to employ it in all sorts of applications. There are some representatives for vocal excitation signal have been raised. Preference to them may depend on the type of work to be conducted. As the interest of this thesis lies in achieving understanding and exploiting distinct features to distinguish different speakers, we are not concerned with many other respects that much. An expedient way for us is to take use of the LP residual signal to provide the speaker-specific source-related characteristics. Thus, the discussion of excitation signal in this section refers to the LP residual signal.

4.3.1 Effects of band-pass filtering and periodicity

In the previous section, an AM-FM signal $x(n) = A(n)cos[\Theta(n)] = A(n)cos[\Omega_c n + \sum_{r=1}^{n} q(r) + \phi]$ was used to explain the signal demodulation algorithm. It was defined as a signal of fairly narrow bandwidth in the frequency domain, thus can be viewed as a mono-component signal. In practice, a voicing source signal always contains a sum of such AM-FM signals, thus to track the envelope and the frequency modulation of it, we need to separate the signal prior to demodulation. In addition to the need of band-pass filtering, voiced source signals share another common characteristics, namely pitch periodicity. Effects caused by the presence of pitch periodicity in the excitation signal and that due to band-pass filtering in the course of processing are two of the main problems need to be tackled when estimating the instantaneous envelopes and frequencies in the source signals. Potamianos and Maragos in [140] have mentioned similar problems in estimating parameters for speech resonances as well.

Effects of band-pass filtering

It is shown by many that a voiced speech signal is composed of several resonances, every single resonance is extracted by a band-pass filter centered around

the formant frequency, and thought to be of AM-FM structure. Although no resonation occurs in the voice production process, the vibration of vocal folds yields quite a few mono-component signals that are of AM-FM type. This can be found from the very basic method of expression for an LP residual signal, i.e., the Fourier spectrum. If look into the spectrum from the lower to the higher frequency region, it is found that the signal contains a bundle of various frequency components, either the pitch harmonics, the transition from harmonically-related elements to the non-harmonic zone, or the noise-like region.

Before applying the demodulation algorithm to an excitation signal, we must have the individual AM-FM element picked out through band-pass filtering. Gamma-tone filters, which are generally used to model auditory filters, as mentioned in Chapter 2, fit our need well. Mathematically, a gamma-tone filter is a linear filter described by an impulse response as the product of a gamma distribution and sinusoidal tone, as that expressed in Equation 4.27.

$$g(t) = at^{(N-1)}e^{-2\pi bt}cos(2\pi f_c t + \phi), \tag{4.27}$$

where a is an arbitrary factor used to normalize the peak amplitude to unity, N is order of the filter, b is the bandwidth that determines the duration of the impulse response, f_c is the center frequency and ϕ is the phase of the sinusoidal tone. A fourth-order filter is found to fit best with a wide range of human masking data, thus is frequently employed in auditory models. Figure 4.5 illustrates the impulse response of a fourth-order Gamma-tone filter whose center frequency and bandwidth are 1000 Hz and 125 Hz, respectively, and with zero phase offset assumed.

Considering the filtering process during signal decomposition, its effects on the re-synthesized signal was discussed in [140]. The updated estimates of amplitude and instantaneous frequency, after taking into account the filter response parameters, are mentioned. Because our focus is on analyzing the inclusive signals and extracting useful cues from them for speaker discrimination purpose, we will not pore deeply on this matter at most of the time, expect in the dis-

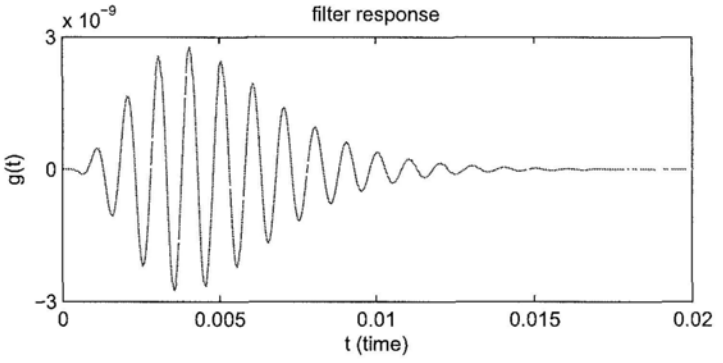


Figure 4.5: *Impulse response of a Gamma-tone filter.*

cussions of Section 4.3.3.

Effects of pitch

Let us start from observations made on synthetic signal. Given two composite sinusoidal signals where each composed of two sine waves as follows:

- $s_1(n) = cos(2\pi f_1 n/f_s) + cos(2\pi f_2 n/f_s)$
- $s_2(n) = cos(2\pi f_3 n/f_s) + cos(2\pi f_4 n/f_s)$

where $f_1 = 10Hz$, $f_2 = 20Hz$, $f_3 = 80Hz$, $f_4 = 90Hz$, and $f_s = 1000Hz$. Their waveforms and spectra are shown in the left and right column of Figure 4.6, respectively.

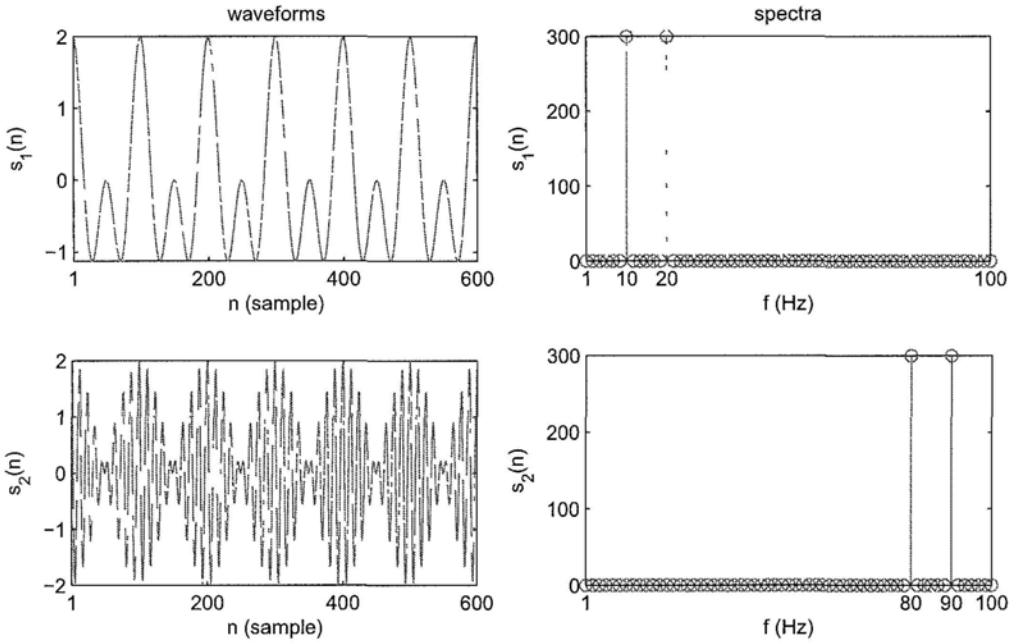


Figure 4.6: *Waveforms and spectra of the composite sinusoidal signals* $s_1(n)$ *and* $s_2(n)$.

Then, the estimated envelopes and instantaneous frequencies of the two signals are depicted in Figure 4.7, where the corresponding waveforms are shown in magenta for the purpose of reference as well. This figure gives evidence of the effects caused by the fundamental frequency of a periodic signal in the course of demodulation. It is found not easy to smoothly track the envelope and instantaneous frequency quantities for composite sinusoids where common fundamental frequency exists. This is actually the main problem encountered by the formant- or pitch-trackers which are based upon signal demodulation algorithms. Referring to the two signals $s_1(n)$ and $s_2(n)$ here, it is indicated that $s_1(n)$ is with lower frequency, and $s_2(n)$ is its higher frequency counterpart, the main points we have learnt from Figure 4.7 are summed up as follows:

- For both $s_1(n)$ and $s_2(n)$, the *instantaneous frequency* estimates keep closed to their real centers, except for the peaks due to periodicity.
- In $s_1(n)$, the *envelope* can reveal the signal period; while, in the scenario

taken by $s_2(n)$, the envelope indicates exactly wave $\left|cos(2\pi f_0 n/f_s)\right|$, where $f_0 = 10Hz$ is found to be the greatest common divisor of the two frequencies included.

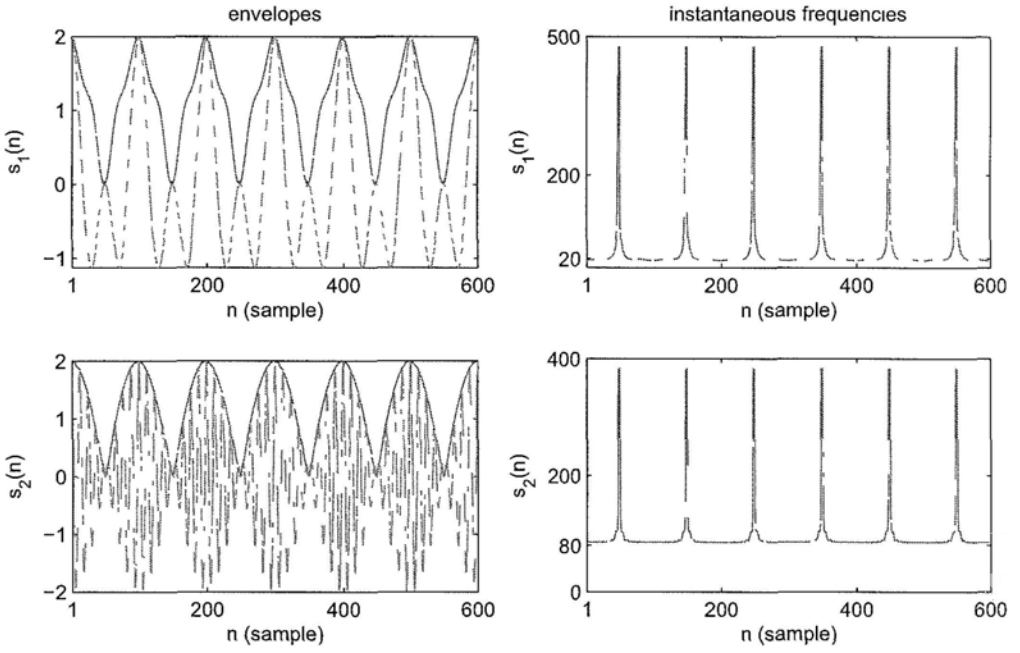


Figure 4.7: *Envelopes and instantaneous frequencies of the composite sinusoidal signals* $s_1(n)$ *and* $s_2(n)$.

4.3.2 Observations on real speech data

The above mentioned phenomenon that caused by the fundamental frequency is referred to as the pitch effect when processing speech or excitation signals. A segment $s(n)$ of speech sound /i/ uttered by a female speaker and its residual signal $e(n)$ by LP analysis are shown in Figure 4.8, note that they are both normalized to the range [-1, 1]. The residual signal is decomposed into 10 channels with a bank of Gamma-tone filters whose center frequencies and bandwidths are listed in Table 4.1.

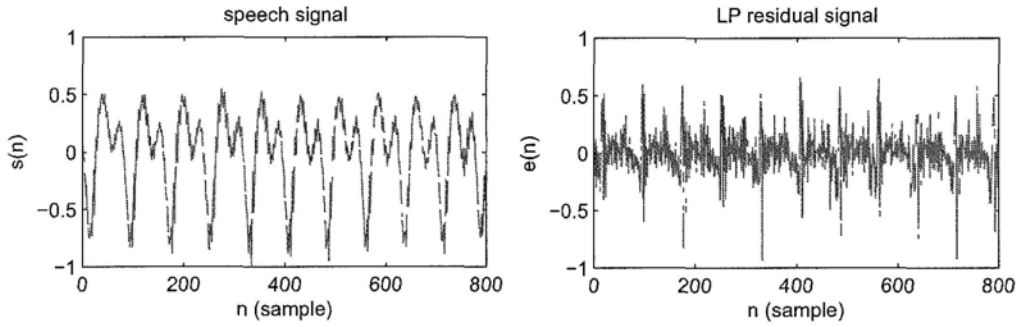


Figure 4.8: *Speech segment /i/ and its LP residual signal from a female speaker.*

Table 4.1: *Center frequencies and ERB bandwidths of a 10-channeled Gammatone filter bank spaced on [100Hz, 4k Hz] (in Hz).*

k	1	2	3	4	5	6	7	8	9	10
$f_c(k)$	3046.8	2308.5	1736.5	1293.5	950.4	684.6	478.7	319.2	195.7	100
$ERB(k)$	353.8	274.0	212.2	164.4	127.3	98.6	76.4	59.2	45.8	35.5

Figure 4.9 manifests the IF estimates in these subbands of residual signal. The five curves in the left column corresponds to the $k = 1, \ldots, 5$ subbands, while those with $k \in [6, 10]$ are shown in the right. It is clearly seen that in almost all subbands where excitation activity occurs, there are peaks present. Most peaks present themselves roughly at the pitch period in this observation, which is a vivid illustration of the pitch effects that take place in real speech.

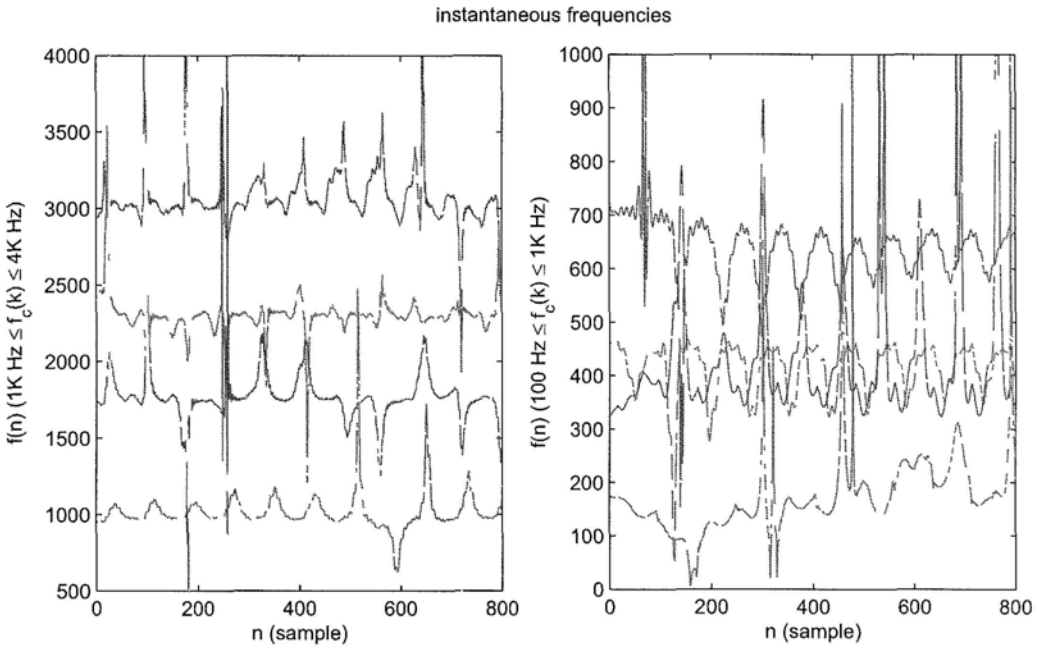


Figure 4.9: *Instantaneous frequency estimates in different subband signals.*

4.3.3 Discussion on excitation signal re-synthesis

The two factors affecting the estimation of pertinent envelopes and instantaneous frequencies in an LP residual signal have been introduced. For signal reconstruction, it is necessary to accurately estimate the AM-FM model parameters and get rid of any consequences from these effects. With the purpose of extracting speaker-distinctive features from the amplitude and frequency estimates, it is not a necessity to have the mentioned effects entirely compensated. Nevertheless, a brief study on them will be adequate.

Scenario 1: effects by band-pass filtering

Suppose each subband of the excitation is a mono-component AM-FM signal $x(n) = A(n)cos[\Theta(n)]$, as that laid out in Equation 4.9. We may rewrite it in the following form,

$$x(n) = A(n)cos\big[\Omega_c n + \phi(n)\big] = A_1(n)cos(\Omega_c n) - A_2(n)sin(\Omega_c n). \qquad (4.28)$$

Similarly, there is

$$A(n)sin\big[\Omega_c n + \phi(n)\big] = A_1(n)sin(\Omega_c n) + A_2(n)cos(\Omega_c n). \tag{4.29}$$

Referring to formula $asinx \pm bcosx = \sqrt{a^2+b^2}sin[x \pm arctan(b/a)]$, we can get the solutions for $|A(n)|$ and $\phi(n)$ as below,

$$|A(n)| = \sqrt{[A_1(n)]^2 + [A_2(n)]^2}, \quad \phi(n) = arctan\left[\frac{A_2(n)}{A_1(n)}\right]. \tag{4.30}$$

The Gamma-tone filter centered at carrier frequency f_c that present in Equation 4.27 will transform to the representation $g(n) = a(\frac{n}{f_s})^{(N-1)}e^{\frac{-2\pi bn}{f_s}}cos(\frac{2\pi f_c n}{f_s} + \phi)$ in the discrete time domain. It can be rewritten as the product of a low-pass filter $g_l(n)$ and a sinusoidal tone signal $cos(\Omega_c n) = cos(\frac{2\pi f_c n}{f_s})$ as follows,

$$g(n) = g_l(n)cos(\Omega_c n). \tag{4.31}$$

The filtered signal $\tilde{x}(n) = x(n) \otimes g(n)$ is still of AM-FM structure as denoted by Equation 4.32, where the carrier is Ω_c,

$$\tilde{x}(n) = \tilde{A}(n)cos\big[\Omega_c n + \tilde{\phi}(n)\big]. \tag{4.32}$$

Under this condition, it is approximately found that

$$\begin{aligned}
\tilde{x}(n) &= \big\{A_1(n)cos(\Omega_c n) - A_2(n)sin(\Omega_c n)\big\} \otimes \big[g_l(n)cos(\Omega_c n)\big] \\
&= \big[A_1(n)cos(\Omega_c n)\big] \otimes \big[g_l(n)cos(\Omega_c n)\big] - \big[A_2(n)sin(\Omega_c n)\big] \otimes \big[g_l(n)cos(\Omega_c n)\big] \\
&\approx \frac{1}{2}\big[A_1(n) \otimes g_l(n)\big]cos(\Omega_c n) - \frac{1}{2}\big[A_2(n) \otimes g_l(n)\big]sin(\Omega_c n).
\end{aligned} \tag{4.33}$$

The envelope $|\tilde{A}(n)|$ and $\tilde{\phi}(n)$ of $\tilde{x}(n)$ are therefore estimated by Equations 4.34 and 4.35.

$$|\tilde{A}(n)| \approx \frac{1}{2}\sqrt{[A_1(n) \otimes g_l(n)]^2 + [A_2(n) \otimes g_l(n)]^2}, \tag{4.34}$$

$$\tilde{\phi}(n) \approx arctan\left[\frac{A_2(n) \otimes g_l(n)}{A_1(n) \otimes g_l(n)}\right]. \tag{4.35}$$

Scenario 2: peaks by pitch effects

Most peaks present in the envelopes and instantaneous frequencies in some bands of the speech signal are due to the pitch effect. This actually provide cues for pitch determination. Because the involving of a bank of band-pass filters, this method is categorized to be a time-frequency domain pitch estimation approach, it is widely employed in speech segregation and harmonic enhancement researches [141], [142]. Inevitably, there exist discontinuities and disturbing noises caused by errors in parameter estimation, they are ought to be removed in a proper way, without affecting the real underlying frequency components. Then the primary frequency components can be extracted subsequently. There are methods can take up this task. We have achieved this via smoothing the estimate sequences, detailed description is represented in Section 4.4.1.

4.4 Analysis for Extracting Source Features

Cepstral coefficients are widely used in recognizing phonemes and discriminating speakers, such as Mel-frequency cepstral coefficients (MFCC). Good performance has been achieved using this type of features, however, it is found that they have bias towards the content of the speech unit [143], and are sensitive to the environmental variations. This motivates the exploration of new features that can offer assistances to the conventional cepstral coefficients in practical applications. Features derived for capturing the LP residual characteristics are proven complementary information source to the vocal tract based parameters, as revealed by the survey on source parameters in Chapter 2.

Recently, the modulation property of the vocal tract system has been involved in discriminating individual speakers. For example, the spectral centroids that depends on the FM of signals by Paliwal *et al.* in [144], average of instantaneous frequencies weighted by amplitudes by Dimitriadis *et al.* in [55], etc. Published results indicate that the information captured by the FM features can offer assistance to enhancing recognition performance when jointly used with the conventional amplitude-based features. However, most of these features focus on the temporal modulation properties of the formants in the vocal tract system, and exclude the excitation characteristics. The sinusoidal model [130], on the other hand, fits the speech waveform by composing a set of sinusoids which are harmonically related to the fundamental frequency of the speech signal. This method is primarily employed for pitch tracking [145], and is used in excitation coding technique. Other models were proposed, but mostly take advantage of the modulation properties of the excitation source and pay little attention to their speaker discrimination potentials.

In this section, we attempt to explore the speaker-relevant characteristics of the modulation phenomena in the excitation source of speech. Unlike the synthesis and coding systems, whose focus are speech intelligibility, waveform matching or transmission load, our method concentrates on characterizing the slow temporal (envelope and frequency) modulations in the LP residual signal, by using multi-band analysis and the nonlinear signal processing method.

We propose a new set of modulation features, which are estimated from the multi-band AM-FM model of the residual signal. The parameters are noted as Averaged Instantaneous Envelope of Residues (RAIE) and Averaged Instantaneous Frequency of Residues (RAIF), respectively. In a multi-stream speaker recognition system, these features are used as the complementary speech features to MFCC.

4.4.1 Characteristics of modulation parameters

As an initial observation on the envelope and instantaneous frequency estimates for a set of subband signals from the LP residual signal in Section 4.3, we mostly mentioned the problems came across and say less about other aspects. In this section, we will dig into and fetch out the important excitation-related characteristics present in the estimates. As a further investigation, study aimed to distinguishing the essential components from others are conducted for the sake of speaker discrimination.

Let us have a close examination on the estimates of modulation parameters in a single subband first, for instance, the $8th$ subband of the LP residual signal shown in Figure 4.8. By looking up Table 4.1, it is found that the center frequency $f_c(8) = 319.2Hz$. The waveform, instantaneous envelope, frequency estimates of the signal are given in Figure 4.10, additionally exhibited are the mean values of the two instantaneous sequences which are marked in magenta.

Observations on primary IE, IF components:

- The *primary frequency* of the subband signal is quite clearly delivered by the IF estimate, except for those peaks mentioned before in Section 4.3. This frequency might be accounted as one of the most essential and significant characteristics when indicating an AM-FM signal.

- The *frequency deviation* of the IF from the carrier frequency, about 100 Hz in this case, records the variation of the instantaneous frequency in an AM-FM signal.

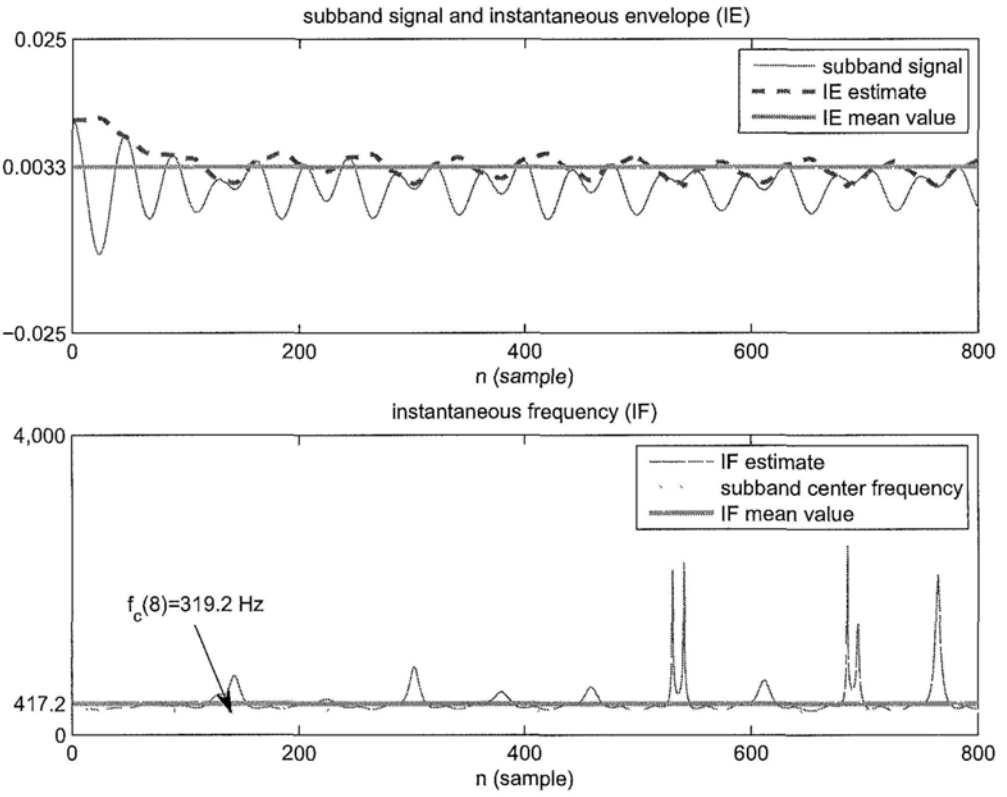


Figure 4.10: *An example subband signal, its IE, IF estimates and the corresponding mean values.*

- The *envelope* represents the absolute amplitude of the AM-FM signal, it roughly can be viewed as a primary measurement of the amplitude modulation modulus.

Attaining the primary IE, IF elements:

As also illustrated in Figure 4.10, the mean values of the IE, IF sequences are evaluated over a time window. To be precise, the smoothing process is described as follows:

- **Step 1 - Removing interruptions by smoothing:** to eliminate the peaks due to pitch effects and other disturbing noises present. Without losing too much temporal characteristics of the sequences, a 21-point median filter is chosen to conduct the smoothing work.

- **Step 2 - Retaining principal elements by taking average:** to retain the primary values for the sequences, their means value are taken within a proper window size.

4.4.2 Source features derivation

In Figure 4.11, a flow chart of the vocal excitation modulation feature extraction is shown. Besides, the composition of the source-related RAIE, RAIF parameters, as well as a complex set of them, parameter set RAIEF are exhibited.

The process of computing the RAIE and RAIF features is summarized as follows:

1. *Voicing decision*: The RAIE and RAIF features are extracted from voiced speech only. The voicing status is detected using Talkin's Robust Algorithm for Pitch Tracking [146].

2. *LP filter estimation*: To obtain the prediction filter coefficients $\{a_k\}, k = 1, 2, \ldots, p$ for yielding the LP residual signal.

3. *LP inverse filtering*: The residual signal $e(n)$ is obtained for each frame by taking LP inverse filtering. A voiced segment is divided into overlapping frames with 30 msec duration and 10 msec frame shift. To diminish intra-speaker variation, the amplitude of the residual segment is normalized to the range of $[-1, 1]$.

4. *Filter bank filtering*: Applying a bank of K Gamma-tone filters on the residual signal $e(n)$ to produce the subband signals. The center frequency $f_c(k)$ ranges from 4 kHz to 80Hz with k increase from 1 to K.

5. *Multi-band demodulation*: Teager's energy separation algorithm is employed in obtaining the instantaneous envelope (IE) sequence $|A(n)|$ and the instantaneous angular frequency $\frac{2\pi}{f_s} f(n)$ (IF) on a frame basis for each subband signal.

6. *Smoothing of the IE and IF sequence*: A 21-point median filter is applied to remove the abrupt impulses in the frames of IE and IF sequence.

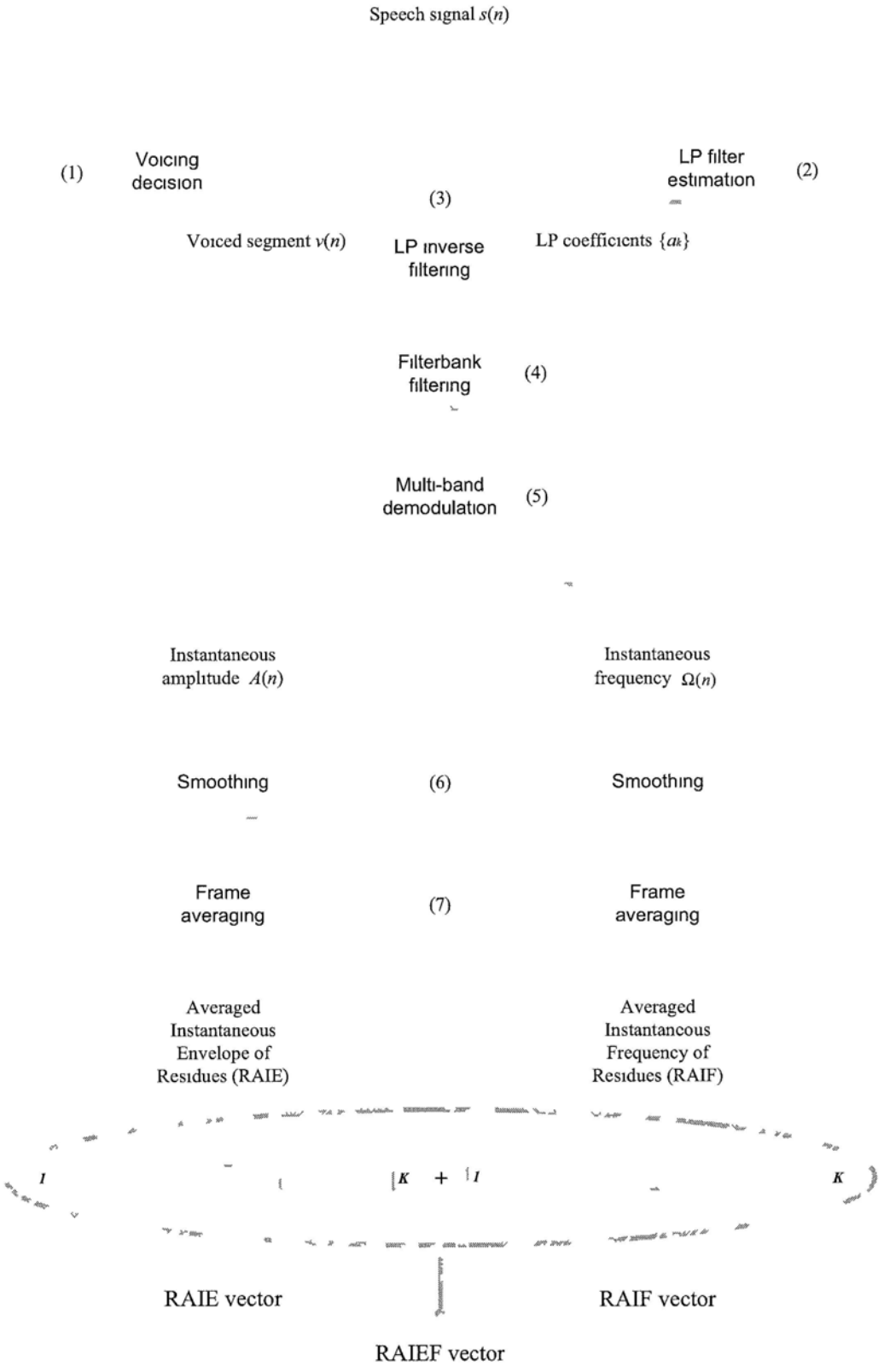


Figure 4.11 *Block diagram for the extraction of RAIE, RAIF and RAIEF feature vectors*

7. *Frame averaging of the smoothed IE and IF*: An averaging operation is done on the smoothed IE and IF sequences for the frames in each subband. In this step, we remove the fluctuations of the IE and IF sequences, and track the amplitude and frequency of the most significant frequency components in each subband frame by frame.

At the end of the above procedures, the K dimensional feature vectors RAIE and RAIF are derived. Given one particular frame, the RAIE and RAIF vectors are viewed as the amplitude and frequency distributions of the principal components over the frequency bands. The RAIE and RAIF feature vectors are described as

$$RAIE = \{RAIE(1), RAIE(2), \ldots, RAIE(K)\},$$

and

$$RAIF = \{RAIF(1), RAIF(2), \ldots, RAIF(K)\}.$$

In order to include both RAIE and RAIF-related parameters in one single vector, we create another set of features, that is, the RAIEF vector, whose composition is demonstrated through Figure 4.11.

The speech data used in the experiments of this paper are sampled at 8 kHz, and thus the speech has the highest frequency of 4 kHz. Since there is no strict derivation to determine an optimal subband number K, and the frequency resolution by the 20-channel Gammatone filterbank can separate the harmonically related frequency components for the data we used, we consider to use $K = 20$ here.

4.4.3 Feature analysis

The RAIE and RAIF features derived above are considered to have captured the amplitude and frequency characteristics of the principal components of the different subbands. In this part, we would like to evaluate these parameters by examining some typical speech properties. These properties cover the variation in F0, the difference in pitch epoch shape, and the relevant excitation details

existing between the adjacent pitch epochs For this purpose, we have first synthesized four signals to assume the desired pitch-related characteristics, they are noted as $e_1(n)$ through $e_4(n)$, respectively The detailed specifications about these signals are listed in Table 4 2

Table 4 2 *Specifications of the synthetic excitation signals ($f_s = 8kHz$)*

	$F0$ (Hz)	Epoch shape	Details?
$e_1(n)$	86 2	Impulse	No
$e_2(n)$	172 4	Impulse	No
$e_3(n)$	172 4	Triangular pulse	No
$e_4(n)$	172 4	Triangular pulse	Yes

Moreover, the F0 value of the synthetic signals and center frequencies of some band-pass filters are settled specially, as shown below

$f_c(18) = F0 = 172\ 4Hz$	$f_c(15) = 2\ 1F0$	$f_c(13) = 3\ 2F0$

Then, a set of experiments are designed specifically to evaluate the features in the following three aspects

- **Pitch variation**

F0 is one of the most primary properties in distinguishing different speakers by human ears B S Atal in [23] used pitch contour to identify speakers, and there have been pitch-related features proposed throughout the years for similar purposes, as we have reviewed in Chapter2 In this experiment, we generate two impulse trains to approximate the pitch-periodicity of the excitation signal By taking these different pitched signals as input, we intend to see whether the AIE and AIF feature vectors can produce discriminative information

In Figure 4 12, there are two impulse trains which are denoted as $e_1(n)$ and $e_2(n)$, respectively Their waveforms are shown on the top row $e_2(n)$ has chosen F0 to be 172 4Hz, which is the same as $f_c(18)$ (center frequency of the 18th filter), it is equivalent to 2 1 times of $f_c(15)$ and 3 2 times of $f_c(13)$ $f_c(k)$

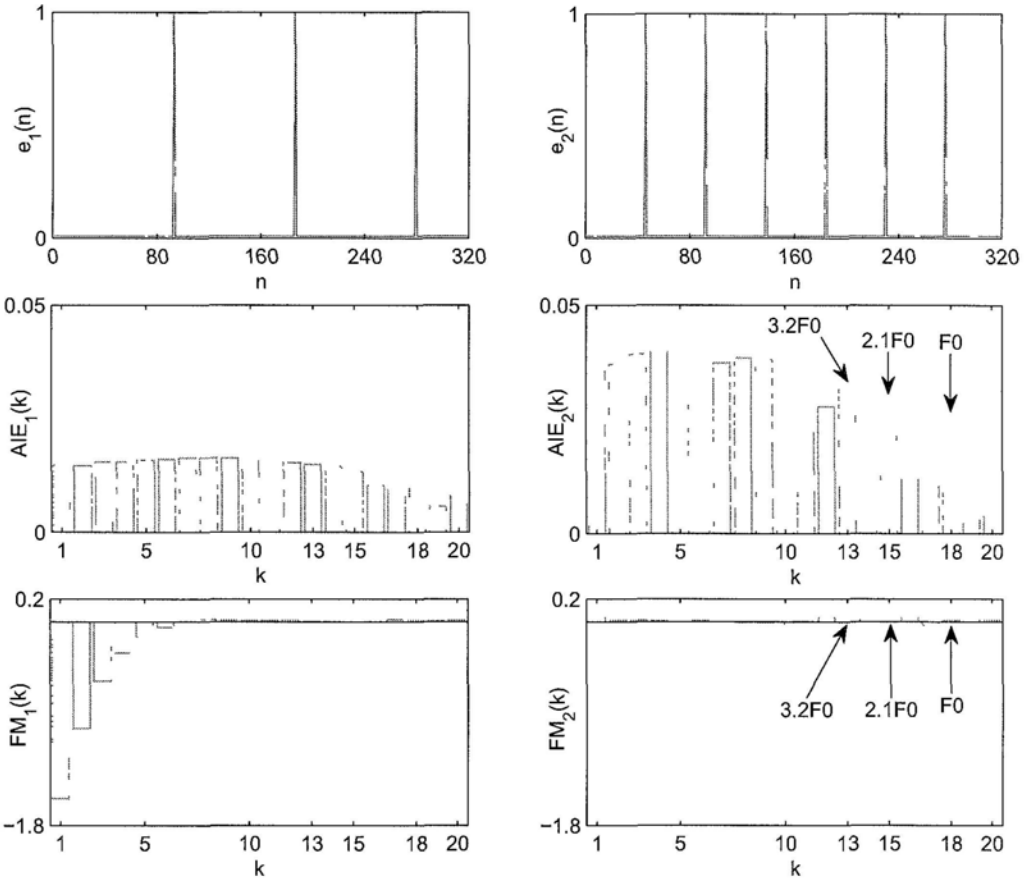


Figure 4.12: *AIE and FM for artificial excitation signals with different F0: $e_1(n)$ and $e_2(n)$.*

decreases as k increases. The F0 of signal $e_1(n)$ is set to be half of $e_2(n)$. On the middle row, the AIE of $e_1(n)$ and $e_2(n)$ are illustrated. It is observed that in the lower frequency region, for $e_2(n)$, the $k = 18, 15, 13$ frequency bands have small peaks for AIE values, and there are no peaks for $e_1(n)$. In the high frequency region, most k in $e_2(n)$ give much higher AIE values than $e_1(n)$, and the AIE values of $e_1(n)$ are more evenly distributed among different k's. On the bottom row, the FM values are revealed to have different distributions over k for $e_1(n)$ and $e_2(n)$. As another expression of the primary frequency in a subband indexed by k, FM component alternatively delivers the deviation of $AIF(k)$ from $f_c(k)$. While it is found that the FM is in general small for

$e_2(n)$, at subbands $k = 18, 15, 13$, it tends to be very close to zero. Differs from what $e_1(n)$ behaves, $e_2(n)$ shows presence of greater negative FM in the higher frequency bands.

- **Pitch epoch shape**

Aside from the pitch period variation among speakers, it is believed that the shape of the pitch epoch and the details between adjacent epochs also play roles in discriminating different speakers. As we have briefly surveyed in Section 4.1, the shape of pulse is essential parameter in almost all voicing source models. In this experiment, these properties will be focused on separately.

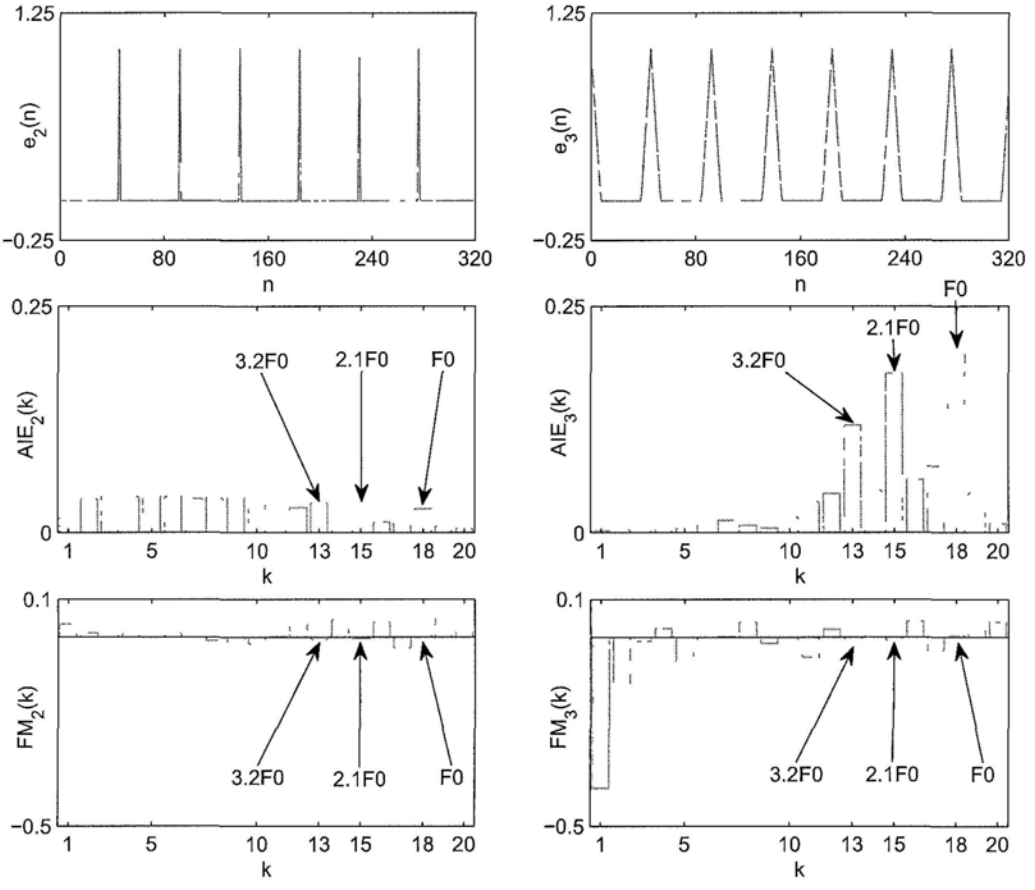


Figure 4.13: *AIE and FM for artificial excitation signals with different epoch shapes:* $e_2(n)$ *and* $e_3(n)$.

In Figure 4.13, in addition to signal $e_2(n)$, another signal $e_3(n)$ and their AIE, FM vectors are illustrated in the three rows, respectively, where both signals share the same F0=172.4 Hz. For the comparison between $e_2(n)$ and $e_3(n)$, it is revealed that although they have the same F0 value, their difference in pulse shape makes their AIE vectors differ a lot. We observe that the lower frequency amplitude is emphasized in $e_3(n)$, and the peaks at $k = 18, 15, 13$ bands are also more prominent for $e_3(n)$ than that for $e_2(n)$. With regard to the FM components on the bottom row, both signals exhibit peaks around the three bands concerned, however, the larger FM values for $e_3(n)$ in the higher frequency bands indicate the drop of energy in these bands.

- **Details between epochs**

Figure 4.14 displays another set of synthetic pulse trains $e_3(n)$ and $e_4(n)$ sharing a same F0. The comparison between them focuses on the effects of the embedded details among the epochs. It is seen that the AIE peaks present in $e_3(n)$ at $k = 18, 15, 13$ appear for $e_4(n)$ as well, but the enlarged amplitudes for $e_4(n)$ in the higher frequency bands, compare with that for $e_3(n)$, are obviously resulted from the additional noise. Likewise, on the bottom row that refers to FM components, it is found that the noise has brought greater effects for bands centered above 1000 Hz than for others. This is actually consistent with our observations from Figure 4.13 in that the noise in a way reduces the FM in some bands, possibly in the higher frequency region.

Indicated by the above experiments and compared with the conventional spectral analysis results, we can see that the AIE vector can reveal the amplitude modulation information in different frequency bands which is absent from the flat spectrum of the excitation signal. On the other hand, AIF provides phase-related information for vocal excitation.

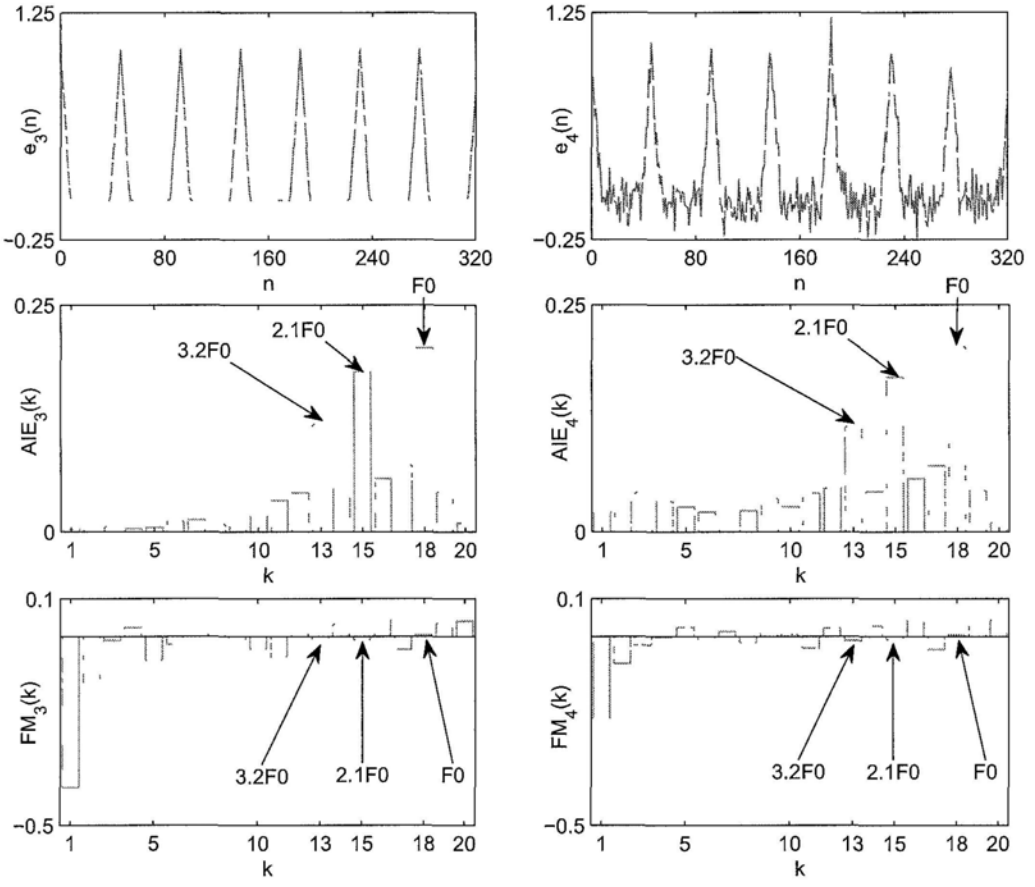


Figure 4.14: *AIE and FM for artificial excitation signals with and without details between adjacent epochs:* $e_3(n)$ *and* $e_4(n)$.

4.5 Evaluation of Excitation Modulation Features in Speaker Recognition

In this section, we will evaluate the proposed features RAIE and RAIF as the complementary parameters to the MFCC features through speaker identification (SID) and speaker verification (SV) experiments on a speech database CU2C. The standard MFCC features we used contain 39 components: the log energy, 12 static coefficients, and their dynamic and acceleration coefficients.

4.5.1 Speech database: CU2C

CU2C is a dual-condition Cantonese speech database developed for speaker recognition research at the Chinese University of Hong Kong [147] in 2005 (`http://dsp.ee.cuhk.edu.hk/html/cuothers.html`). It contains parallel data collected under two different acoustic conditions: the wideband desktop microphone and public fixed-line telephone channel. In the recording process, each speaker was asked to read the same materials twice under the two recording conditions, one after the other immediately. These two kinds of data can be used separately to develop different applications. Thus, it provides a proper platform for the study of channel effects in speaker recognition systems. We use part of CU2C, which contains the speech data from 50 male speakers. Each speaker has 18 sessions of recordings which were made over a time span of 4-9 months. There are 6 utterances in each session. Three sub-corpora with different contents are provided in CU2C: ID numbers, digit strings and sentences. We use the digit strings sub-corpora, where each of the utterance contains a sequence of 14 randomly generated digits. The utterance length is about 5 to 6 seconds. The original sampling frequency of the microphone data was 16 kHz, and they were down-sampled to 8 kHz to be used in this work. The telephone data were sampled at 8 kHz. Noisy speech data are generated by digitally adding white Gaussian noise signals to the CU2C utterances at controlled SNR level. The noise signals are taken from the NOISEX-92 Database [148].

4.5.2 Experimental set-up

For all speakers, 6 out of the 18 sessions are used to train the speaker models, and the remaining data are used for performance evaluation. The standard approach for UBM-GMM training is adopted. Separate systems are built based on MFCC, RAIE, RAIF, and RAIEF, respectively. The features RAIE, RAIF and RAIEF are used as complementary counterparts of MFCC, respectively, in the three sets of experiments. The score-level fusion technique is used to combine the contributions of the systems by MFCC and one of the source-related parameter set. The final decision is determined by the overall combined score. In the identification tasks, the log-likelihood score of each test is a linear combination of the log-likelihood scores from the MFCC and RAIE/RAIF/RAIEF features, with weighting parameters w_M and w_R (i.e., $L = w_M L_M + w_R L_R$). Meanwhile, in the verification tasks, the fusion is performed on the log-likelihood ratio scores, that is, $\lambda = w_M \lambda_M + w_R \lambda_R$. In both tasks, w_M and w_R are related by $w_M + w_R = 1$. The weighing strategy is described as follows: initially, let $w_M = 0$, and $w_R = 1$. Next, we empirically increase w_M by a step of $\frac{1}{50}$, and repeat this for 50 times, with $w_R = 1 - w_M$ satisfied. Finally, we identify the optimum parameter set $[w_M, w_R]$ with the best recognition results.

For SID and SV tasks, the identification error rate (IDER) and equal error rate (EER) are used as the primary performance indicators, respectively.

4.5.3 Experimental results

In this part, the proposed vocal excitation modulation feature sets RAIE, RAIF, RAIEF will be evaluated in terms of their individual speaker discriminative power as well as their complementarity with the the conventional MFCC features. Factors in extracting the source-related modulation parameters will be studied through experiments as well.

♦ **Benchmark results**

The individual performance of the MFCC and RAIE, RAIF parameter sets in clean environment under the protocols of speaker identification and verifica-

tion are shown in Table 4.3, respectively. Their results are therefore used in the baseline system to provide the benchmark record, where the MFCC features are of 39-dimension and the RAIE, RAIF feature vectors each contains 20 parameters, respectively.

Table 4.3: *Speaker recognition performance of individual feature sets: IDER & EER (in %).*

feature configuration	IDER	EER
MFCC	2.44	1.52
RAIE_*20*	40.72	13.17
RAIF_*20*	35.11	10.42

♦ Results by fusing with MFCC

The fusion of MFCC features with either RAIE or RAIF parameters leads to improved performance over the MFCC-alone speaker recognition system. Table 4.4 gives the results of the proposed two sets of source-related parameters after combining with MFCC. For reference, the averaged weighting parameter sets $[w_M, w_R]$ in getting these results are given in bottom row of the table.

Table 4.4: *Speaker recognition performance of combined feature sets: IDER & EER (in %).*

feature combination	IDER	EER
MFCC+RAIE_*20*	2.39	1.49
MFCC+RAIF_*20*	2.28	1.24
$[w_M : w_R]$	[0.65 : 0.35]	[0.63 : 0.37]

♦ Effects of feature dimension

In Table 4.5, the recognition performance of feature sets RAIE, RAIF with an increased subband number 40 are evaluated. An additional parameter set RAIEF with the same vector dimension is examined as well. The set of RAIEF

feature is composed of the RAIE and RAIF parameters under dimension 20 scenario, where the vector composition of it can be found in Figure 4.11. Averaged weighting parameter set $[w_M, w_R]$ in getting these results are given in bottom row of the table as well.

Table 4.5: *Effects of feature dimension on speaker recognition performance: IDER & EER (in %).*

feature configuration/combination	IDER	EER
RAIE_*40*	27.28	9.46
RAIF_*40*	19.67	8.01
RAIEF_*40*	22.50	8.17
MFCC+RAIE_*40*	2.44	1.44
MFCC+RAIF_*40*	**2.06**	**1.27**
MFCC+RAIEF_*40*	2.17	1.36
$[w_M : w_R]$	[0.58 : 0.42]	[0.63 : 0.37]

4.5.4 Analysis of results

♦ **Recognition accuracy analysis**

From the SID and SV results shown in Table 4.3, the performance of individual features RAIE and RAIF are far from comparable with those of MFCC features. With score-level fusion with the MFCC features, these two sets of 20 dimensional parameters achieve some improvements over the MFCC-only scenario, as shown in Table 4.4. When the dimension of the feature vectors, i.e., the number of frequency band for signal demodulation, becomes 40, their combinations with MFCC in Table 4.5 indicate enhanced performance than the lower dimensional cases.

♦ **Effects of feature dimension**

When looking through either feature RAIE or RAIF in terms of their individual performance, it is found that both of these parameters have attained improved results after their dimension was doubled. The enhancement might

be examined by comparing the corresponding IDERs and EERs indicated in Table 4.3 and Table 4.5. In Figure 4.15, there are two subfigures exhibiting the relative reduction of IDER and EER caused by increasing the dimension of the feature vectors. Figure 4.15(a) illustrates the improvement over the results of RAIE_*20* assumed by RAIE_*40* and RAIEF, respectively. Similarly, in Figure 4.15(b), the improvements achieved by RAIF_*40* and RAIEF are measured based on the benchmark results of RAIF_*20*.

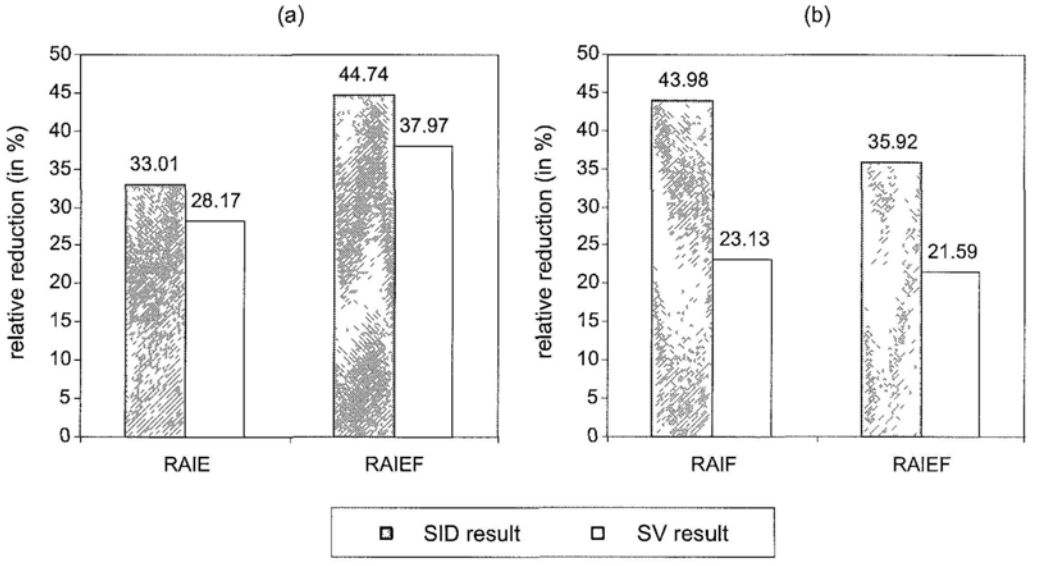


Figure 4.15: *Effects of feature dimension on experimental results.*

However, the case of RAIEF is dissimilar with the above two. On one hand, its dimension is the same with both RAIE_*40* and RAIF_*40*, however, the subband number in signal decomposition is still 20. On the other hand, it is a complex of the RAIE_*20* and RAIF_*20* sets. Although the two 20 dimensional features are found not effective no matter in individual or combination scenarios, their complex can attain considerable enhancement in both cases.

- **AIF *vs.* AIE**

As we can see from Figure 4.16 that the performance of the three feature configurations: AIE, AIF and AIEF that applied the excitation signals tend

to behave similarly for SID and SV tests. A same trend is observed in both the individual results of these three feature sets and their performance after fusing with MFCC, that is, the parameter set AIF outperform the other two, and the set AIEF shows superiority than that of AIE in terms of recognition performance.

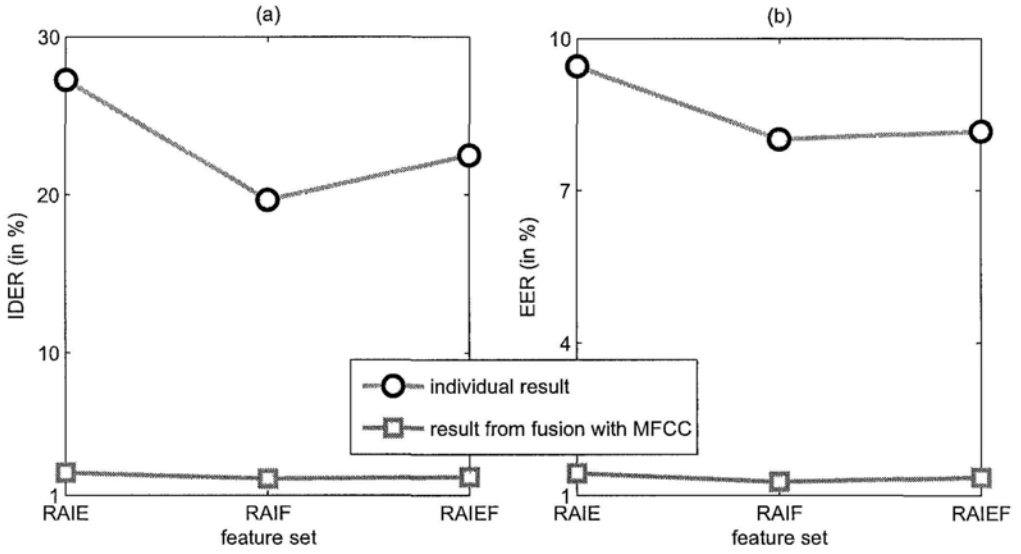


Figure 4.16: *Results of RAIE, RAIF and RAIEF features when employed individually and combining with MFCC for : (a) SID experiments; (b). SV experiments.*

♦ Feature complementarity analysis

Concerning the combined SID and SV results recorded in Tables 4.4 and 4.5, the averaged weighting factor w_R is around $0.35 \sim 0.42$ in both experiments. This reveals the high complementary relationship between the MFCC and RAIE/RAIF features in representing the acoustic characteristics of an individual speaker.

4.6 Summary

This chapter aims to explore the vocal excitation source properties in terms of the temporal modulations in both amplitude and frequency. The study to characterize individual speakers' vocal excitation pattern indwelling is conducted under the framework of multiple frequency-band decomposition and modeling. In extracting useful excitation-related modulation parameters from the LP residual signals, the amplitude and frequency components are separated over the time, their distribution across the multiple bands are then parameterized into feature vectors. It is observed via the designed experiments on synthetic signals that the proposed features are capable of capturing the vocal differences in terms of F0 variation, pitch epoch shape, and relevant excitation details between epochs. Subsequently, these spectro-temporal parameters are evaluated through simulations on real database. It is found that (1). multi-band amplitude and frequency modulation parameters are capable to capture the time-frequency vocal excitation characteristics; (2). modulation-related source parameters are complementary with the relevant vocal tract features; and (3). speaker recognition accuracy provided by the spectral-based features can be further improved by combining the proposed source features in a multi-stream recognition system.

Chapter 5

Speaker Discrimination using Phase Information of Speech Signal

Auditory experiments show insensitivity of human ears to phase information in perceiving phonetic content of speech signal. However, the discarded phase information may provide useful acoustic cue for identifying individual speaker, this is especially useful for speaker recognition systems operated with degraded magnitude in adverse conditions. This chapter is therefore motivated to derive phase-related features for reliable speaker recognition performance. A pertinent representation for most dominant primary frequencies present in the speech signal is first built. It is then applied to frames of the speech signal to derive effective speaker-discriminative features. Through a set of specifically designed experiments on synthetic vowels, it is observed that the proposed features are capable of differentiating the inclusive formants, pitch harmonics from other components, and expressing the vocal particularities in various-shaped formants. By combining with standard cepstral parameters, these phase-related features have shown to evidently reduce the identification error rate and equal error rate in the context of Gaussian mixture model-based speaker recognition system.

In Section 5.1, representations characterizing the vocal tract resonances are reviewed at first. With attempt to derive phase-related speaker representa-

tives, in Section 5.2, we employ the set of frequencies that dominate in the primary components as a representation of speech phase, which then in Section 5.3 is evaluated in its capability of delivering potential speaker-discriminative properties. Finally in Section 5.4, the pertinent representation is applied on a frame-by-frame basis in the front-end to be assessed under speaker recognition protocols. Section 5.5 summarizes the chapter.

5.1 Vocal Tract Resonance Characterization

An efficient and compact representation for the speech properties that keep stationary within a period of time is of distinct importance for distinguishing different speakers as well as for speech recognition. Similarly as we have done in the previous chapter to build a model for conveying the excitation-related vocal attributes, we are going to explore the concerned vocal tract characteristics here through an adequate manner of expression. There are a variety of approaches, such as, linear prediction analysis, formant and bandwidth tracking [149], [150], articulatory models [151], [152], etc, have been developed for characterizing vocal tract properties, most of them are focused on capturing the concerned characteristics of formants, which are resonances of the vocal tract. The speech modeling method that models the pole frequencies of speech signal or transfer function of the vocal tract based on source-filter model is referred to as formant synthesis [153].

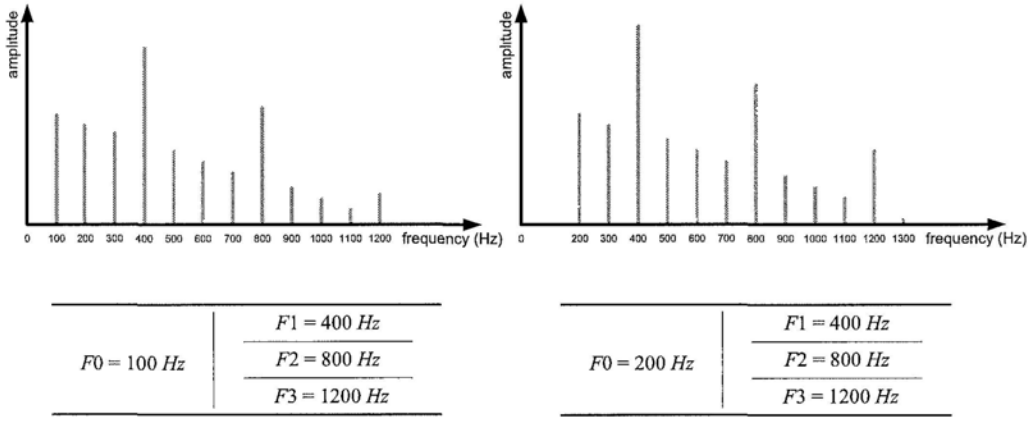


Figure 5.1: *Interaction between the excitation source and formants.*

The human vocal system generally comprises of an excitation generator and a set of band-pass filters. The excitation generator operates in at least two modes: pitch-controlled oscillator, or a noise generator, while the vocal tract exhibits resonant modes that emphasize some frequencies and simultaneously suppressing others. Each formant is usually modeled with a two-pole resonator which

enables both the formant frequency (pole-pair frequency) and its bandwidth to be specified. Acoustically, it is the formants that make the vowels sound different one from another, and in pronouncing a same vowel, inter-personal differences in the anatomical properties such as the length and cross-section of the tube of air that comprises the vocal tract, shape the voices with discernable attributes. Figure 5.1 depicts the interaction between the excitation source and the formants in generating an exemplary vowel. The two sounds are determined to share identical formant frequencies F1, F2, F3, but excited by different fundamental frequency F0, which as a result, exhibits discernable excitation-formant interaction behaviors.

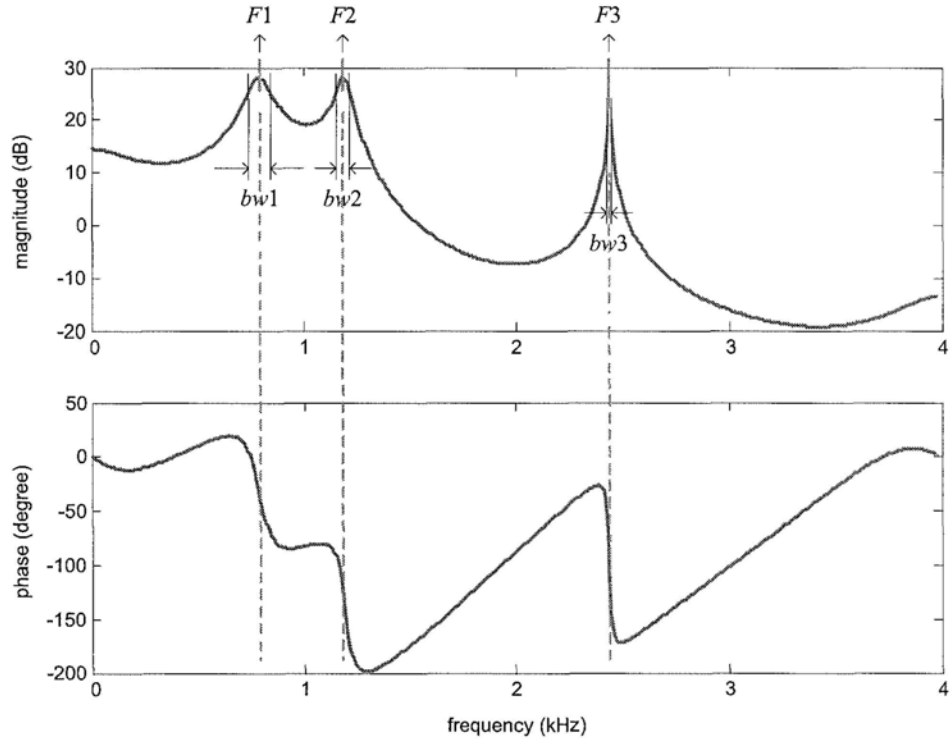


Figure 5.2: *Formant structure expressed by the frequency response of an all-pole model.*

Another essential attribute that marks the spectral envelope of a voiced speech sound is the frequencies and bandwidths of the primary formants, as illustrated by Figure 5.2.

5.1.1 Conjugate pole-pair model

The formants are viewed as the resonances in vocal tract system. In representing a resonance spectrally, autoregressive (AR) model is a frequently used method. For speech signal processing, the two poles in a second-order formant AR model usually constitute a conjugate pair that located within the unit circle [19]. A typical complex resonant frequency of the vocal tract is

$$s_k, s_k^* = -\sigma_k \pm j2\pi F_k. \tag{5.1}$$

While, the corresponding conjugate poles represented in the discrete time would be

$$z_k, z_k^* = exp\big(-\frac{\sigma_k}{f_s}\big)exp\big(\pm j\frac{2\pi}{f_s}F_k\big). \tag{5.2}$$

Center frequency of the kth resonance is F_k, the bandwidth is approximately $2\sigma_k$. In the z transform domain, the bandwidth is determined by the radius of the poles, i.e.,

$$r_k = |z_k| = exp\big(-\frac{\sigma_k}{f_s}\big), \tag{5.3}$$

while, the angles of the conjugate poles from the origin are

$$\theta_k = \pm\frac{2\pi}{f_s}F_k. \tag{5.4}$$

Therefore, the transfer function of a specific formant in terms of the conjugate pole pair ought to be expressed as in Equation 5.5,

$$R_k(e^{\jmath\omega}) = \frac{1}{1 - 2r_k cos\theta_k e^{-\jmath\omega} + r_k^2 e^{-\jmath 2\omega}} = \frac{1}{\big[1 - (r_k e^{\jmath\theta_k})e^{-\jmath\omega}\big]\big[1 - (r_k e^{-\jmath\theta_k})e^{-\jmath\omega}\big]}. \tag{5.5}$$

Accordingly in the time domain, the resonance is noted as

$$r_k(n) = r_k^n \cdot \frac{sin\big[(n+1)\theta_k\big]}{sin\theta_k} \cdot u(n). \tag{5.6}$$

Obviously, Equation 5.6 can be approximated alternatively as product of an amplitude and sinusoid as follows [57],

$$r_k(n) = \left[\frac{r_k^n}{sin\theta_k}\right] cos\left[(n+1)\theta_k - \frac{\pi}{2}\right]. \tag{5.7}$$

5.1.2 AM-FM representation

Taking up the formulation of kth formant of the vocal tract system in Equation 5.2, given an impulse-train to excite it, it will lead to a bandlimited signal $F_k(n)$, which can be represented by

$$F_k(n) = A_k(n)exp\Big\{j\Big[\Theta_k(n)\Big]\Big\}, \tag{5.8}$$

with the formant being characterized by two sequences:

- $A_k(n)$ – Amplitude of formant;
- $\Theta_k(n)$ – Phase of formant.

In practice, the vocal tract system has long been known as a modulation system, where individual resonances modulate the vocal excitation at separate frequencies, which as a result present peaks in the spectrum and affect the phase properties. The formant signal as we have denoted in Equation 5.8 is a bandlimited signal centering around the formant frequency and spanning approximately over the formant bandwidth. Being a similar case to the vocal cords' vibration studied in Chapter 4, the resonant voice caused by resonation of vocal tract with the excitation source at a specific frequency follows the law of AM-FM modulation as well. Therefore, the bandlimited resonant signal is viewed as a mono-component AM-FM signal. Message conveyed here is revealed consistent with that in Equation 5.7 too, in that the formulation is determined by sequences of amplitude and phase quantity.

As far as the time-varying property of a speech signal is concerned, the amplitude and phase of a resonant signal are deemed to be time-variant sequences. Referring to the expression in Equation 5.7, these sequences could subsequently be delivered by the following terms:

$$A_k(n) = \frac{r_k^n}{sin\theta_k}, \tag{5.9}$$

$$\Theta_k(n) = (n+1)\theta_k - \frac{\pi}{2}. \tag{5.10}$$

Thus, we can rewrite Equation 5.7 to be $r_k(n) = A_k(n)cos[\Theta_k(n)]$, in a similar manner with that in Equation 4.9. When making equal the phase quantity

of Equations 4.11 and 5.10, i.e., $(n+1)\theta_k - \pi/2 = \Omega_c(k)n + \sum_{r=1}^{n} q_k(r) + \phi$, we get

$$\theta_k = \Omega_c(k) + q_k(n), \tag{5.11}$$

where $\Omega_c(k)$ denotes the carrier frequency (center) of the AM-FM signal, and $q_k(n)$ records the deviation of the instantaneous frequency from the center. ϕ is initial phase offset. The frequency of the formant here is obviously an instantaneous quantity, thus, by employing the formulation of mono-component AM-FM representation in Equation 4.9, we literally replace θ_k with symbol $\Omega_k(n)$, and obtain the resonant AM-FM signal expression as

$$r_k(n) = A_k(n)cos\big[\Omega_c(k)n + \sum_{r=1}^{n} q_k(r) + \phi\big]. \tag{5.12}$$

In practice, quite a few works on formant modeling/tracking and formant synthesis techniques have taken advantages of representing the vocal tract resonances by means of AM-FM form.

5.1.3 Observations on synthetic vowels

A specific formant in the vocal tract system is usually determined by its center frequency and bandwidth. Alternatively in the time domain, the corresponding resonant signal is delivered by instantaneous envelope and frequency quantities under the framework of AM-FM modeling. Therefore, we could take advantage of resonant signal modeling to deliver concerned vocal tract-related characteristics of a person.

In order to have a good idea of resonance modeling by AM, FM parameters, we inspect the concerned practice on two synthetic vowels at different perspectives. As for the exemplary vowels, scrutinies on them are made for both the vowel as a whole and their constitute parts, i.e., the individual resonant signals. The course of amplitude-frequency separation as introduced in Chapter 4 is followed, with relevant observations in spectral, temporal domains reported therewith.

Let us see waveform of the synthetic vowels in Figure 5.3 at first. The specification by which the vowels are generated are subsequently tabulated in

Table 5.1.

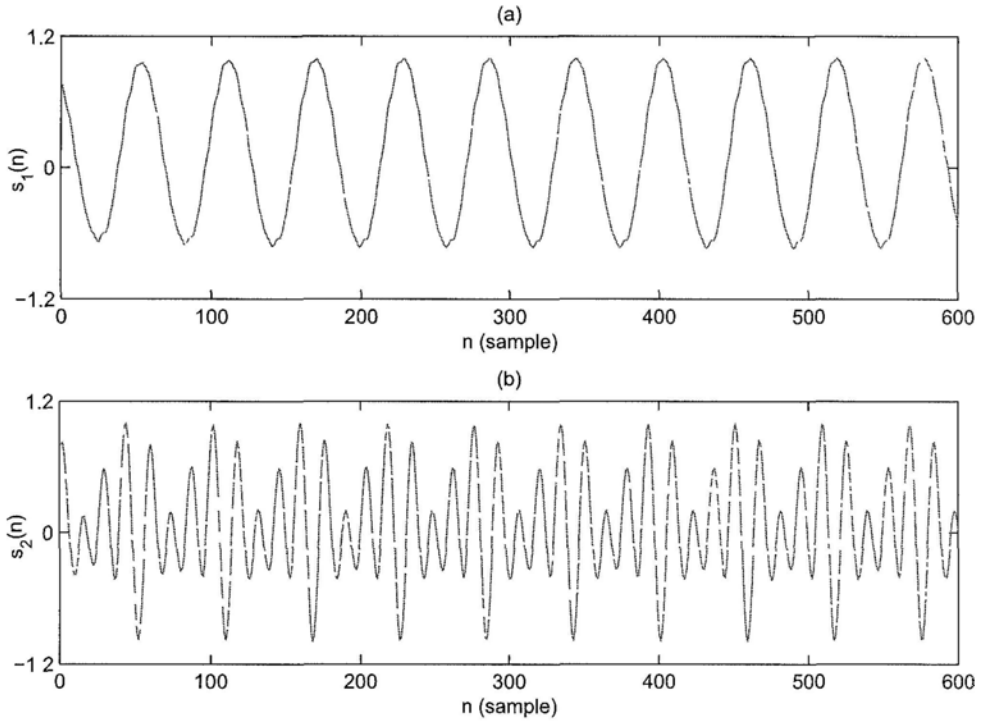


Figure 5.3: *Waveform of synthetic vowels: (a) /i/ and (b) /a/.*

Table 5.1: *Specifications of the synthetic vowels ($f_s = 16kHz$).*

	$F0\ (Hz)$	$F1,\ F2,\ F3\ (Hz)$	$bw1, bw2, bw3\ (Hz)$
/i/	275	270, 2290, 3010	50
/a/	275	730, 1090, 2440	50

In Figure 5.4, the synthetic vowels are depicted through AR modeling. Formants inclusive are present in the form of spectral peaks. Via applying separate band-pass filters on signals $s_1(n)$ and $s_2(n)$, which centered around the predetermined formant frequencies with a constant Q factor $Q_i = F_i/bw_i = 9,\ i = 1, 2, 3$, we pick out a set of three resonant signals for each vowel. Spectral model of these segmented formants are demonstrate in Figure 5.4 as well. Additionally in the time domain, they take the form of resonant signals as shown in Figure 5.5, where the resonant signals are noted as $r_1(n)$, $r_2(n)$ and $r_3(n)$, respectively.

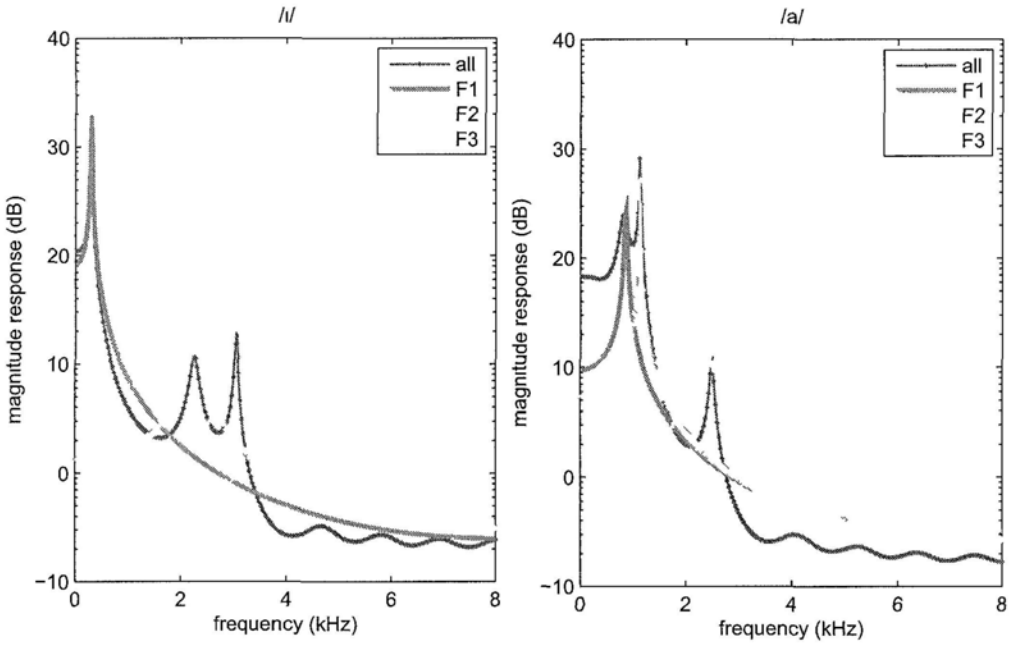


Figure 5.4: *Formant structure of the synthetic vowels /i/ and /a/.*

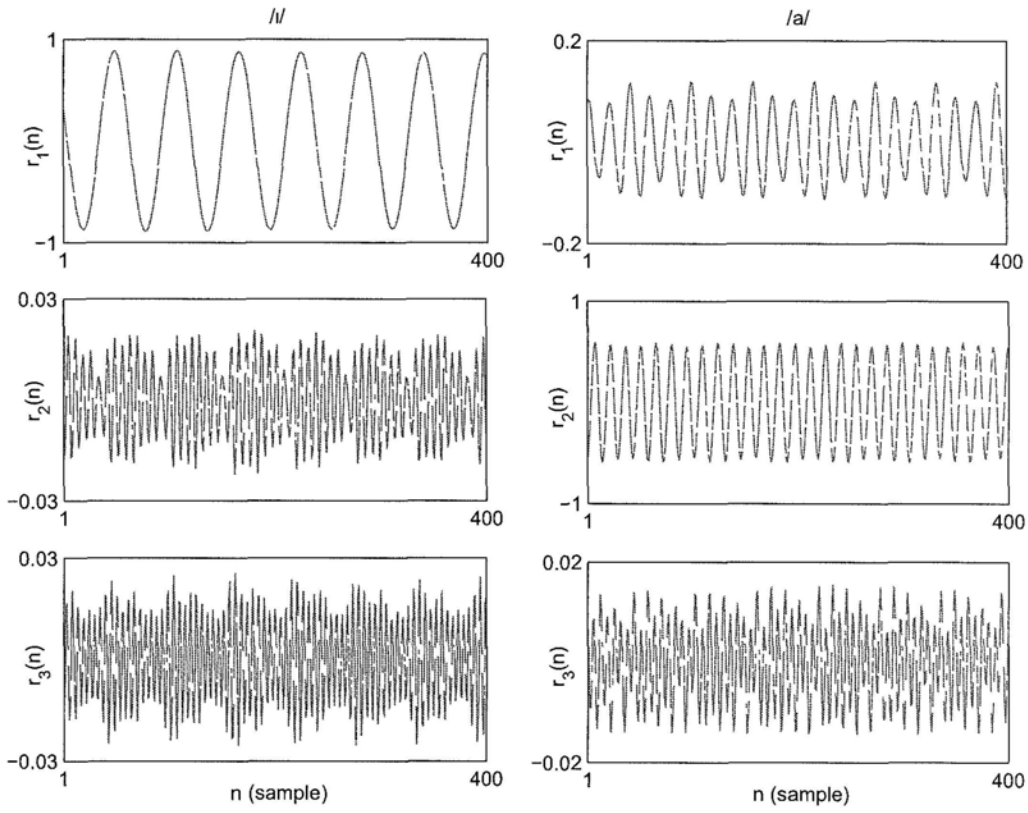


Figure 5.5: *Waveform of resonant signals* $r_1(n)$, $r_2(n)$, *and* $r_3(n)$ *segmented from synthetic vowels /i/ and /a/.*

The algorithm for decomposing the vibratory excitation signals into corresponding envelope and instantaneous frequency sequences, as described in Section 4.2, is found fit for the resonant signals as well. By applying the DESA demodulation algorithm for a mono-component AM-FM signal that described therein, we consequently have captured frequencies that dominate the resonances $r_1(n), r_2(n), r_3(n)$ respectively, on a segment basis and have separated them from the envelopes thereafter. In Figure 5.6, those extracted sequences are laid out, where the upper- and lower-row illustrates the sequences of instantaneous envelope (IE) and instantaneous frequency (IF), respectively.

Observations we make for the extracted IE, IF quantities are given out in following paragraphs.

Observations on primary IE, IF components:

- The *primary frequency* of resonant signals is quite clearly delivered by their IF estimates with little disturbance. For specific resonant signals, mean value of the detected frequency quantities is thereafter found quite close to the predefined spectral centers.

- The *envelope* represents the absolute amplitude of resonant signals, it in general can be viewed as a primary measurement of intensity for specific formants, as well as exhibits the relative strength relation among formants.

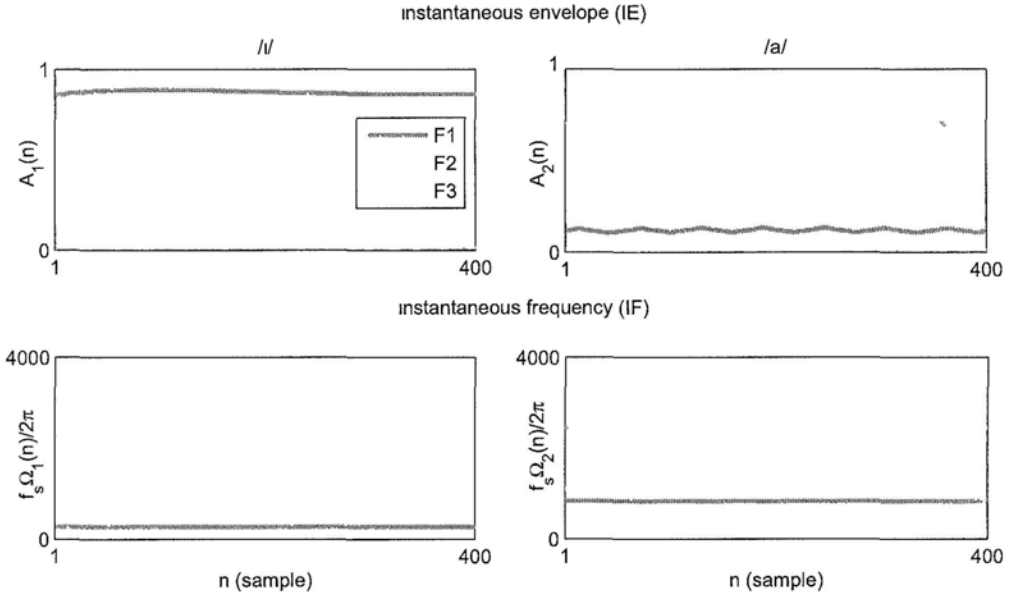


Figure 5.6: *Instantaneous envelopes and frequencies of the resonant signals.*

5.2 Representing Phase Information by Instantaneous Frequencies

Considering the difficulty of extracting useful information from the phase spectrum of speech, an alternative representation for phase is explored for speaker recognition purpose. In this section, we are inspired to employ the set of frequencies that dominate in the primary components as a representation of speech phase.

5.2.1 Identifying primary speech components

In the source-filter speech production model, a periodic impulse sequence is filtered with a glottal filter to produce the periodic part of the excitation signal, while the non-periodicity and the turbulent content are obtained by adding an additional white noise source. The resulting signal excites the vocal tract system, which is characterized by its formant structure. As pointed out by O'Shaughnessy that the spectral formant and harmonic structure models only spectral behavior of these speech properties [154], the spectro-temporal characteristics of the primary speech elements should be addressed instead.

A typical formant in the vocal tract system is formulated as a mono-component AM-FM term $A(n)cos[\Theta(n)]$, as elaborated in the earlier section. Besides the principal formants, the formant structure possesses a number of other components, for instance, the spread of spectral envelope and transitions between formants, etc. A pitch harmonic can also be interpreted as an AM-FM component. However, besides the formants, the pitch and its principal harmonics, there are other primary components that result from the interferences among all different harmonics and from the interactions within the vocal tract. Speaker-discriminative properties of a speech signal is therefore being possessed by these components. A speech signal can thus be written as a linear combination of amplitude and frequency modulated components which we called the

primary components,

$$\begin{aligned} s(n) &= \sum_{k=1}^{K} A_k(n) cos\big[\Theta_k(n)\big] + \eta(n) \\ &= \sum_{k=1}^{K} A_k(n) cos\Big\{\Big[\Omega_c(k)n + \sum_{r=1}^{n} q_k(r)\Big]\Big\} + \eta(n), \end{aligned} \tag{5.13}$$

where $A_k(n)$ denotes the instantaneous amplitude of the kth primary component and $\Theta_k(n)$ denotes its instantaneous phase. With the backward difference between $\Theta_k(n)$ and $\Theta_k(n-1)$, the instantaneous frequency (IF) sequence is defined as $\Omega_k(n) = \Omega_c(k) + q_k(n) = \frac{2\pi}{f_s} f_c(k) + q_k(n)$, where f_s is the sampling frequency, $q_k(n)$ is the frequency modulation (FM) component. $\eta(n)$ takes into account additive noise and errors of modeling, especially errors due to the finite summation. This model has also been described as the elementary waveform speech model [130].

Depending on the application, the number of primary components required for processing may vary. For coding purposes, synthesis-by-analysis coders based on the sinusoidal representation use a fairly large number of primary components, even for unvoiced sounds. But in the representation of vocal properties of speaker, the relevant components are usually identified with the formants and pitch harmonics.

In Figure 5.7, the primary AM-FM components extracted from multi-bands of a speech segment as well as the corresponding carriers are illustrated. For a clear presentation, the carriers are displayed in dash, and the tracked AM-FM components are in real line. Each arrow in the plane expresses an AM-FM element, where its length stands for the mean value of segmental amplitude and the horizontal distance from origin indicates the averaged frequency quantity. The carriers are of unit amplitude for all frequency bands, while their frequencies are delivered by $\Omega_c(k)$, $k = 1, 2, \ldots, K$. When taking the kth subband for a close inspection, a space is discerned between the carrier and the detected center, it is therein noted as the FM component that measures the deviation of dominant frequency from carrier.

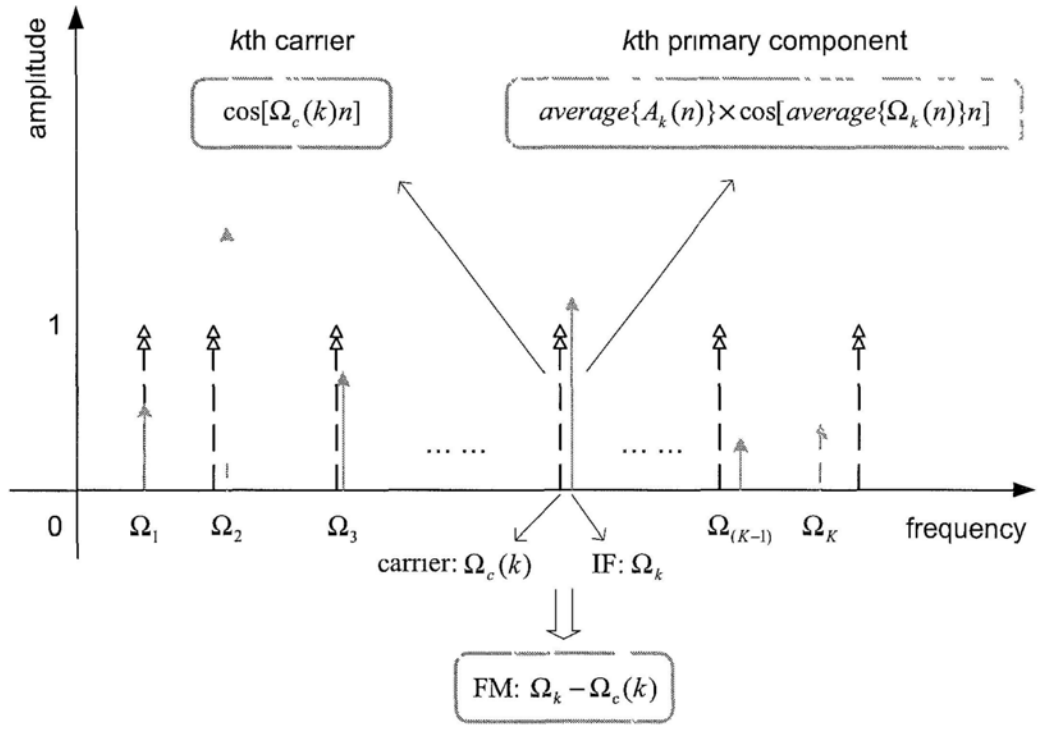


Figure 5.7: *A pictorial illustration for primary speech components under AM-FM framework.*

5.2.2 Representing frequencies present in speech

The phase of the primary components in a speech signal may not be well-captured by its short-time phase spectra. The instantaneous frequency quantities that derived from the multi-band amplitude and frequency demodulation, can essentially get hold of the phase variations in a speech signal, which inspires the usage of the set of dominant instantaneous frequencies as the phase representation of speech.

Let us examine the instantaneous frequency (IF) component in capturing the dominant frequency present in an AM-FM signal. Taking inspection on the synthetic vowels mentioned in Section 5.1.3 at first. This is conducted by means of analysis-by-synthesis method, where we initially determine a few formant-descriptive resonant signals with predefined centers and bandwidths, and then pick out the underlying frequencies included therein to testify the model. Table 5.2 lists the results acquired in the course of processing applied on the synthetic vowels.

Table 5.2: *Formant frequency estimation results for synthetic vowels /i/, /a/, and /u/: symbols A, B, C stands for the formant frequency, frequency estimate and the estimation error rate, respectively.*

		$F1$	$F2$	$F3$
/ɪ/	A	270 Hz	2290 Hz	3010 Hz
	B	281 Hz	2212 Hz	3014 Hz
	C	4 1 %	3 4 %	0 1 %
/a/	A	730 Hz	1090 Hz	2440 Hz
	B	830 Hz	1099 Hz	2430 Hz
	C	13 7 %	0 8 %	0 4 %
/u/	A	300 Hz	870 Hz	2240 Hz
	B	278 Hz	828 Hz	2157 Hz
	C	7 3 %	4 8 %	3 7 %

It is found that the IF in most scenarios referring to might be well taken as an estimate of the center frequency present. Capability possessed by the IF in

identifying primary frequency components owned by an AM-FM speech signal is therefore exhibited.

To clearly investigate the implicit rationale behind representing speech phase by sequences of frequencies in a multi-band speech processing approach, we then move the experiments forward to real speech data. Considering that a speech signal always contains quite a few primary frequency components that spanning over its bandwidth, we apply a bank of Gammatone filters on the test data first.

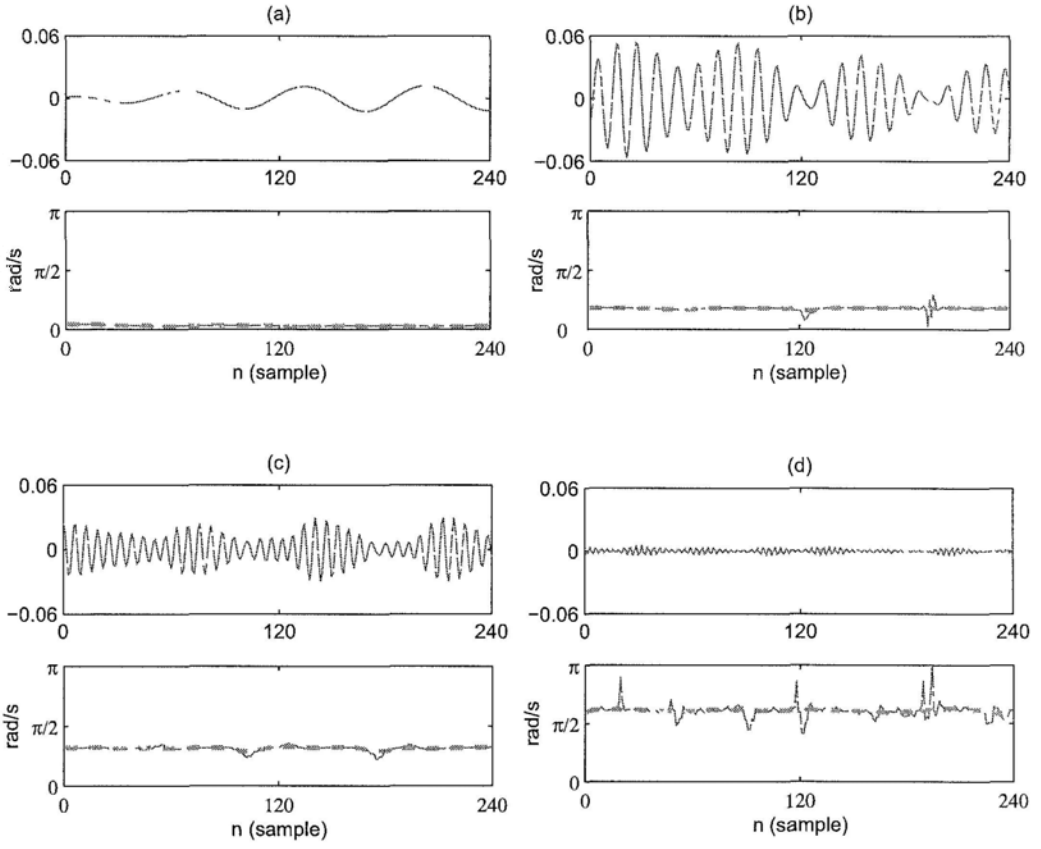


Figure 5.8: *Subband signals and their instantaneous frequency sequences from a male speaker. The subbands are of ERB bandwidth, with their center frequencies are:* (a). 0.03π; (b). 0.18π; (c). 0.31π; (d). 0.61π. *(π is the Nyquist frequency)*

In Figure 5.8, a voiced speech segment is inspected in four frequency bands. In each band, the band-passed signal and its IF sequences are shown in the upper and lower row, respectively. Two sets of IF sequences are presented for

each subband, which are marked by a solid line and a dashed line, respectively. As previous studies [155] found that the abrupt impulses presence along the time line should be avoided in examining the dominant frequency component, we have therefore obtained the dashed line as a smoothed version of the solid one by extrapolation.

Distributions of the IF sequence in 40 Equivalent Rectangular Bandwidth (ERB) bands across the frequency range [80 Hz, 4000 Hz] have been worked out. It is found that in a band which captures either the pitch harmonic or formant, the standard deviation of the IF from its mean value is pretty small (e.g., 0.03π for (d) being the highest among those referred to in Figure 5.8). Thus, the mean value of the smoothed IF sequence within a short time interval is of high confidence to be used as the frequency estimate in a specific subband.

5.3 Phase-related Modulation Parameters

Short-time Fourier analysis provides important magnitude characteristics for speech, however, since it cannot effectively capture the phase variation in speech, it is hard to quantify the phase spectrum and distinguish useful components contained therein. Previous studies on speech synthesis and coding have employed the multi-band demodulation framework to analyze and quantify speech components in terms of instantaneous amplitude and frequency [130], [131]. Estimation of the primary formants and pitch tracking also involved the multi-band decomposition of speech signal in the frequency domain [145], [149]. Frequency modulation related parameters of speech signal have been used for identifying phonemes and speakers [55], [57], [144], [156]. These researches inspire the exploration of proper representation for phase information in speech signals.

5.3.1 Instantaneous frequency-based features

The phase-related parameter set that derived by using the multi-band demodulation method is noted as Averaged Instantaneous Frequency of Speech (SAIF). The process of extracting the SAIF parameters is illustrated in Figure 5.9 and summarized as follows.

1. *Voicing decision*: The SAIF features are extracted from voiced speech only. The voicing status is detected using Talkin's Robust Algorithm for Pitch Tracking [146].

2. *Filter bank filtering*: Applying a bank of K Gamma-tone filters on the voiced signal $v(n)$ to produce the subband signals. The center frequency $f_c(k)$ ranges from 4 kHz to 80Hz with a k increase from 1 to K.

3. *Multi-band demodulation*: Teager's energy separation algorithm is employed in obtaining the instantaneous angular frequency $\frac{2\pi}{f_s} f(n)$ (IF) on a frame basis for each subband signal.

4. *Smoothing of the IF sequence*: A 21-point median filter is applied to remove the abrupt impulses in the frames of IF sequence.

5. *Frame averaging of the smoothed IF*: An averaging operation is done on the smoothed IF sequence for the frames in each subband. In this step, we remove the fluctuations of the IF sequences, and track the most significant frequency component in each subband frame by frame.

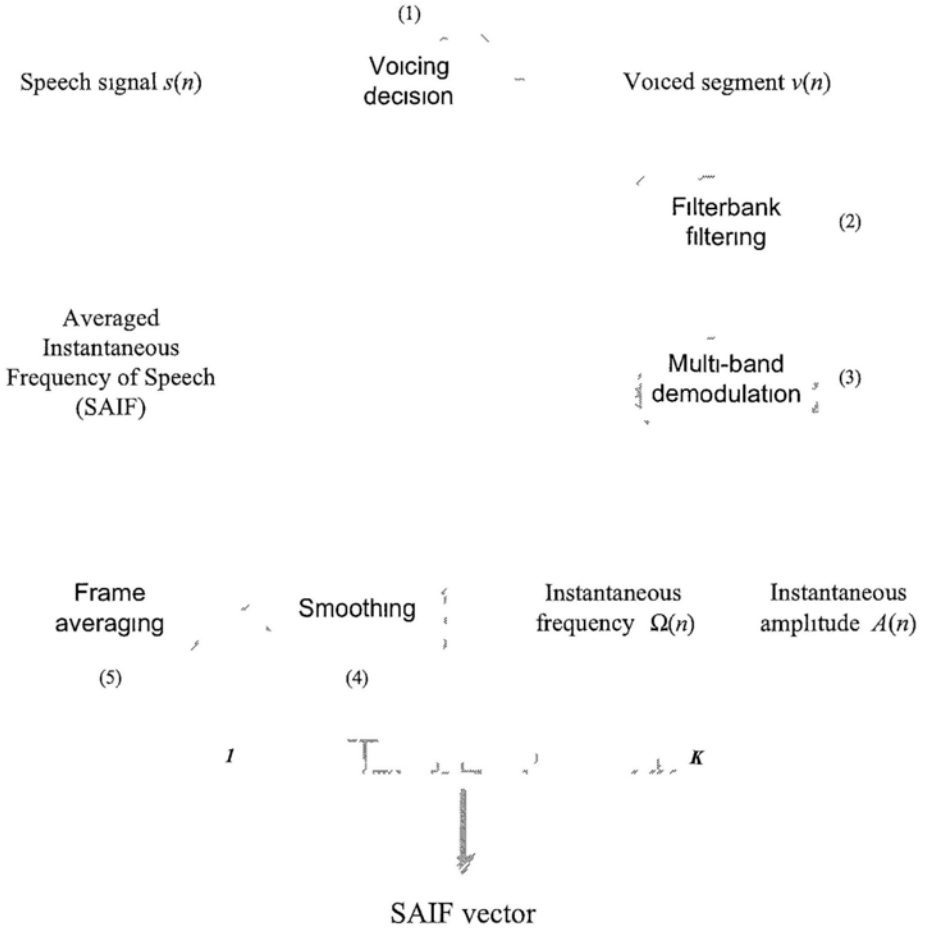


Figure 5.9: *Block diagram for the extraction of SAIF features.*

5.3.2 Feature analysis

Vocal properties that distinguish different speakers are essentially carried by the interrelated formants and harmonic structure. In this part, we will show that the SAIF parameter set is able to capture these typical speech properties.

- **Formants, harmonics and their interactions**

The major primary components of a voiced sound include the pitch component and some of its first harmonics. However, they are ignored by most of the classical algorithms which pre-emphasize the signal. The other primary components encompass naturally the formants of a speech signal. In this experiment, we use a vowel synthesized at 8 kHz sampling frequency with the specifications listed hereunder:

- F0 = 172.4 Hz = $f_c(36)$
- F1 = 773.8 Hz = $f_c(22)$, F2 = 1161.8 Hz = $f_c(17)$, F3 = 2446.2 Hz = $f_c(7)$

Moreover, the F0 value of the synthetic signals and center frequencies of some band-pass filters are settled specially, as shown below:

$f_c(36) = F0 = 172.4Hz$	$f_c(30) = 2.1F0$	$f_c(26) = 3.2F0$	$f_c(23) = 4.1F0$

Figure 5.10 shows the primary components in the synthetic vowel $s_1(n)$ in terms of averaged amplitude and FM values in the upper and lower rows, respectively. To help identifying those excitation-related components captured, corresponding amplitude and frequency values from the excitation signal are also shown. When k decreases, the subband center frequency $f_c(k)$ increases. It is observed that with smaller $f_c(k)$, the primary components mainly capture the first several harmonics of F0, which can be evidently seen from the small peaks in the amplitude and valleys in FM around the 36th, 30th and 26th subbands. This observation is consistent with what we have reported in [61]

that the subband which exactly gets hold of a pitch harmonic will result in a peak amplitude and a valley-bottom FM value among the others nearby. With the center frequency increasing, firstly, the higher order F0 harmonics are coexistent with the emerging formants, for instance, the 23rd and 22nd subbands contain the F1 and the 4th F0 harmonic simultaneously. Gradually, the formants begin to dominate, which can be seen from the prominent peaks in amplitude around the 17th and 7th subbands. Likewise, when the dominant frequency in a subband exactly matches a formant, its FM tends to go to zero. Therefore, we can see that the primary frequency components in speech can track not only the phase variation among frequencies, which is absent from the power spectrum, but can also give rise the essential vocal characteristics conveyed by the excitation source.

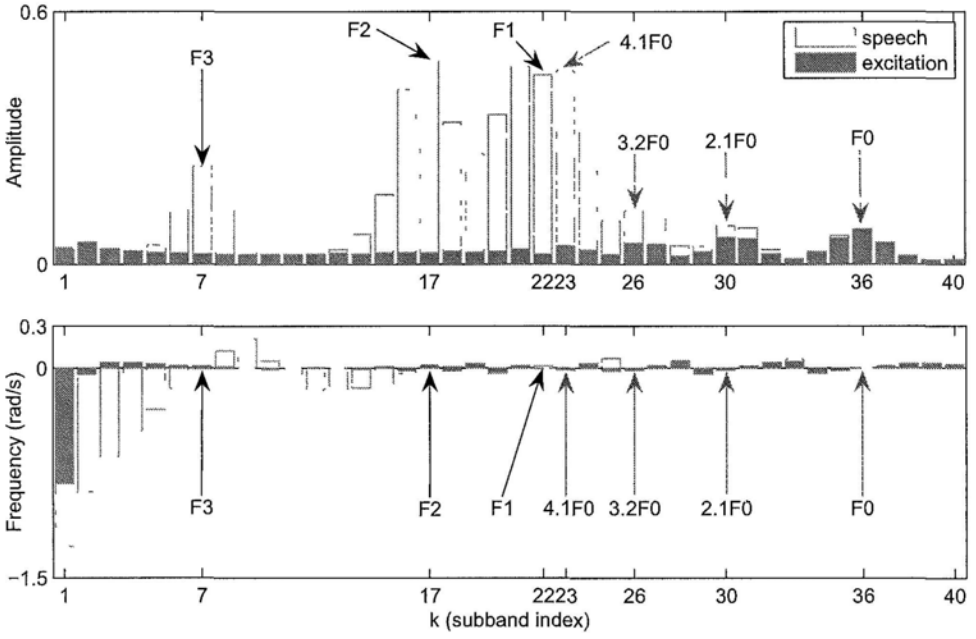


Figure 5.10: *Amplitude and frequency of primary speech components: formants, harmonics and their interactions.*

- **Formant bandwidth effect**

As indicated by the classical synthesis model that the vocal characteristics of a formant depend not only on its central frequency but on its bandwidth as well. It is almost without exception that the formants can characterize fairly well different voiced phones. But when uttering a similar voiced sound, speakers may also differ in formant bandwidth. In Figure 5.11, the SAIF parameter sets of speech signals $s_1(n)$ and $s_2(n)$ are displayed. The specifications of $s_2(n)$ are the same with that of $s_1(n)$ except the formant bandwidth, which is 10 Hz for $s_1(n)$ and 200 Hz for $s_2(n)$. To be focused, we take F3 for further inspection. It is clearly revealed that the transition around the formant frequency, e.g., 7th subband is quite different for $s_1(n)$ and $s_2(n)$. The stronger formant that resulted from the narrower bandwidth in $s_1(n)$ exhibits greater influence on its neighbors. These particularities of formants that determine the formant structure are legibly retrieved from the derived parameter set.

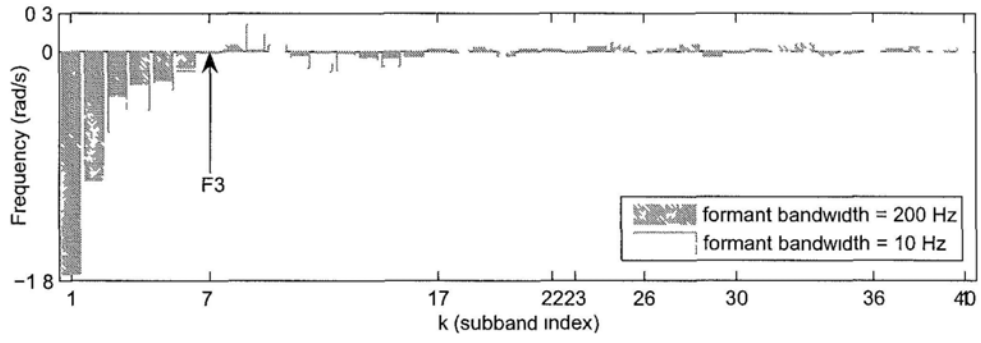


Figure 5.11: *Frequency of primary speech components: formant bandwidth effect.*

5.4 Performance of Phase Information for Discriminating Speakers

In this section, we will evaluate the proposed feature SAIF as the complementary parameters to MFCC features through speaker identification (SID) and speaker verification (SV) experiments on a speech database CU2C. The standard MFCC features we used contain 39 components: the log energy, 12 static coefficients and their dynamic and acceleration coefficients.

5.4.1 Experimental set-up

For all speakers, 6 out of the 18 sessions are used to train the speaker models. For the remaining data, 6 sessions are used as development data, while the last 6 sessions are employed in performance evaluation. The standard approach for UBM-GMM training is adopted. Two separate systems are built based on MFCC and SAIF, respectively. Score-level fusion technique is used to combine the contributions of the two systems and produce the final decision. For the identification tasks, the log-likelihood score of each test is the linear combination of the log-likelihood score L_M from MFCC and L_S from the SAIF features, i.e., $L = w_M L_M + w_S L_S$, where w_M and w_S are the weights on MFCC and SAIF, respectively. Meanwhile, in the verification tasks, the fusion is performed on the log-likelihood ratio scores λ_M and λ_S, i.e., $\lambda = w_M \lambda_M + w_S \lambda_S$. In both tasks, w_M and w_S are related by $w_M + w_S = 1$. The optimal values of w_M and w_S are determined such that they achieve the best identification/verification performances for the development data. This is done by exhaustive search with a step size of 0.02, over the interval of $[0, 1]$.

For SID and SV tasks, the identification error rate (IDER) and equal error rate (EER) are used as the primary performance indicators, respectively.

5.4.2 Experimental results

The proposed phase-related parameter set SAIF is about to be examined from a few perspectives based on their performance in dependant experiments.

♦ **Benchmark results**

Performance of the MFCC and SAIF parameter sets in clean environment under the protocols of speaker identification and verification are shown in Table 5.3, individually. Their results are therefore used in the baseline system to provide the benchmark record, where the MFCC features are of 39-dimension and the SAIF feature vectors containing 40 parameters.

Table 5.3: *Speaker recognition performance of individual MFCC, SAIF features under clean environment: IDER & EER (in %).*

feature configuration	IDER	EER
MFCC	**2.44**	**1.52**
SAIF_*20*	6.33	3.72
SAIF_*40*	**4.78**	**2.70**

♦ **Effects of frame length**

The SAIF feature vectors are extracted from fix-length frames of speech signals. In Table 5.4, the recognition performance of the SAIF features that extracted with doubled and tripled size of frame are evaluated.

Table 5.4: *Effects of frame length on speaker recognition performance: IDER & EER (in %).*

feature configuration	IDER	EER
SAIF_*bil_40*	6.39	3.64
SAIF_*tril_40*	9.22	4.54

- **Supplementing cepstral features MFCC**

The fusion of MFCC features with either SAIF_*40*, SAIF_*bil_40* or SAIF_*tril_40* parameters leads to improved performance over the MFCC-alone speaker recognition system Table 5 5 gives the results of the above three sets of phase-related parameters after combining with MFCC For reference, the averaged weighting parameter sets $[w_M, w_S]$ in getting these results are given in bottom row of the table

Table 5 5 *Performance by fusing SAIF features with MFCC IDER & EER (in %)*

feature combination	IDER	EER
MFCC+SAIF_*40*	1 83	1 16
MFCC+SAIF_*bil_40*	1 89	1 21
MFCC+SAIF_*tril_40*	1 78	1 33
$[w_M \quad w_S]$	[0 63 0 37]	[0 69 0 31]

When there is white noise present in the speech data, the SAIF parameters manifest similar recognition accuracy with the MFCC features in both SID and SV experiments For example, with SNR = 10 dB training/test data, the EERs of SAIF_*40* and MFCC features are 3 99% and 4 03%, the IDERs are 8 00% and 8 33%, respectively Feature combination in this scenario yields 2 61% EER and 5 44% IDER, where the improvements over MFCC are 35 24% and 34 69% for SV and SID, respectively

- **Combining with source-related characteristics**

It is interesting to investigate the complementary effects between the SAIF_*40* features and the source-related parameters RAIE_*40*, RAIF_*40* in discriminating different speakers The feature vectors RAIE_*40* and RAIF_*40* were proposed and evaluated in the previous chapter Individual performance of RAIE_*40* and RAIF_*40* have already been reported in Table 4 5, however, they

are given below in Table 5.6 again for the purpose of convenient reading. Averaged weighting parameter set $[w_R, w_S]$ in getting these results are given in bottom row of the table as well.

Table 5.6: *Performance by fusing SAIF features with source feature sets RAIE, RAIF: IDER & EER (in %).*

feature configuration/combination	IDER	EER
RAIE_*40*	27.28	9.46
RAIF_*40*	19.67	8.01
RAIE_*40*+SAIF_*40*	5.33	2.69
RAIF_*40*+SAIF_*40*	4.61	2.65
$[w_R : w_S]$	[0.56 : 0.44]	[0.25 : 0.75]

5.4.3 Analysis of results

♦ **Recognition performance**

In Table 5.3, it is found that the SAIF features perform well in both SID and SV tests. The best performed SAIF parameter set that listed in Tables 5.3 and 5.4, i.e., SAIF_*40*, is of comparable dimension with MFCC, while it provides phase characteristics for individual speaker that are independent to the magnitude information carried by MFCC. With a simple linear combination method, 25.00% and 23.68% relative reductions in IDER and EER have been achieved, respectively. Under additive noise, the SAIF feature set is found to show comparable or even higher robustness than MFCC. Besides, combination of the two information sources under both scenarios can offer noticeable improvements.

♦ **SAIF features: in scrutiny**

SAIF parameter set quantifies frequencies of the relevant primary components within a time interval of speech signal, where it depends on two factors: the number of primary components involved and the frame duration for parameter estimation. For our data, the primary SAIF_*40* feature set involves

40 subbands and 30 *msec* frame length. To observe the effects of the two factors on the usability of the SAIF parameters in delivering phase characteristics of speech signal, we further derive three sets of SAIF parameters: SAIF_*20*, where the density of the Gammatone filters is reduced by half; SAIF_*bil_40* and SAIF_*tril_40*, where the frame length is doubled (i.e., 60 *msec*) and tripled (i.e., 90 *msec*), respectively. It is indicated that the SAIF_*40* can take account most of the principal frequencies present, and these frequencies are relatively stable within a duration of 30-60 *msec*. This implies that the phase characteristics quantified in this manner possess reliable speaker-discriminative power. The effects of frame length on the global performance of MFCC-SAIF combinations may be reflected by Figure 5.12, where the relative reductions therein produced based on the MFCC baseline in SID and SV tests are numerically marked.

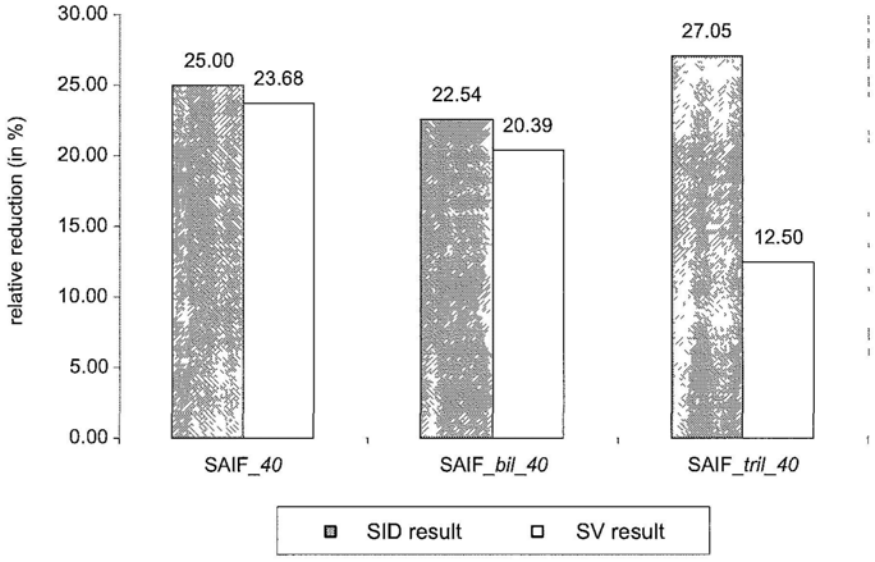


Figure 5.12: *Effects of frame length on experimental results.*

♦ Complementary effects in fusion

Since the SAIF features are carrying phase-based information of speech signals, to exert its potential in distinguishing speakers, sources of speech information that are of complementary effects with SAIF are necessary. The underlying

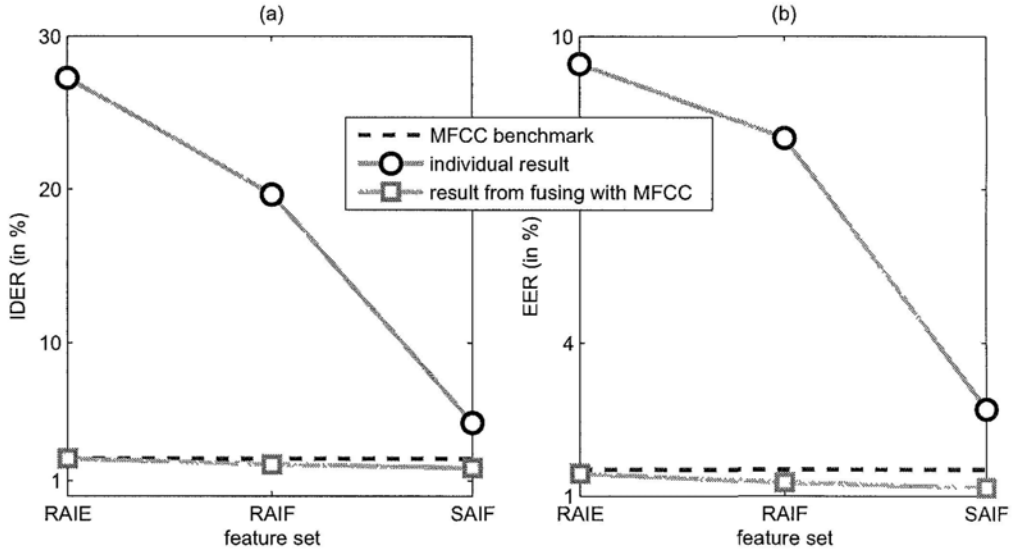


Figure 5.13: *Results of SAIF, RAIE and RAIF features when employed individually and combining with MFCC for : (a) SID experiments; (b). SV experiments.*

complementary effects exist in SAIF's fusion with the magnitude-specific MFCC features are found pretty good, as demonstrated through experimental results in Table 5.5 and illustrations in Figure 5.12, respectively. Even so, it is intriguing to inspect the supplement RAIE, RAIF parameters, as another important information source in speech, could offer to SAIF. By referring to the benchmark and relevant results recorded in Tables 5.3, 5.5 and 5.6, it is noted that the SAIF-RAIF fusion can offer some assistance to improve the SAIF-only results, but the support received herein is much smaller than that from fusing with MFCC. While, fusion SAIF-RAIE is revealed ineffective in this case. Figure 5.13 gives a comparison on performance of the three sets of modulation parameters RAIE, RAIF, and SAIF when working individually or together with MFCC, respectively.

From another point of view, put aside the assistance from MFCCs, SAIF could produce better results in comparison with the other two sets of modulation parameters. Furthermore, the combination of SAIF and RAIF provides the optimal performance we could get from the application of AM-FM framework

on speech signals.

♦ SAIF *vs.* RAIF

When inspecting the effectiveness of the AIF parameter set derived from the residual signal or speech signal, i.e., the RAIF and SAIF features, it is found that the SAIF outperforms the other part, as indicated by Figure 5.13. Although under quite similar flow path in generation, the phase-related parameter set that capturing primary frequencies present in speech signals apparently show higher discriminative power in discriminating different speakers. Taking them individually first, it is possible that a part of harmonics that conveyed by the RAIF set is also captured by the SAIF vectors, where the latter then take advantage of both vocal excitation- and tract-related attributes and win. Regarding the complementary employment with MFCC, the SAIF features are obviously of greater predominance than RAIF. Finally, the usefulness of effectively extracted phase characteristics of speech is confirmed.

5.5 Summary

Magnitude feature-oriented front-ends provide information for understanding the speech content, but neglect the potential of using phase information as speaker representatives. Considering the difficulty of extracting useful information from the phase spectrum of speech, an alternative representation for phase is needed for speaker recognition purpose. In this chapter, we are inspired to employ the multi-band instantaneous frequency quantities as the speaker-discriminative parameters. Through analytical comparisons and simulation results, it is revealed that the proposed features are capable of: (1). capturing the formant and pitch related vocal differences among speakers; and (2). offering consistent complementary assistance in recognizing speakers to spectral features under both clean and additive noise scenarios.

Chapter 6

Performance Evaluation on the Robustness of Modulation Speaker Features

Phase information provide useful acoustic cue for identifying individual speakers. Speaker verification employing the instantaneous phase-related features that perform well in clean or matched noise/channel conditions degrades dramatically when encounter unexpected communication environments. These adverse effects can distort the short-term distributions of the speaker parameters. It is observed that by mapping each feature stream to a target distribution over a specific time interval, their robustness to environmental or channel mismatch is enhanced. Through speaker verification experiments on microphone and telephone data, it is observed that the proposed robust feature extraction front-end consistently reduces the equal error rate.

In this chapter, with attempt to characterize and discriminate different speakers in various environmental and communication scenarios, we first study the additive and convolutive noise effects on the phase-related SAIF parameter set in Section 6.1. The mechanism of feature mapping method, its implementation flowchart, and new perspectives needed in order to treat the modulation parameters are introduced in Section 6.2. Enhanced speaker features from distribution mapping are evaluated subsequently in terms of their speaker verifi-

cation performance under mismatched conditions in Section 6.3. Section 6.4 summarizes this chapter.

6.1 Environmental Effects on Phase-related Parameters

It is always expected that a pattern classification system like speaker verification system, can consistently produce accurate results no matter operate under what conditions in the world. Nevertheless, speaker representatives such as the cepstral coefficients have been revealed to perform unreliably in presence of noise or unexpected channels. It is highly desirable if the phase-related parameters which named averaged instantaneous frequency of speech (SAIF) could offer some assistance to compensate for the deficiency. In attempt to work out an efficient solution under the circumstances, it is essential to investigate the undesirable effects endured by speech parameters that may cause by the mismatched training-test scenarios.

The effects resulted from various environmental variations such as noises and channel variability that degrade the cepstral features, have been discussed extensively with regard to speech/speaker recognition and language identification, for instance, in [38], [157], [158]. We have made a study on their robustness to additive noises for the purpose of speaker recognition in a previous work [103]. Therefore in this section, we will focus on observing and analyzing the alteration of statistical characteristics for the SAIF parameters regarding specific noise effects.

6.1.1 Speaker distinction under adverse conditions

In the previous chapter, the set of phase-related SAIF parameters are found useful in complementing the MFCC features with clean microphone data in speaker recognition tasks. Our pilot study on speech data corrupted by additive noise and those over a telephone network reveals that this set of SAIF parameters can still help enhancing the MFCC-based speaker verification performance, if matched training and test data are used. Detailed results can be found in Figure 6.1, where the individual performance of SAIF features under both clean and matched channel/noise training-test conditions as well as its combination with

MFCC parameters in speaker verification task are demonstrated.

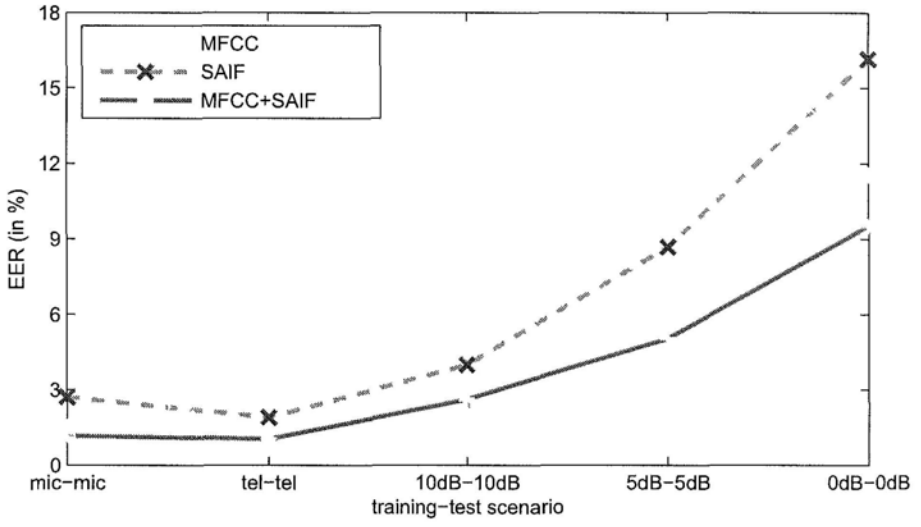


Figure 6.1: *Speaker verification performance under clean and matched channel/noise conditions: EER (in %).*

Table 6.1: *Speaker verification performance under mismatched noise/channel conditions: EER (in %).*

Training data	Test data	MFCC	SAIF_40	MFCC + SAIF_40	$[w_M : w_S]$
Mic.: clean	Tel.	18.94	24.67	18.35	[0.72 : 0.28]
Mic.: clean	Mic.: 10dB	17.41	23.47	15.93	[0.62 : 0.38]

However, performance of the speaker recognizers using these acoustic features degrades dramatically in presence of mismatched conditions, e.g., an unexpected transmission channel or environmental noise. Take those shown in Table 6.1 for instance, where the results of MFCC, SAIF features that produced under two conditions are shown. The clean microphone data is used for training in all scenarios. The test data in the noise mismatch condition are with Gaussian noise at 10dB SNR level, while telephone data are used for channel mismatch

tests. It is seen that the MFCC and SAIF features are severely deteriorated, their joint contribution yet cannot offer adequate help to improve the conditions. Therefore, it is essentially needed to exploit speech parameters that are speaker-specific and robust to noise, channel or transducer effects for speaker verification systems operating in actual applications.

6.1.2 Observations on noise contamination

The pertinent phase representation of a speech signal as we built in an earlier stage, gets hold of the dominant frequencies of primary AM-FM speech components. In order to make clear the noise effects on concerned speech parameters, we make an observation on the frequency components owned by the contaminated speech segment as well as the noises.

Figure 6.2 illustrates additive noise corruption on the amplitude and frequency modulation parameters of a speech segment. The corruption by Gaussian noise is at two different SNR levels, which are separately shown in Figure 6.2(a) and (b). Columns in Figure 6.2(a) and (b) each, from left to right, corresponds to the segment of clean speech $s(n)$, noisy speech $y(n)$ and the contaminating noise $d(n)$, respectively. For a specific segment, the waveform, averaged instantaneous envelope (AIE) parameters, and frequency modulation (FM) parameters across subbands are displayed in the three rows from the top down. The concerned characteristics of AIE and FM components from speech signals have been discussed in Chapter 5. In the rightmost columns of Figure 6.2(a) and (b), as for the corrupting noises, it is found that their energy distribution over the subbands is more or less alike their spectra in the frequency domain, which is widespread among frequencies yet rich in high frequency components. The FM components of noises tend to be close to zero for most bands. These distinct attributes of noises render transformations on the parameters of speech, which present in the central columns accordingly. Therefore, the AIE from the noisy speech segment is revealed visibly different from the clean one, as well as the FM parameters. It is found that the AIE component in bands

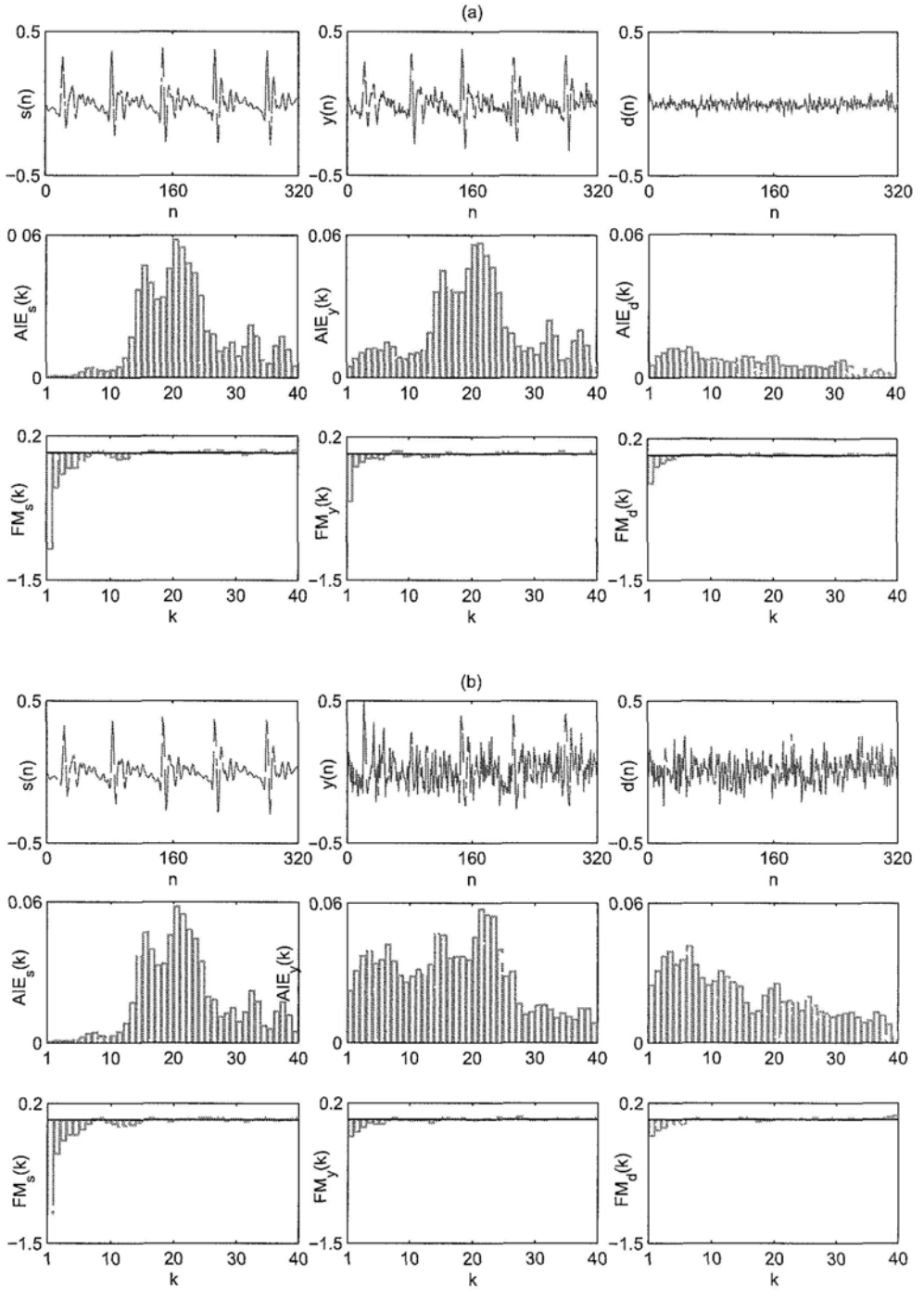


Figure 6.2: *Speech segment, averaged instantaneous amplitude and frequency quantities from a male speaker under additive Gaussian noises: (a)* $SNR = 10dB$ *and (b)* $SNR = 0dB$.

that affected is capturing content from both speech and noise, which resultantly increase its amplitude; while for the case of frequency modulation, the noise in many bands generally tends to diminish the distance between the dominant frequency and the carrier, which results in decreased FM components. On the other side, when the contamination getting worse, like moving from case in Figure 6.2(a) to case in Figure 6.2(b), the noise effects become heavier, but basically in similar manners.

It is difficult to quantify the amplitude-frequency modulation under noise conditions strictly, however, considering the short-term distributions of related parameters, the remaining parts in this section further observes the concerned effects by noises.

6.1.3 Additive noise effects

In Chapter 5, a primary component of a speech signal is represented as an AM-FM signal $A_k(n)cos\{[\Omega_c(k)n + \sum_{r=1}^{n} q_k(r)]\}$ (Equation 5.13). In addition, we found that the frequency of a primary component in a voiced speech signal is relatively stable for a short time interval, thus, we use an estimate of the dominant instantaneous frequency in a specific subband as the phase-related quantity, from which the SAIF feature vector is constructed. Accordingly, for the kth primary speech component $s_k(n)$, where the dominant frequency is Ω_k^s, i.e., the kth parameter in the corresponding SAIF vector, the bandwidth, spectral envelope of $s_k(n)$ are all determined by its amplitude sequence $A_k^s(n)$. This founding leads us to scrutinize the noise effects on the primary frequency components of a speech signal by viewing them as AM signals with fixed carriers, i.e.,

$$s_k(n) = A_k^s(n)cos(\Omega_k^s n). \tag{6.1}$$

In Figure 6.3, a drawing is given to convey this idea.

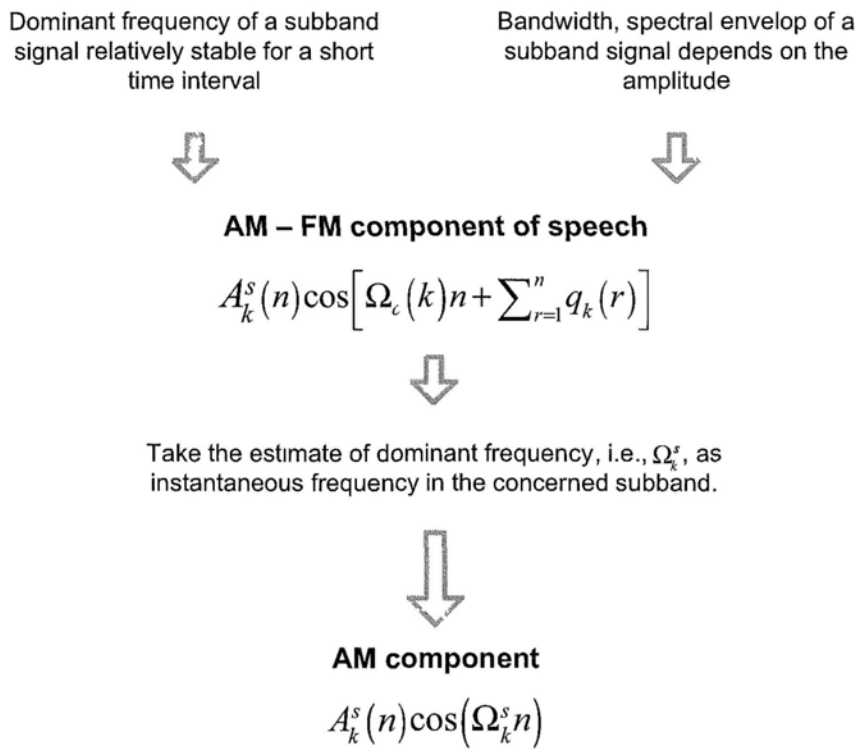


Figure 6.3: *An illustrative AM representation for an AM-FM speech component.*

In speech coding, usually the summation of a large number of AM-FM components are used to model the portion of unvoiced speech, because it is considered that a noise-like signal consist of numerous narrow-band frequency components. Similarly, under additive noise $d(n)$, for a particular speech component with amplitude sequence $A_k^s(n)$ and frequency estimate Ω_k^s, usually there are several affecting noise bands. Figure 6.4 visualizes the general effects of additive noise on a speech component. It is seen that most of the narrow-band noise components deviate from the carrier Ω_k^s. Thus, all the amplitudes and frequency quantities in the noise signal will affect the measurement of the dominant frequency of a speech component. These slowly-varying noise components around Ω_k^s generally have the effect of reducing the variance and distorting the distribution of the kth SAIF parameter.

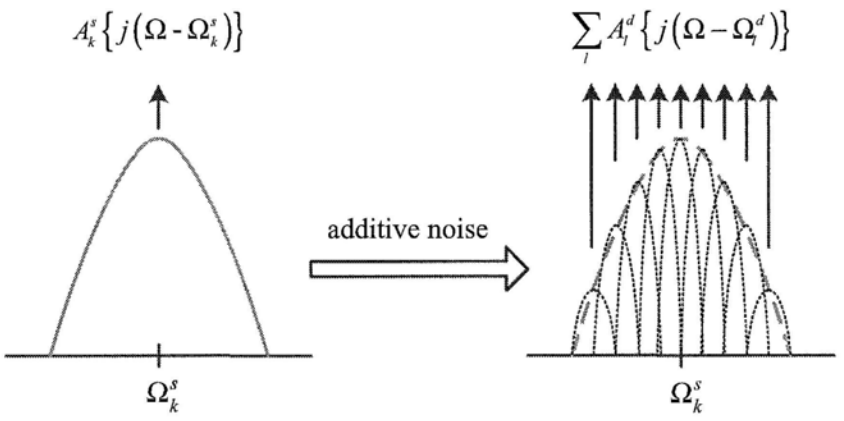


Figure 6.4: *Additive noise effects on kth frequency component of a speech signal. ($A_k^s(j\Omega)$ is the Fourier transform of the amplitude sequence in the kth subband of speech, $A_l^d(j\Omega)$ is the Fourier transform of the amplitude sequence of one narrow-band noise signal centered at Ω_l^d)*

6.1.4 Convolutive noise effects

In telephone network, the speech signal $s(n)$ is transmitted through a channel with impulse response $c(n)$. Following the formulation of Equation 6.1, if we assume $c_k(n) = A_k^c \delta(n - n_k^c)$, the transmitted AM signal $s_k(n)$ is noted by $z_k(n)$ as in Equation 6.2,

$$\begin{aligned} z_k(n) &= s_k(n) \otimes c_k(n) \\ &= \left\{ A_k^s(n) cos(\Omega_k^s n) \right\} \otimes \left\{ A_k^c \delta(n - n_k^c) \right\} \\ &= \left\{ A_k^c A_k^s (n - n_k^c) \right\} \cdot \left\{ cos(\Omega_k^s n - \Omega_k^s n_k^c) \right\}. \end{aligned} \tag{6.2}$$

It can be found that the assumed channel response generally has the effect of changing only the amplitude of a speech component. However, in real communication systems, where the channel models are much more complicated than that we assumed, and there are usually some forms of additive noise exist, thus, the estimation of the instantaneous dominant frequency Ω_k^s might be affected inevitably.

6.2 Robust Feature Extraction

There are a number of methods have been proposed to deal with either the additive noise effects in the spectral domain or to compensate for the linear channel effects in the cepstral domain. These methods have shown their effectiveness in enhancing the discrimination power of the cepstral features in speech recognition applications. Among them, cepstral mean subtraction (CMS) [9] is found to be the most popular one in suppressing the linear channel effects. Modulation spectrum analysis [159] aimed to determine the relative importance of the spectral components under mismatched channel conditions in speaker verification system, however, with limited robustness to additive noise. Distribution normalization of single cepstral features over a short interval by removing its mean and scaling its standard deviation has been proposed as an extension to the CMS method in speaker detection tasks [160], and robust speech recognition applications [97]. Pelecanos *et al* in [106] proposed a novel feature mapping approach that had demonstrated improvements for speaker verification over a number of enhancement methods such as CMS, modulation spectrum processing, short-term windowed CMS and variance normalization. This feature mapping method warps the distribution of each cepstral feature stream to a standard normal distribution over a specified time interval.

Although the phase-related features SAIF are proved useful in complementing the magnitude-based MFCC features in clean and matched noise conditions, their performance is inadequate in mismatched conditions. Feature enhancement methods that developed for cepstral features are obviously not fit for SAIF, and there are very little efforts have been made to understand the undesirable noise or channel effects on phase-related parameters. Considering the difficulty of detecting and quantifying the noise/channel distortions for phase parameters, it is ideal if an enhancement approach requiring no precise distortion model can be applied to SAIF features. Through looking into the distributions of individual SAIF parameters, it is found that their uni-modal nature makes feature mapping a proper post-processing for them, in order for improved immunity to environmental or transmission variations. We are therefore inspired

to investigate the robustness of this approach across different recording/channel environments other than what the system is developed

6.2.1 Mechanism of feature mapping

The feature mapping method aims to construct a more robust representation of each feature distribution It is achieved by conditioning and conforming the individual feature streams such that they follow a specific target distribution over a window of speech frames In [106], the observed cepstral feature distribution over a specified speech interval is mapped so that the accumulated distribution is similar to a target distribution The procedures in this process are summarized as follows

1 *Frame windowing* a sliding window is used to isolate N frames of speech features where the nth frame that in the window center is the current frame Step 2) to Step 4) repeat each time the sliding window shifts by 1 frame sequentially

2 *Feature sorting* for each windowed stream from a K dimensional feature set, the feature values are sorted descendingly to assume the rank from 1 to N Each stream is mapped independently, where the current stream is noted as the kth, $1 \leq k \leq K$

3 *Cumulative distribution matching* find the *relative* position for the feature in the kth stream of the nth frame in the target distribution It is achieved by matching the cumulative integration of the measured feature distribution $f_m(x)$ with that of the target distribution $f_t(y)$ in the manner shown by Equation 6 3

$$\int_{x=1}^{u} f_m(x)dx = \int_{y=1}^{v} f_t(y)dy, \tag{6 3}$$

where u, v are the ranks of the current feature in the measured and target distributions, respectively

4. *Feature revaluing*: find the *absolute* feature value of the parameter that ranked the vth in the target distribution.

Figure 6.5 illustrates the procedures in mapping the feature frames of a specific type.

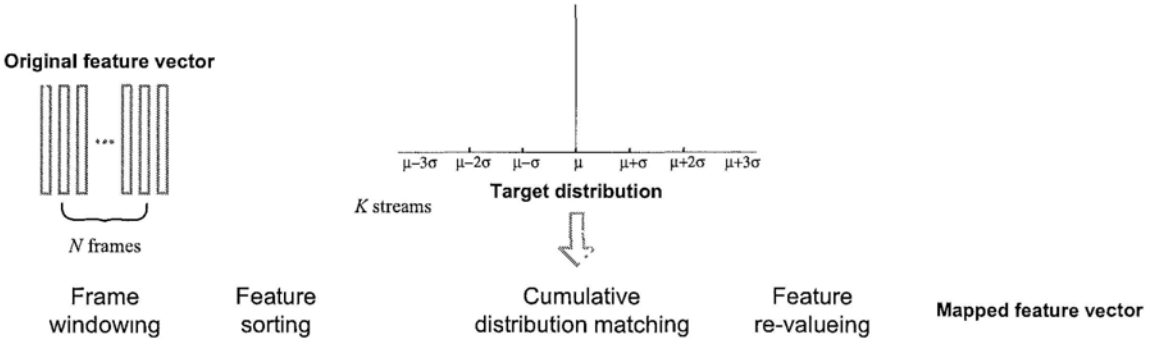


Figure 6.5: *Flowchart of the feature mapping approach.*

6.2.2 Normal distribution warping

Feature mapping is carried out on features so as to construct a more robust representation for each stream of feature distribution. Figure 6.6 illustrates how to condition the distribution of a feature stream such that it will observe a target distribution. Thus, in the implementation of feature mapping, the first step is to choose an appropriate target distribution for the features.

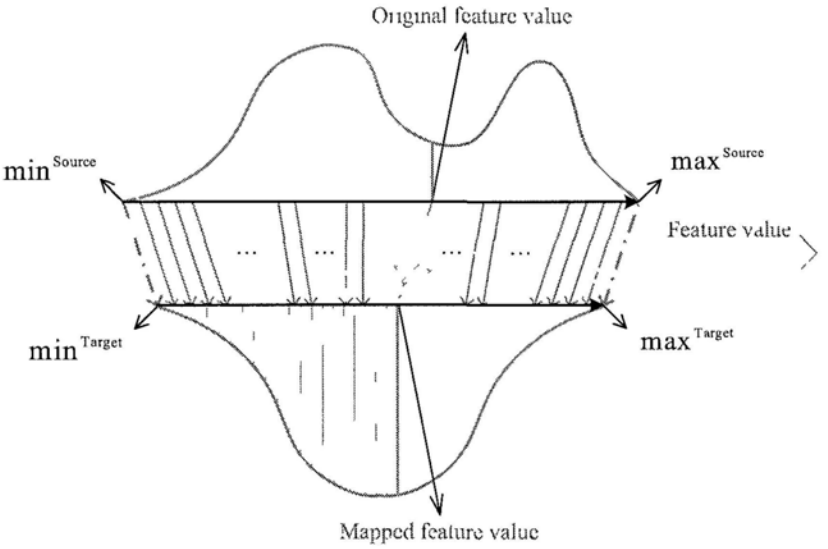


Figure 6.6: *Mapping of features according to a target distribution.*

MFCC features are known to be multi-modal, thus the ideal target for a MFCC feature is ought to be of some form that consist of multi-modal Gaussian components. However, under various channel and noise interference, this distribution might be corrupted. For an efficient mapping scheme, it is usually assumed that the target speech features conform to a specific distribution shape, like single mode Gaussian distribution shown in Figure 6.7. This normal distribution warping scheme has been applied to cepstral features in speech, conceptually similar to the histogram equalization processing that frequently used in image analysis.

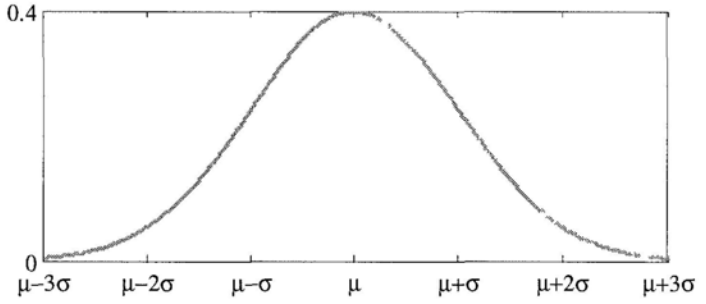


Figure 6.7: *A normal distribution with mean μ and standard deviation σ.*

6.2.3 New perspectives for modulation parameters

Rather than MFCC, the underlying distribution of the SAIF parameters has not been inspected so far. Since SAIF parameters bear implicit physical meaning, it is necessary to observe and measure their source distributions prior to assigning targets for them.

Considering that parameters in a certain SAIF feature set are derived from a series of frequency bands, for example, 40 for SAIF_*40*, in examining the genuine distribution of SAIF parameters, statistics should be made across various frequency bands independently. Figure 6.8 inspects their histogram statistics in four frequency regions, through sub-figures noted from (a) to (d), which corresponds to the low-, medium- and high-frequency regions, respectively. In each sub-figure, histograms of four streams of SAIF parameters are exhibited, the statistics are made over an utterance which contains 564 frames in total. It is found that the individual parameter in a SAIF set, which records the most dominant frequency in certain subband of the speech signal, might well be of single Gaussian distribution in nature. Besides, the center and spread of features varies, from one stream to another, and they are observed to be related with the center frequency and bandwidth of concerned subband of speech signal.

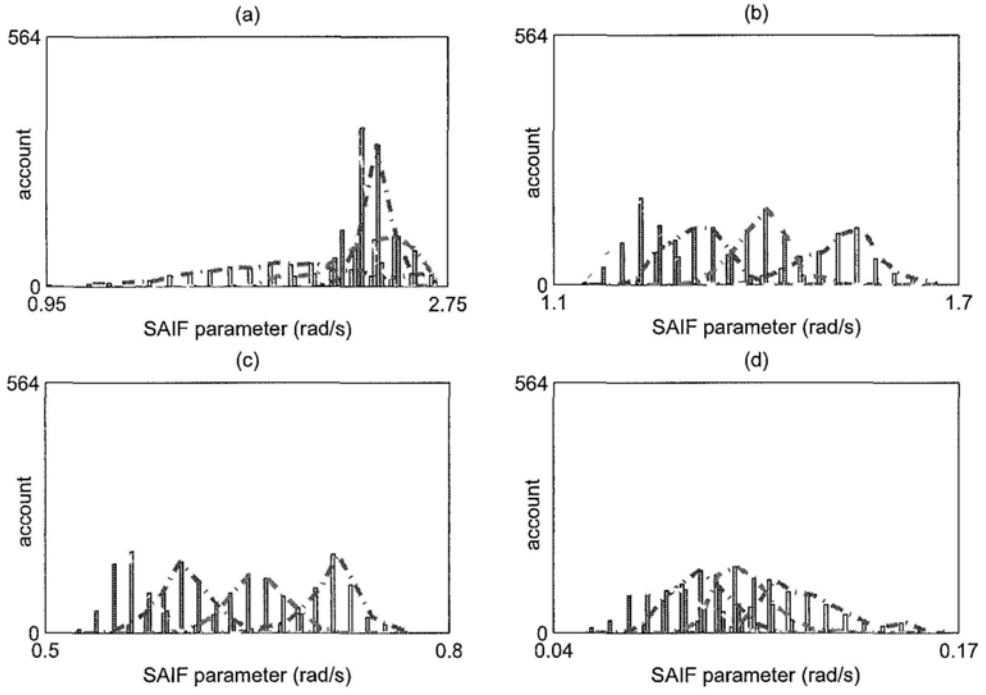


Figure 6.8: *Histogram statistics of certain streams of SAIF features over an utterance: (a). $k = 1, \ldots, 4$, (b). $k = 10, \ldots, 13$, (c). $k = 20, \ldots, 23$, and (d). $k = 37, \ldots, 40$ (k as subband index).*

The SAIF features is therefore being mapped so that their distributions are of single Gaussian distributions. For MFCC, all the feature streams share the standard normal distribution as their target in the mapping, as in [106]. While, for the SAIF features, parameters in different subbands are treated independently. For the SAIF stream from the kth subband, the target mean and standard deviation are chosen to be the center frequency $\Omega_c(k)$ and one third of the corresponding filter bandwidth, respectively.

In Figure 6.9, SAIF feature vectors under clean, 10dB and 0dB SNR noise conditions as well as the ones after mapping are shown. It is found that prior to the mapping process, difference among the three sets of parameters is relatively large in the high frequency region, although it is small in other regions. As for the mapped ones, their difference among all the frequency bands is small. It is

therefore thought that the feature mapping approach can establish more robust distributions for individual feature streams.

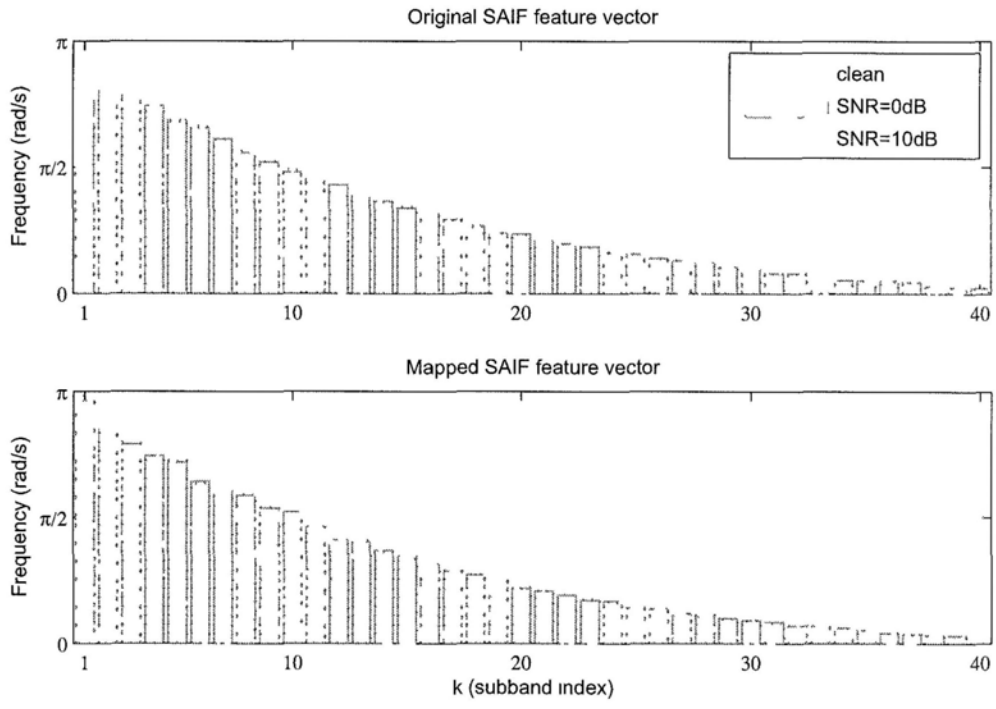


Figure 6.9: *Original and mapped SAIF feature vectors from clean and additive noise condition (with $SNR = 10dB$ and $SNR = 0dB$).*

6.3 Performance Evaluation of Robust Features

In this section, we evaluate the robust MFCC and SAIF parameters that extracted by the proposed front-end through speaker verification experiments on a dual-conditional speech database CU2C. The additive noises are from the database NOISEX-92 [148]. The standard MFCC features we used contain 39 components: the log energy, 12 static coefficients and their dynamic and acceleration coefficients.

6.3.1 Experimental set-up

For all speakers, 6 out of the 18 sessions are used to train the speaker models. For the remaining data, 6 sessions are used as development data, while the last 6 sessions are employed in performance evaluation. The standard approach for UBM-GMM training is adopted. Speaker models are trained from clean speech data, while two parallel sets of test data are used, one set is microphone data with additive noise, the other is telephone data. For both speaker model training and verification tests, we used the MFCC and SAIF features that extracted as described in Chapter 5.

Two separate systems are built based on MFCC and SAIF, respectively. Score-level fusion technique is used to combine the contributions of the two systems and produce the final decision. For the verification tasks, the log-likelihood ratio score of each test is the linear combination of the log-likelihood ratio score λ_M from MFCC and λ_S from the SAIF features, i.e., $\lambda = w_M\lambda_M + w_S\lambda_S$, where w_M and w_S are the weights on MFCC and SAIF, respectively. w_M and w_S are related by $w_M + w_S = 1$. The optimal values of w_M and w_S are determined such that they achieve the best verification performance for the development data. This is done by exhaustive search with a step size of 0.02, over the interval of $[0, 1]$. Equal error rate (EER) is used as the primary performance indicator.

6.3.2 Experimental results

The speaker verification experiments are carried out in the mismatched noise/channel conditions, where the additive noise is at $SNR = 10dB$ level.

♦ **Results from original features**

The performance of the MFCC and SAIF features that achieved under mismatched channel/noise conditions are referred to as benchmark in the following evaluation. MFCC features used are of 39-dimension and the SAIF feature vectors containing 40 parameters, respectively. Relevant results have already been reported in Table 6.1.

♦ **Results with mapped features**

Feature mapping could be applied over the whole utterance or over a relatively small window below one second. Having a long normalization window is thought to limit the robustness to noise/channel variation, while, a shorter window may reduce the compensation effectiveness to the feature distribution. The feature vectors sum up to 1 $sec \sim 3$ sec in length for most of the utterances in our evaluation data set. Thus, we implement the feature mapping with four kinds of window size. Figure 6.10 shows the speaker verification results of the mapped MFCC and SAIF features under two types of mismatch conditions, i.e., clean microphone training data with $SNR = 10dB$ test data, and telephone test data, respectively. The weight of SAIF in the combination, i.e., w_S is found to vary in the range $0.64 \sim 0.66$ and $0.40 \sim 0.50$ for the mismatched noise and channel conditions, respectively.

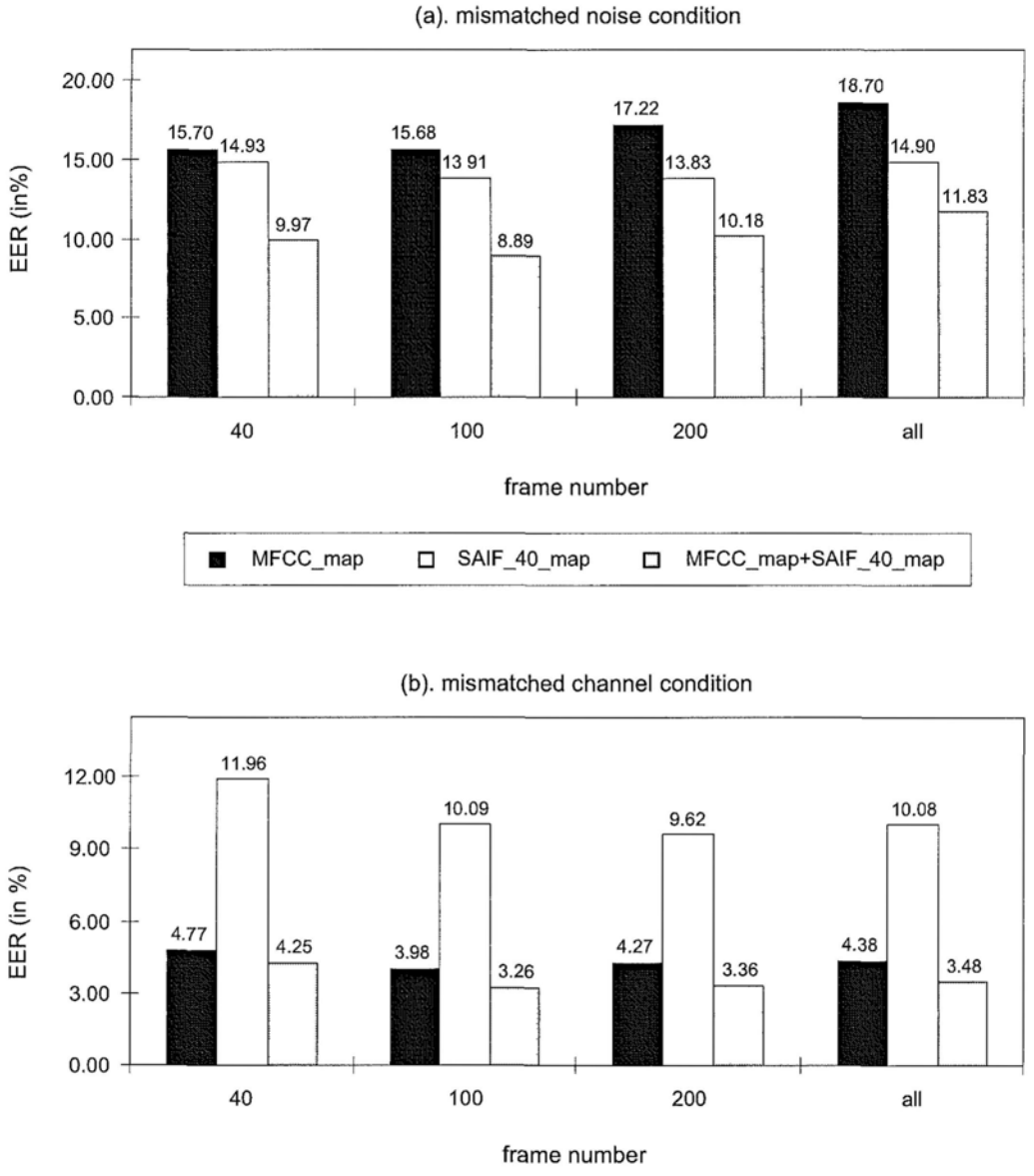


Figure 6.10: *Speaker verification results with mapped feature sets under mismatched noise/channel conditions.*

6.3.3 Analysis of results

- **Recognition performance**

As shown in Table 6.1, the training and test data mismatch in either environmental noise levels or transmission channels prevent feature set SAIF from providing adequate complement to MFCC in speaker verification experiments. By importing a mapping approach that can alter the original feature distribution as a post-processing step in the front-end, the robustness of both sets of parameters are enhanced. Through comparing the results in Table 6.1 and Figure 6.10, it is clearly revealed that for both mismatch conditions: (1). performance of individual MFCC and SAIF features is improved; and (2). combination of the magnitude-based and phase-related parameters demonstrate visible advantage, to put it concretely, the overall improvements achieved over the MFCC benchmark are 41.30% and 81.05% for the noise and channel mismatch scenarios, respectively.

- **Feature mapping effectiveness**

Through a comparison of individual feature's performance under the two mismatch scenarios, it is found that the mapping method generally offers more assistance in enhancing channel robustness of the features, especially for MFCC, where the EER is reduced by 77.03% relatively on an average than the 57.68% relative reduction by SAIF. While, generally speaking, this method is fit well for them both. Besides, for the additive noise effects, the SAIF features are observed to get more benefit than MFCC from the mapping, where the EER is averagely reduced by 38.69% for SAIF than by 3.33% for MFCC.

As an implementation issue, the results reveal that a sliding window size covering 100 frames (approximated 1 *sec* in length) may be a proper choice for speech data which are similar with ours. To illustrate the observations clearer, the individual parameter set MFCC, SAIF, as well as their combination are independently treated that each is measured based on its own baseline result that indicated by Table 6.1. The details are shown in Figure 6.11, where in comparing with the individual MFCC and SAIF features, their combination is

revealed to make the most of the effectiveness from introducing the feature mapping approach into the front-end of a speaker verification system. Ultimately, the usefulness of robust phase-specific parameters in distinguishing speakers is confirmed.

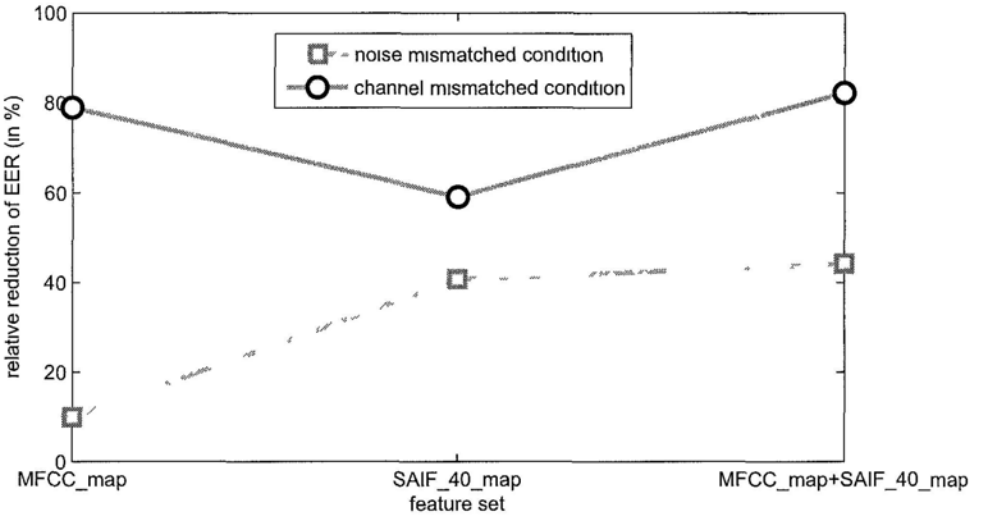


Figure 6.11: *Relative reductions of EER (with window size = 100 frames) under the channel/noise mismatched conditions (in %).*

6.4 Summary

In order to improve the cepstral feature-based speaker verification system in tackling the training-test mismatches caused by environmental/transmission variations, a robust feature extraction front-end is proposed. The front-end is primarily capable of producing instantaneous phase-related parameters from a speech signal that are proven containing complementary speaker characteristics with the cepstral features. It can also conform the distribution of speaker feature into a representation that is more robust to either additive noise effects or transmission channel variations. Theoretical analysis about the undesirable environmental effects on the phase parameter sets are given, and evaluation results under speaker verification protocol in various mismatched scenarios using the enhanced features confirm the effectiveness of the proposed method.

Chapter 7

Conclusion

7.1 Discussion and Conclusion

In this thesis, an efficient yet effective solution to the robustness problem of speaker authentication system has been provided from a feature-based perspective. In this course of study, a series of work has been conducted, where the main steps taken are as follows.

- **Identify and extract useful vocal characteristics**

Other than the signal magnitude, there are complementary information sources in speech samples that possess speaker-distinguishable characteristics. It is, however, intriguing to identify a proper and effective way to take advantage of these essential vocal properties. The central question is to derive primary components contained in segments of speech through penetrating their inner structure with a suitable tool. This tool should take account of the typical attributes of speech signals, and makes the separation of target components from others easy.

We take an analysis-by-synthesis approach to achieve this. It is adopting from the successful cases of speech synthesis/coding in which the speech segments are scrutinized through decomposition across various subbands. We decide to extract the instantaneous quantities of envelope and frequency in different subband signals through multi-band amplitude-frequency demodulation

algorithm. The sequences from the decomposed LP residual signals are found to possessing vibration patterns of voicing source; while, information of constituent formants can also be deduced from the reconstructed speech segments.

- **Design effective feature forms**

The next step ahead is to develop appropriate forms to parameterize the identified vocal characteristics into vectors, which are essential for the sake of speaker recognition.

It is found that for a particular speech segment, the envelope and instantaneous frequency sequences record the critical amplitude and frequency along the time line in each subband. This amplitude-frequency interrelation is therefore carrying relevant internal time-frequency characteristics. Besides the amplitude information, in each subband of the speech signal, the sequence is dominated by a primary frequency component that appears as the center in general. These dominant frequencies across all bands together constitute a new representation for speech. It is an unique and novel form for the concerned vocal properties, in that it seizes the temporal variation as well as their distribution across various frequency regions.

- **Inspect and analyze feature contents**

In order to examine the capability of the novel feature sets in capturing important speaker-dependent properties, a series of experiments are specifically designed for that purpose. For the excitation-related RAIE and RAIF feature vectors, they are examined in their ability of catching the F0 and principal pitch-harmonics, differentiating from varied pitch epochs, and detecting details present between epochs. The phase-concerned SAIF feature sets are evaluated to reflect formant-related characteristics, such as, harmonic-formant interaction, formant bandwidth effects, etc.

It is validated by experiments that the RAIE vectors are highly correlated with the averaged amplitude of signals that reside within specific frequency regions, which are absent from the spectrum of excitation signal. Parameters

in RAIF and SAIF vectors record the primary frequencies present in subbands of LP residual signal and speech signal, respectively.

- **Evaluate features in discriminating speakers**

The derived feature vectors are examined through speaker recognition tasks on speech database under clean environment. With the benchmarks produced by the conventional MFCCs are 2.44% IDER and 1.52% EER, the newly proposed features achieve the following results: RAIE with 27.28% IDER and 9.46% EER, RAIF with 19.67% IDER and 8.01% EER, SAIF, which outperforms the other two, attain 4.78% IDER and 2.70% EER. The thus performed features turn out offer considerable assistance to improve the MFCC-based baseline performance, among which, the optimal results are obtained by the SAIF with 1.83% IDER and 1.16% EER, where the relative improvements are 25.00% and 23.68%, respectively. In this way, the present speaker discrimination system is noticeably enhanced in terms of feature effectiveness.

- **Enhance robustness of features**

Considering the severe degradation occurs when speaker authentication facing mismatched training-test scenarios, it is expected that feature extraction front-end of a system should produce the required robust features. In our work, this is achieved by building a more robust representation for each stream of features in the vector sets independently. At the very beginning, we choose to observe and identify the adverse noise's effects on the short-term distribution of feature streams, it is then found that at most of the times, the centers are moved and the standard deviations decreased. A warping-based method is raised to maintain the distribution of parameters as consistent as possible before and after any ambient interruptions such as background noise or distorting transmission. Eventually, a robust feature extraction front-end which including a vocal source- and phase-related feature extractor and a feature enhancement post-processing module is submitted as a solution to the problem raised in the beginning of this thesis.

In speaker authentication experiments carried out at this stage, under the noise- and channel-mismatch scenarios respectively, we achieve improvements over the benchmarks from 17.41% to 10.22%, and from 18.94% to 3.59%.

We summarize the work conducted in this thesis by the following points:

1. In attempt to derive excitation-specific speaker features to complement MFCCs, the vocal attributes related with the vibrations of a speaker's vocal cords are exploited. The goal is achieved by applying multi-band AM-FM demodulation on excitation signals. The advantage of this approach lies in: first, it is a complete description of the signal rather than only noting some parameters, like the pitch-period or F0; second, it is convenient to represent the inclusive components in terms of their time-varying amplitude and frequency quantities, rather than fitting the waveform or spectrum; third, it is easy to investigate the individual components for capturing useful information. Experimental results prove *the application of this AM-FM modeling technique on deriving excitation modulation features, is not only a novel method but successful for the purpose of speaker characterization.*

2. Apart from pitch-harmonics, the formants and other constituent elements in a voiced speech signal are found to follow the law of AM-FM modulation as well. They are interpreted as the primary components in a multicomponent AM-FM speech model, where there exists a sequence of dominant frequency in each component. The phase-specific speaker-discriminative properties of speech is therefore being possessed by these primary frequencies that present in the signal. Feature vectors derived from them achieve pretty good recognition performance. It verifies that *phase information in speech signals contain speaker-dependent cues, if given well-developed extraction approach, they can be parameterized into effective speaker parameters.*

   Besides, the phase-related information in the speech signal, compared with

other modulation parameters, are more effective for discriminating different speakers. This is revealed by the performance on their own as well as combination with the cepstral features MFCCs.

3. The complementary application of MFCCs and modulation features provide a speaker recognition system with higher accuracy than using MFCCs alone in clean and matched conditions. A feature enhancement method that deployed as a post-processing for the front-end of a speaker recognition system can produce robust features which are able to noticeably alleviate the degradation of system performance in mismatched scenarios. It is clearly shown that *feature-based solution to robust speaker authentication problem is feasible and efficient.*

7.2 Contributions of This Thesis

This work has either initialized or pushed forward relevant research in the following aspects:

First of all, in attempt to prompt a source signal model for delivering useful excitation-related amplitude and frequency parameters, we have made a systematic review covering relevant works through recent years. These works mostly come from research fields such as speech coding, synthesis and analysis, etc, which are looked into in this thesis for the potential usage on source modeling, rather than for their original purposes. It is expected to provide researchers unfamiliar to this area a brief introduction.

Secondly, we introduce an idea to investigate the speaker-discriminative power of speech properties through analysis-by-synthesis approach. Traditional speaker representatives mostly focus less on reconstruction by a closed form. Nevertheless, a summation-based re-synthesis from the concerned parameters provide a new and proper channel for inspecting the features in terms of their content and capability of retaining the inclusive vocal attributes.

Thirdly, it is shown in this research that the phase information in speech signals not only can be employed as speaker representatives, but also convey some degree of robustness, given a proper way to extract and quantify them.

Fourthly, efforts are made to develop a novel feature extraction front-end which includes two successive modules, the first module is responsible for extracting effective source- and phase-related parameters, while the other is capable of conditioning the feature vectors into more robust representations that are easier to maintain in whatever conditions.

Finally, the newly derived feature sets not only show advantages in clean and matched conditions, but provide considerable improvements to the baseline system under various noise/channel mismatch.

7.3 Suggestions of Future Work

Toward the ending of this thesis, we give out some suggestions to extend our work in future.

- **Flexible feature forms**

More flexible feature forms can be adopted. We have employed a parametrization to record a fixed bands of amplitude and frequency parameters. Actually, through a closer look at the feature vectors, it is found that some streams of the features are more distinguishable and less affected by interruptions. A weighting scheme of some proper manner among the streams in a vector or a selecting criterion may help to produce features with higher efficiency.

- **Generating features from unvoiced segments**

Apart from the voiced segments of speech which are taken use in extracting vocal source- or phase-related parameters, we consider to refine the method to include the remaining unvoiced parts. Given that we generate features from instantaneous amplitude and frequency sequences, the rapid-varying characteristics therein can be captured with a smaller and variable analysis window. Besides, we have known from the summation-based speech coding scheme that a large number of subbands are involved in modeling unvoiced segments, it is possible to introduce more subbands than that for the voiced in delivering the inclusive properties owned by these segments.

The speech corpus employed in simulating speaker recognition tasks in this thesis is CU2C, information about it was briefly described in Chapter 4. In comparison with other data sets, this dual-conditional database additionally plays a distinct role, that is, to provide a platform for investigating channel effects on speaker recognition systems. This particularity has been well taken in our work. However, the evaluation of speaker features utilizing this database may be limited in terms of number of speakers and text-constraint of data (only

digital numbers contained). It is considered beneficial if we build our system on a larger-scaled data set in future.

The method we proposed in this thesis can also be applied in several other problems, two of them are listed as below for reference.

1. **Source signal modeling**: We have talked about the derivation of amplitude- and frequency-specific source parameters through multi-band decomposition procedures. It therefore makes the other side of this problem possible, that is, to establish a waveform modeling approach through the concerned amplitude-frequency characteristics. This idea is also inspired by the sinusoidal speech representation, where a summation of sinusoids is involved to build a speech waveform model. If putting the relevant vocal tract properties aside, the summed sinusoidal model can be used for delivering source signal waveform.

2. **Phonetic class determination**: A set of primary frequencies dominating a speech signal in various subbands are taken as an expression of speaker's phase characteristics. In our pilot study, it is found that this type of representative which captures frequency components present in a segment of speech signal, not only convey speaker-specific information, but behaves quite differently for various phonetic classes. It is easy to understand that the inner time-frequency organization of different phonetic classes are distinct for one and other class, for instance, vowels usually rich in low frequency frequency components which covering as low as the fundamental frequency and harmonics, while, high frequency components are sometimes absent, for the fricatives, its noise-like structure determines that plenty of high frequency components are predominant. As an extension to this task, considering the phase-related modulation features are in general characterizing the phase properties of formant-structure information in speech signals, they are thought applicable to speech recognition as well. Further study in this direction will provide us with new perspectives on exploitation of speech properties in future.

Bibliography

[1] A. Tapus, M. J. Matarić, and B. Scassellati, "Socially assistive robotics: The grand challenges in helping humans through social interaction," *IEEE Robotics & Automation Magazine*, pp. 35–42, 2007.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[3] L. G. Kersta, "Voiceprint identification," *Nature*, 1962.

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

[5] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, "Speaker authentication and identification: A comparison of spectrographic and auditory presentations of speech material," *J. Acoustical Society of America*, vol. 44, pp. 1596–1607, 1968.

[6] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-26, no. 1, pp. 43–49, 1978.

[7] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," in *Proc. ICASSP*, 1985, pp. 387–390.

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

[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 2, pp. 254–272, 1981.

[10] F. K. Soong and B. H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. ICASSP*, 1984, pp. 37–40.

[11] C. S. Liu, M. T. Lin, W. J. Wang, and H. C. Wang, "Study of line spectrum pair frequencies for speaker recognition," in *Proc. ICASSP*, 1990, pp. 277–280.

[12] S. Furui, "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 29, no. 3, pp. 342–350, 1981.

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

[14] M. Schmidt and H. Gish, "Speaker identification via support vector machines," in *Proc. ICASSP*, 1996, pp. 105–108.

[15] V. Wan, "Speaker verification using support vector machines," Ph.D. dissertation, University of Sheffield, 2003.

[16] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, 1997, pp. 963–966.

[17] T. F. Quatieri, *Discrete-time speech signal processing: principles and practice*, ser. Prentice Hall Signal Processing Series, A. V. Oppenheim, Ed. Upper Saddle River: Prentice Hall, 2002.

[18] P B. Denes and E. N. Pinson, *The speech chain: The physics and biology of spoken language*, 2nd ed. New York: W. H. Freeman and Company, 1998.

[19] L. R. Rabiner and R. W. Schafer, *Digital processing of speech signals*, ser. Prentice Hall Signal Processing Series, A. V. Oppenheim, Ed. Englewood Cliffs: Prentice Hall, 1978.

[20] B S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear predicition of the speech wave," *J. Acoustical Society of America*, vol. 50, no. 2, pp. 637–655, Apr. 1971.

[21] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoustical Society of America*, vol. 49, no. 2, pp. 583–590, 1971.

[22] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, ser Prentice Hall Signal Processing Series, A. V. Oppenheim, Ed. Englewood Cliffs: Prentice Hall, 1993.

[23] B. S. Atal, "Automatic speaker recognition based on pitch contours," *J Acoustical Society of America*, vol. 52, no. 6, pp. 1687–1697, 1972.

[24] J. W. Picone, "Signal modeling techniques in speech recognition," *Proc IEEE*, 1993

[25] J. J. P. Campbell, "Speaker recognition: A tutorial," *Proc. IEEE*, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[26] T. Robinson, "Speech analysis," webpage: http://svr-www.eng.cam.ac.uk/ ajr/SpeechAnalysis/SpeechAnalysis.html, 1998.

[27] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, ser. Springer Series in Information Sciences. New York: Springer-Verlag, 1990.

[28] B. C. J. Moore and B. R. Glasberg, "A revision of zwicker's loudness model," *Acta Acustica United with Acustica*, vol. 82, no. 2, pp. 335–345, 1996.

[29] R. D. Patterson, M. Allerhand, and C. Giguere, "Time-domain modeling of peripheral auditory processing: A modular architecture and software platform," *J. Acoustical Society of America*, vol. 98, pp. 1890–1894, 1995.

[30] T. Irino and H. Kawahara, "Signal reconstruction from modified auditory wavelet transform," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3549–3554, 1993.

[31] R. D. Patterson, "Auditory filter shapes derived with noise stimuli," *J. Acoustical Society of America*, vol. 59, no. 3, pp. 640–654, Mar. 1976.

[32] R. D. Patterson, I. Nimmo-Smith, D. L. Weber, and R. Milroy, "The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram, and speech threshold," *J. Acoustical Society of America*, vol. 72, no. 6, pp. 1788–1803, Dec. 1982.

[33] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer, Perception Group, Tech. Rep. 35, 1993.

[34] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *J. Acoustical Society of America*, vol. 101, no. 1, pp. 412–419, Jan. 1997.

[35] P. Rose, *Forensic speaker identification*, ser. Taylor & Francis forensic science series. New York: Taylor & Francis, 2002.

[36] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoustical Society of America*, vol. 51, no. 6 (Part 2), pp. 2044–2056, 1972.

[37] A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 1, pp. 4–37, 2000.

[38] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[39] H Hermansky, "Perceptual linear prediction (PLP) analysis for speech," *J. Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[40] G Fant, J. Liljencrants, and Q Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 26, no. 4, pp. 1–13, 1985.

[41] P Thevenaz and H. Hugli, "Usefulness of the LPC-residue in text-independent speaker verification," *Speech Communication*, vol. 17, no 1-2, pp. 145–157, Aug. 1995.

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

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

[44] T. Kinnunen and P. Alku, "On separating glottal source and vocal tract information in telephony speaker verification," in *Proc. ICASSP*, 2009, pp. 4545–4548.

[45] J. Gudnason and M. Brookes, "Voice source cepstrum coefficients for speaker identification," in *Proc. ICASSP*, 2008, pp. 4821–4824.

[46] S. Prasanna, C. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, no. 10, pp. 1243–1261, 2006.

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

[48] I. Magrin-Chagnolleau, G. Durou, and F. Bimbot, "Application of time-frequency principal component analysis to text-independent speaker identification," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 371–378, Sep. 2002.

[49] N. Malayath, H. Hermansky, S. Kajarekar, and B. Yegnanarayana, "Data-driven temporal filters and alternatives to GMM in speaker verification," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 55–74, 2000.

[50] T. Kinnunen, C. Koh, L. Wang, H. Li, and E. Chng, "Temporal discrete cosine transform: Towards longer term temporal features for speaker verification," in *Proc. ISCSLP*, 2006.

[51] J. C. R. Jankowski, T. F. Quatieri, and D. A. Reynolds, "Formant AM-FM for speaker identification," in *Proc. IEEE International Symposium on Time-frequency and Time-scale Analysis*, 1994, pp. 608–611.

[52] R. Kumaresan and A. Rao, "Model-based approach to envelope and positive instantaneous frequenecy of signals and application to speech," *J. Acoustical Society of America*, vol. 105, no. 3, pp. 1912–1924, Mar. 1999.

[53] L. Atlas and S. Shamma, "Joint acoustic and modulation frequency," *EURASIP J. on Applied Signal Processing*, vol. 2003, no. 7, pp. 668–675, 2003.

[54] D. Dimitriadis, P. Maragos, and A. Potamianos, "Auditory teager energy cepstrum coefficients for robust speech recognition," in *Proc. Signals, Systems and Computers*, 2003, pp. 2078–2082.

[55] D. V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition," *IEEE Signal Processing Letters*, vol. 12, no. 9, pp. 621–624, Sep. 2005.

[56] V. Tyagi, "FEPSTRUM: An improved modulation spectrum for ASR," in *Proc. Interspeech*, 2007, pp. 1114–1117.

[57] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of FM components from speech signals using all-pole model," *Electronics Letters*, vol. 44, no. 6, pp. 449–450, 2008.

[58] Y. Kubo, S. Okawa, A. Kurematsu, and K. Shirai, "A comparative study on AM and FM features," in *Proc. Interspeech*, 2008, pp. 642–645.

[59] K. V. S. Narayana and T. V. Sreenivas, "Comparison of AM-FM based features for robust speech recognition," in *Proc. Interspeech*, 2008, pp. 1545–1548.

[60] N. Wang, P. C. Ching, and T. Lee, "Exploitation of phase information for speaker recognition," in *Proc. Interspeech*, 2010, pp. 2126–2129.

[61] ——, "Exploration of vocal excitation modulation features for speaker recognition," in *Proc. Interspeech*, 2009, pp. 892–895.

[62] A. Adami, "Modeling prosodic differences for speaker recognition," *Speech Communication*, vol. 49, no. 4, pp. 277–291, 2007.

[63] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP*, 2003, pp. 788–791.

[64] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Communication*, vol. 46, no. 3-4, pp. 455–472, 2005.

[65] G. Doddington, "Speaker recognition based on idiolectal differences between speakers," in *Proc. Eurospeech*, 2001, pp. 2521–2524.

[66] W. Andrews, M. Kohler, J. Campbell, J. Godfrey, and J. Hernández-Cordero, "Gender-dependent phonetic refraction for speaker recognition," in *Proc. ICASSP*, 2002, pp. 149–152.

[67] K. Y. Leung, M. W. Mak, M. H. Siu, and S. Y. Kung, "Adaptive articulatory feature-based conditional pronunciation modeling for speaker verification," *Speech Communication*, vol. 48, no. 1, pp. 71–84, 2006.

[68] A. Higgins, L. Bhaler, and J. Porter, "Voice identification using nearest neighbor distance measure," in *Proc. ICASSP*, Minneapolis, MN, 1993, pp. 375–378.

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

[70] 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.

[71] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis.* New York: Wiley, 1973.

[72] J. L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[73] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, 1995.

[74] J. P. Eatock and J. S. Mason, "A quantitative assessment of the relative speaker discriminating properties of phonemes," in *Proc. ICASSP*, 1994, pp. 133–136.

[75] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1-2, pp. 91–108, Aug. 1995.

[76] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-27, no. 2, pp. 113–120, Apr. 1979.

[77] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*, 1979, pp. 208–211.

[78] R. J. Mcaulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-28, no. 2, pp. 137–145, Apr. 1980.

[79] P. Lockwood and J. Boudy, "Experiments with a non-linear spectral subtractor (NSS), Hidden Markov Models and the projections, for robust recognition in cars," *Speech Communication*, vol. 11, no. 2-3, pp. 215–228, 1992.

[80] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 2, pp. 126–137, Mar. 1999.

[81] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[82] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-26, no. 3, pp. 197–210, Jun. 1978.

[83] J. Meyer and K. U. Simmer, "Multi-channel speech enhancement in a car environment using wiener filtering and spectral subtraction," in *Proc. ICASSP*, 1997, pp. 1167–1170.

[84] K. Paliwal and A. Basu, "A speech enhancement method based on Kalman filtering," in *Proc. ICASSP*, 1987, pp. 177–180.

[85] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 4, pp. 373–385, Jul. 1998.

[86] S. D. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. ICASSP*, 2002, pp. 4164–4167.

[87] N. B. Yoma and M. Villar, "Speaker verification in noise using a stochastic version of the weighted Viterbi algorithm," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 3, pp. 158–166, Mar. 2002.

[88] Y. Chen and Y. Rui, "Real-time speaker tracking using particle filter sensor fusion," in *Proc. IEEE*, vol. 92, no. 3, Mar. 2004, pp. 485–494.

[89] H. G. Hirsch and C. Ehrlicher, "Noise estimation techniques for robust speech recognition," in *Proc. ICASSP*, 1995, pp. 153–156.

[90] V. Stahl, A. Fischer, and R. Bippus, "Quantile based noise estimation for spectral subtraction and Wiener filtering," in *Proc. ICASSP*, 2000, pp. 1875–1878.

[91] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[92] I. Cohen, "Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[93] R. J. Mammone, X. Zhang, and R. P. Ramachandran, "Robust speaker recognition: A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, Sep. 1996.

[94] M. G. Rahim, B. H. Juang, W. Chou, and E. Buhrke, "Signal conditioning techniques for robust speech recognition," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 107–109, Apr. 1996.

[95] L. Burget, P. Matějka, P. Schwarz, O. Glembek, and J. Černocký, "Analysis of feature extraction and channel compensation in a GMM speaker recognition system," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 1979–1986, Sep. 2007.

[96] A. E. Rosenberg, C. H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proc. ICSLP*, 1994, pp. 1835–1838.

[97] O. Viikki and K. Laurila, "Ceptral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1-3, pp. 133–147, Aug. 1998.

[98] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Proc. ICASSP*, 1999, pp. 325–328.

[99] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[100] X. Zhang and R. J. Mammone, "Channel and noise normalization using affine transformed cepstrum," in *Proc. ICSLP*, 1996, pp. 1993–1996.

[101] T. F. Quatieri, D. A. Reynolds, and G. C. O'Leary, "Estimation of handset nonlinearity with application to speaker recognition," *IEEE Trans. Speech and Audio Processing*, vol. 8, no. 5, pp. 567–584, Sep. 2000.

[102] C. T. Lu and H. C. Wang, "Enhancement of single channel speech based on masking property and wavelet transform," *Speech Communication*, vol. 41, no. 2-3, pp. 409–427, Oct. 2003.

[103] N. Wang, P. C. Ching, N. Zheng, and T. Lee, "Robust speaker recognition using denoised vocal source and vocal tract features," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 196–205, Jan. 2011.

[104] M. Rahim, Y. Bengio, and Y. LeCun, "Discriminative feature and model design for automatic speech recognition," in *Proc. Eurospeech*, 1997, pp. 75–78.

[105] L. P. Heck, Y. Konig, M. K. Sonmez, and M. Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communication*, vol. 31, no. 2-3, pp. 181–192, Jun. 2000.

[106] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. A Speaker Odyssey - The Speaker Recognition Workshop*, 2001, pp. 213–218.

[107] B. Xiang, U. V. Chaudhari, J. Navrátil, G. N. Ramaswamy, and R. A. Gopinath, "Short-time gaussianization for robust speaker verification," in *Proc. ICASSP*, 2002, pp. 681–684.

[108] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, Jan./Apr./Jul. 2000.

[109] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. ICASSP*, 2003, pp. 49–52.

[110] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 42–54, Jan./Apr./Jul. 2000.

[111] M. Ben, R. Blouet, and F. Bimbot, "A Monte-Carlo method for score normalization in automatic speaker verification using Kullback-Leibler distances," in *Proc. ICASSP*, 2002, pp. 689–692.

[112] M. J. F. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech and Audio Processing*, vol. 4, no. 5, pp. 352–359, Sep. 1996.

[113] L. P. Wong and M. Russell, "Text-dependent speaker verification under noisy conditions using parallel model combination," in *Proc. ICASSP*, 2001, pp. 457–460.

[114] S. Sagayama, Y. Yamaguchi, S. Takahashi, and J. Takahashi, "Jacobian approach to fast acoustic model adaptation," in *Proc. ICASSP*, 1997, pp. 835–838.

[115] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in Text-Independent speaker verification," in *Proc. Interspeech*, 2005, pp. 3117–3120.

[116] R. Teunen, B. Shahshahani, and L. Heck, "A model-based transformational approach to robust speaker recognition," in *Proc. ICSLP*, 2000, pp. 495–498.

[117] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP*, 2003, pp. 53–56.

[118] P. Kenny, M. Mihoubi, and P. Dumouchel, "New MAP estimators for speaker recognition," in *Proc. Eurospeech*, 2003, pp. 2961–2964.

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

[120] P. Kenny and P. Dumouchel, "Experiments in speaker verification using factor analysis likelihood ratios," in *Odyssey - The Speaker and Language Recognition Workshop*, 2004, pp. 219–226.

[121] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

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

[123] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Proc. ICASSP*, 2006, pp. 97–100.

[124] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proc. ICASSP*, 2005, pp. 629–632.

[125] D. G. Childers, "Vocal quality factors: Analysis, synthesis, and perception," *J. Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, Nov. 1991.

[126] G. Fant, "Glottal source and excitation analysis," *Speech transmission laboratory, Royal Institute of Technology, Quarterly Progress and Status Report*, vol. 20, no. 1, pp. 85–107, 1979.

[127] D. H. Klatt, "Review of text-to-speech converstion for English," *J. Acoustical Society of America*, vol. 82, no. 3, pp. 737–793, Sep. 1987.

[128] D. G. Childers, "Gender recognition from speech, Part II: Fine analysis," *J. Acoustical Society of America*, vol. 90, no. 4, pp. 1841–1856, Oct. 1991.

[129] H. M. Hanson, P. Maragos, and A. Potamianos, "Finding speech formants and modulations via energy separation: with application to a vocoder," in *Proc. ICASSP*, 1993, pp. 716–719.

[130] R. J. Mcaulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.

[131] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications, 1996.

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

[133] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the properties of a time-varying quasi-harmonic model of speech," in *Proc. Interspeech*, 2008, pp. 1044–1047.

[134] D. W. Griffin and J. S Lim, "Multiband excitation vocoder," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, Aug. 1988.

[135] A. J. Abrantes, J. S. Trancoso, and M Isabel, "Hybrid sinusoidal modeling of speech without voicing decision," in *Proc. Eurospeech*, 1991, pp. 231–234.

[136] L. B. Almeida and F. M. Silva, "Variable-frequency synthesis: An improved harmonic coding scheme," in *Proc. ICASSP*, 1984, pp. 437 – 440.

[137] J. F. Kaiser, "On a simple algorithm to calculate the "energy" of a signal," in *Proc. IEEE*, 1990, pp. 381–384.

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

[139] H. M. Teager and S. M. Teager, *Speech production and speech modelling.* Boston, MA: Kluwer, 1990, ch. Evidence for nonlinear sound production mechanisms in the vocal tract, pp. 241–261.

[140] A. Potamianos and P. Maragos, "A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation," *Signal Processing*, vol. 37, no. 1, pp. 95–120, May 1994.

[141] G. Hu and D. L. Wang, "Monaural speech segregation based on pitch tracking and amplitude modulation," *IEEE Trans. Neural Networks*, vol. 15, no. 5, pp. 1135–1150, Sep. 2004.

[142] M. Wu, D L. Wang, and G. J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, May 2003.

[143] W. N. Chan, N. Zheng, and T. Lee, "Discrimination power of vocal source and vocal tract related features for speaker segmentation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1884–1892, Aug. 2007.

[144] J. Chen, Y. Huang, Q. Li, and K. K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 258–261, Feb. 2004.

[145] R. J. Mcaulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE*, 1990, pp. 249–252.

[146] D. Talkin, *Speech Coding and Synthesis*. Elsevier, 1995, ch. A robust algorithm for pitch tracking.

[147] N. Zheng, C. Qin, T. Lee, and P. C. Ching, "CU2C: A dual-condition Cantonese speech database for speaker recognition applications," in *Proc. Oriental-COCOSDA*, 2005, pp. 67–72.

[148] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, pp. 247–251, 1993.

[149] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *J. Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, Jun. 1996.

[150] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 36–48, Jan. 1998.

[151] P. Mermelstein, "Articulatory model for the study of speech production," *J. Acoustical Society of America*, vol. 53, no. 4, pp. 1070–1082, Apr. 1973.

[152] H. B. Richards, J. S. Mason, M. J. Hunt, and J. S. Bridle, "Deriving articulatory representations from speech with various excitation modes," in *Proc. ICSLP*, 1996, pp. 1233–1236.

[153] J. L. Flanagan, C. H. Coker, L. R. Rabiner, R. W. Schafer, and N. Umeda, "Synthetic voices for computers," *IEEE Spectrum*, vol. 7, no. 10, pp. 22–45, Oct. 1970.

[154] D. O'Shaughnessy, *Speech communication: human and machine*, ser. Addison-Wesley series in electrical engineering. Reading, MA: Addison-Wesley Publishing Co., 1987.

[155] K. K. Paliwal and B. S. Atal, "Representing frequencies in speech," AT&T Research Labs, Florham Park, NJ, Tech. Report, Jan. 2000.

[156] Y. Wang, J. Hansen, G. K. Allu, and R. Kumaresan, "Average instantaneous frequency (AIF) and average log-envelopes (ALE) for ASR with the Aurora 2 database," in *Proc. Eurospeech*, 2003, pp. 25–28.

[157] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[158] R. Mammone, X. Zhang, and R. Ramachandran, "Robust speaker recognition – A feature-based approach," *IEEE Signal Processing Magazine*, vol. 13, no. 5, pp. 58–71, 1996.

[159] S. van Vuuren and H. Hermansky, "On the importance of components of the modulation spectrum for speaker verification," in *Proc. ICSLP*, vol. 2, 1998.

[160] J. Koolwaaij and L. Boves, "Local normalization and delayed decision making in speaker detection and tracking," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 113–132, 2000.