THESIS REPORT

Master's Degree

Application of Auditory Representations on Speaker Identification

by Taishih Chi
Advisor: S. A. Shamma

M.S. 97-9

Sponsored by
the National Science Foundation
Engineering Research Center Program,
the University of Maryland,
Harvard University,
and Industry

Abstract

Title of Thesis:	APPLICATION OF AUDITORY REPRESENTATIONS ON SPEAKER IDENTIFICATION
Degree candidate:	Taishih Chi
Degree and year:	Master of Science, 1997
Thesis directed by:	Professor S. A. Shamma Department of Electrical Engineering

The noise-robustness of auditory spectrum and cortical representation is examined by applying to text-independent speaker identification tasks. A Bayes classifier residing on M-ary hypothesis test is employed to evaluate the robustness of the auditory cepstrum and demonstrate its superior performance to that of the well-studied mel-cepstrum. In addition, the phase feature of the wavelet-transform based multiscale cortical representation is shown much more stable than magnitude feature in characterizing speakers by correlator technique which is traditionally used in scene matching application. This observation bears consistence to physiological and psychoacoustic phenomena.

The underlying purpose of this study is to inspect the inherent robustness of the auditory representations which are derived from a human perception based model. The experimental results indicate that the biologically motivated features significantly enhance the speaker identification accuracy under noisy environments.

APPLICATION OF AUDITORY REPRESENTATIONS ON SPEAKER IDENTIFICATION

by

Taishih Chi

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland at College Park in partial fulfillment
of the requirements for the degree of
Master of Science
1997

Advisory Committee:

Professor S. A. Shamma, Chairman/Advisor
Professor R. Chellappa
Professor P.S. Krishnaprasad

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 An Overview

Speech plays a major role for human beings in communicating with each other. In general, the speech signal conveys information not only about the spoken words or message, but also the identity of the speaker. While the area of speech recognition concerns about the underlying linguistic message in an utterance, the area of speaker recognition concerns about the identity of the person who is speaking. Depending on the application, the speaker recognition task is divided into two further categories: verification and identification. In verification, the goal is to determine if a person is whom he/she claims. In identification, the goal is to determine exactly whom the person is in a specific group. Furthermore, in either field the speech can be constrained to be a known phrase (text-dependent) or totally unconstrained (text-independent). Success in speaker identification task depends on extracting the speaker-dependent characteristics from the speech signal that can effectively represent different speakers. After the feature extraction, the speaker identification task falls under the general problem of pattern

classification.

Currently, the most popular representation for acoustic signal processing for speaker identification is the cepstrum coefficients, the homomorphic equivalence of the short-time Fourier spectrum, which are motivated by human audition. The idea of using long-term average acoustic features is to average out the speech dependent factors and leave only the speaker dependent components. Since the ultimate goal of the speaker identification system is to perform remarkable ability to recognize speaker by increasing the robustness of the signal representation and making the system relatively insensitive to noise and reverberation, in much the same way as the human *ear*, it is natural to investigate the spectral analysis methods in human auditory system and apply these functional principles to represent speech signal. In recent decades, the adoption of mimicking auditory processes like mel-frequency scale [20] has led to significant improvements in performance over systems using traditional parametric representations, such as linear prediction code (LPC), cepstrum, their temporal derivatives and reflection coefficients [3, 14].

However, the auditory approaches often involve complex, multistage, nonlinear transformations that make theoretical analysis and practical implementation very difficult. In this work, a well-defined auditory model is employed to process the speech signal and the more robust representations of the signal are evaluated for text-independent speaker identification task. Roughly, this model consists two major portions: the peripheral system and the primary cortex. The output of the peripheral auditory system is called *auditory spectrum* in this study. This auditory spectrum can be thought as the spectral estimation of the sound signal based on human perception. Furthermore, the primary cortex functions like a

wavelet-transform-based spectral profile analyzer. This stage generates the multiscale representation which is called *cortical representation* by transforming the input auditory spectrum into its ripple domain. More detailed introduction of this biologically motivated auditory model is provided in Section 1.2. In brief, the main purpose of this paper is to investigate the robustness of the auditory spectrum and cortical representation in solving the speaker identification problem. All the speech data involved in experiments of this work is extracted from TIMIT database and Section 1.3 gives an introduction of this database.

In Chapter 2, some experiments are conducted to compare the performance of the identification system of the well-studied mel-cepstrum feature and the auditory cepstrum feature which is derived from the auditory spectrum and will be defined later. To simplify matters, these two multivariate cepstral features are assumed to be Gaussian distributed. Finally, the Bayes classifier is applied to the pattern recognition problem and the experimental results are depicted in Section 2.4. In Chapter 3, the two-dimensional cortical representation that conveys all information about the auditory spectrum based on different scales is considered as a complex image. Therefore, the correlator approach that is often applied to scene matching problems is employed to examine the robustness of cortical representation in identifying speakers. The experimental evaluations for verifying a particular speaker by cortical representation, auditory spectrum and LPC spectrum are compared in Section 3.4.1. In addition, the significant robustness of the phase response of cortical representation in distinguishing speakers is demonstrated in Section 3.4.2.

The underlying purpose of this paper is to inspect the inherent robustness of the features, auditory spectrum and cortical representation, which are derived

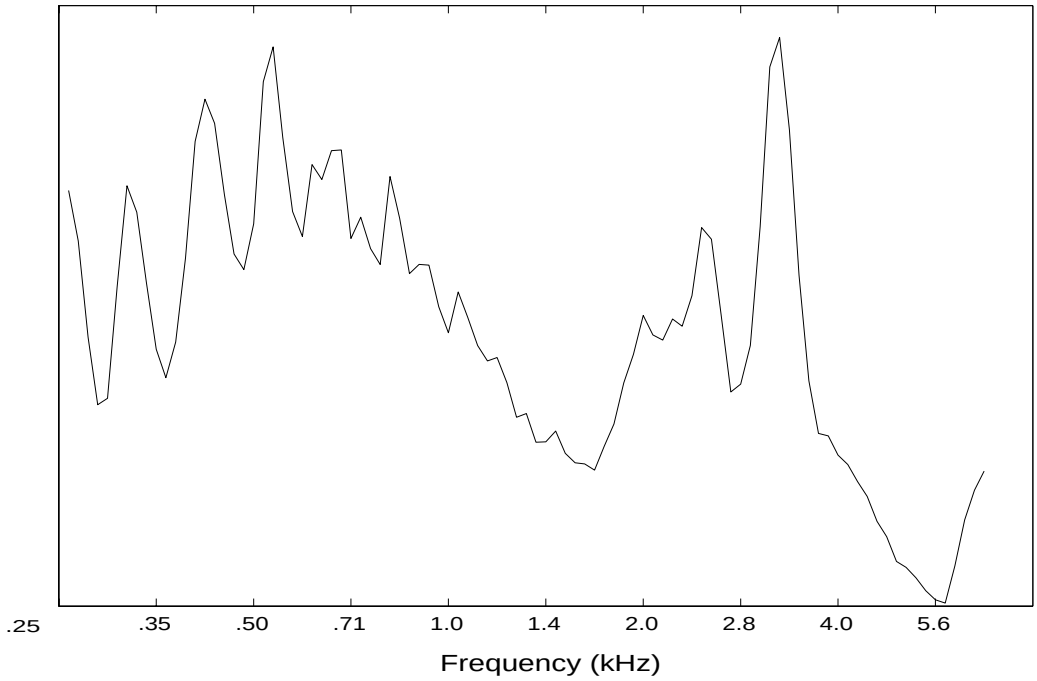

Figure 1.1: The long-term average auditory spectrum from utterance 'Come home right away.' by a male speaker. The amplitude ordinate is normalized to arbitrary units.

from a human perception based model. The experiments conducted in this work elaborate the intuition that it is beneficial to examine speech problem from a psychophysical point of view. Finally, some conclusions and directions for further studies are given in Chapter 4.

1.2 The Auditory Model

The motivation for investigating auditory functional principles is to gain an understanding of the way humans process and decode complex sounds,to be able to implement similarly robust recognition methods for acoustic signal. After years of study, an auditory model consists of peripheral and cortex model is

created to emulate the function of the human auditory system [23, 26].

The peripheral model can be reduced to three major stages : *analysis*, *transduction*, and *reduction*. First, the *analysis* stage converts the acoustic signal into a complex spatiotemporal pattern of displacements along the basilar membrane of the cochlea. It is just like performing an affine wavelet transform on the input time signal and interpreting the continuous spatial axis of the cochlea as the *scale* parameter axis. The *transduction* stage is to model the inner hair cells of the cochlea which transduce the spatiotemporal patterns of basilar membrane vibrations into intracellular hair cell potentials. In the *reduction* stage, the auditory-nerve transmits the sound evoked activity (the hair cell potentials) to the cochlear nucleus of the central auditory system. This stage is implemented biologically by a neural network known as the *lateral inhibitory network (LIN)* which generates a "spectral" profile of the stimulus by rapidly detecting discontinuities along the spatial axis of the auditory-nerve patterns and integrating its outputs over a few milliseconds [28]. In summary, this peripheral auditory process analyzes a complex sound with a topographically organized array of channels that are tuned to different characteristic frequencies. From the systematic viewpoint, the input to this peripheral system model is an ordinary speech waveform and the output is a short-time spectral profile representation of the signal, which will be referred to as the *auditory spectrum* of the signal. This *auditory spectrum* has been shown to be significantly insensitive to wideband distortion and robust to noise [25]. Figure 1.1 shows an example of the long-term average *auditory spectrum* extracted from the utterance 'Come home right away.' spoken by a male speaker in TIMIT speech database. Obviously, pitch, an important feature for identifying speaker, is clearly demonstrated in this example. The schematic

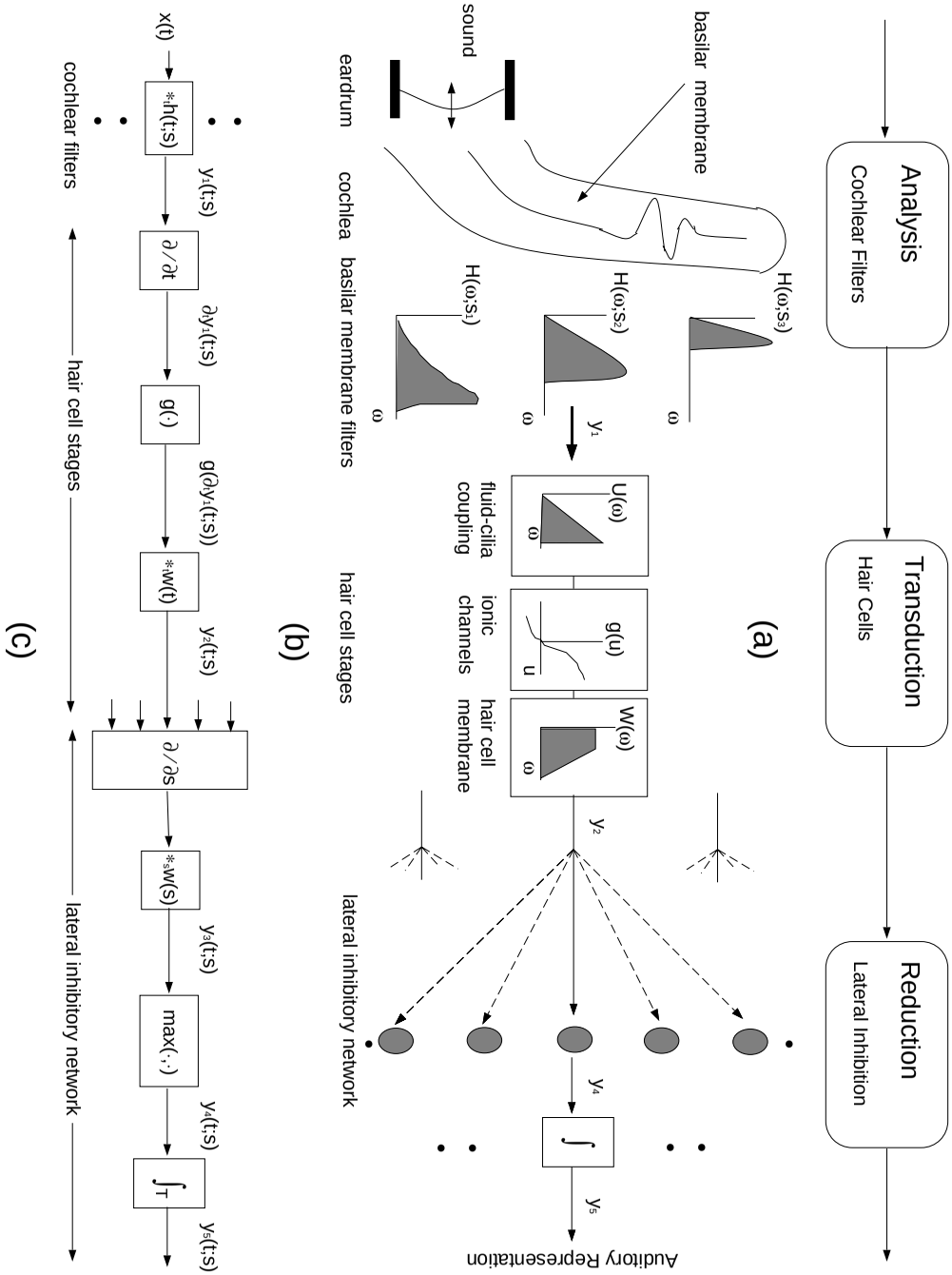

Figure 1.2: Schematic description of the peripheral auditory model (adapted from [28]). (a)Block diagram of the three basic stages in the early auditory system. (b) Quasi-anatomic sketches of the auditory stages. (c) Mathematical model of the different stages.

description of this peripheral model and the corresponding mathematical model for the three stages is shown in Figure 1.2.

In the early stages of auditory processing, a robust representation of the acoustic spectrum, *auditory spectrum*, is extracted in a series of reasonably well understood operations. In addition, a cortical model based on the physiological data [18, 19, 21] and psychoacoustical experiments with human subjects was established and analyzed the spectral profile in the higher auditory stages [27]. According to the model, the auditory cortex functionally performs a complex affine wavelet transform to the *auditory spectrum* and produces a multi-scale representation for the input auditory spectral profile at different levels of resolution. Figure 1.3 shows the different resolution representations of the *auditory spectrum* shown in Figure 1.1. The lower/higher resolution in (a)((d)) represents the frequency response of the sound signal through the channel tuned to lower/higher ripple frequency.

The schematic summary of the whole auditory process is shown in Figure 1.4 where the intermediate Central Auditory Processing stages are regarded as relay stations, and the inputs to the cortex are assumed to have similar profiles as the *auditory spectrum* that comes out from the peripheral auditory model. To sum up, the primary auditory cortex analyzes a spectral profile with a topographically organized array of channels that are tuned to different "characteristic ripple" frequencies. The overall double wavelet transform has been analyzed successfully from signal processing viewpoint. The goal here is to apply multi-scale cortical representations of acoustic signal to speaker identification problem to enhance the recognition accuracy by taking the advantages of noise-robustness characteristic.

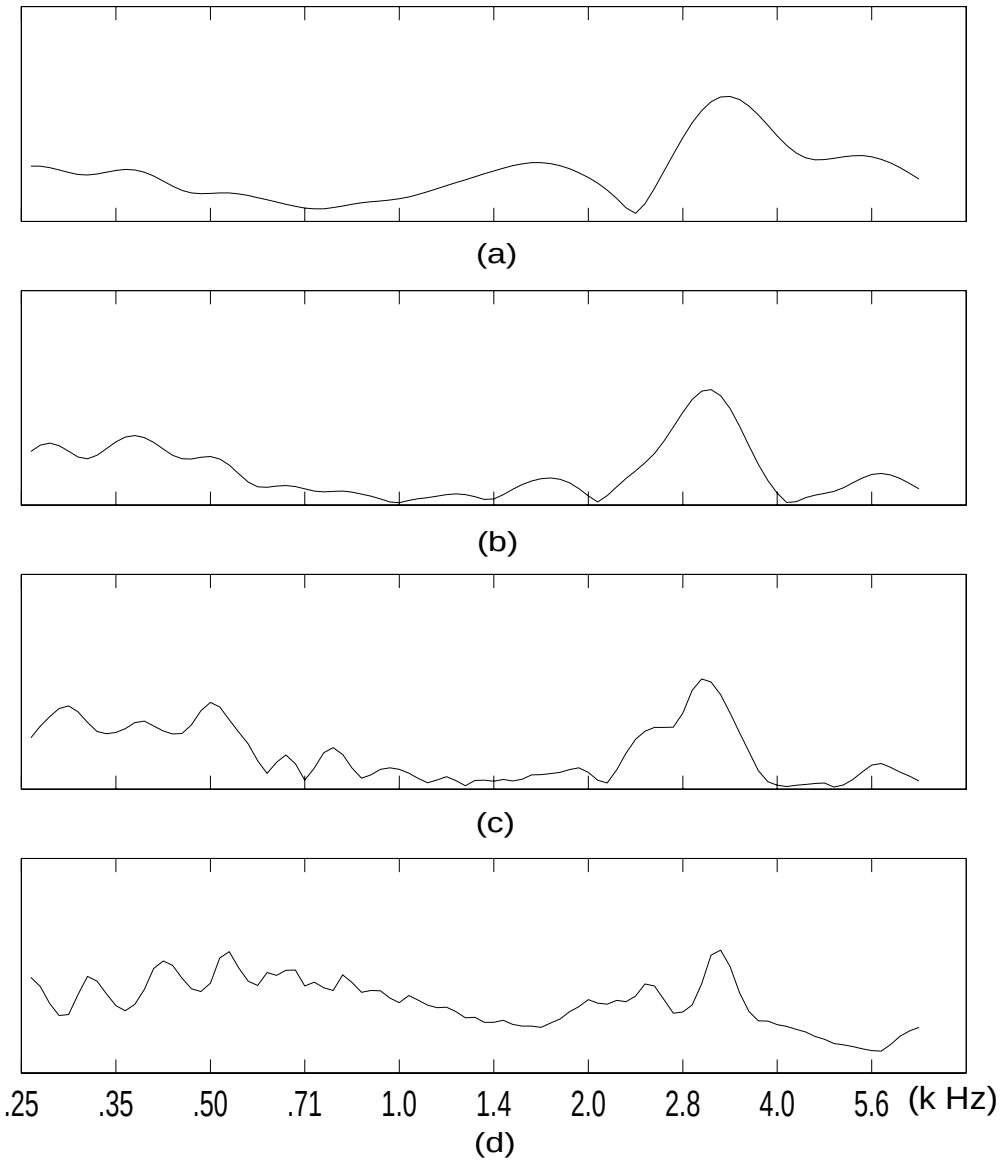

Figure 1.3: Multi-resolution representations for a long-term average auditory spectrum. The lower resolution representation shown in (a) provides the global distribution of the overall spectral energy, and the higher resolution representation in (d) shows the *local* characteristics of the spectrum.

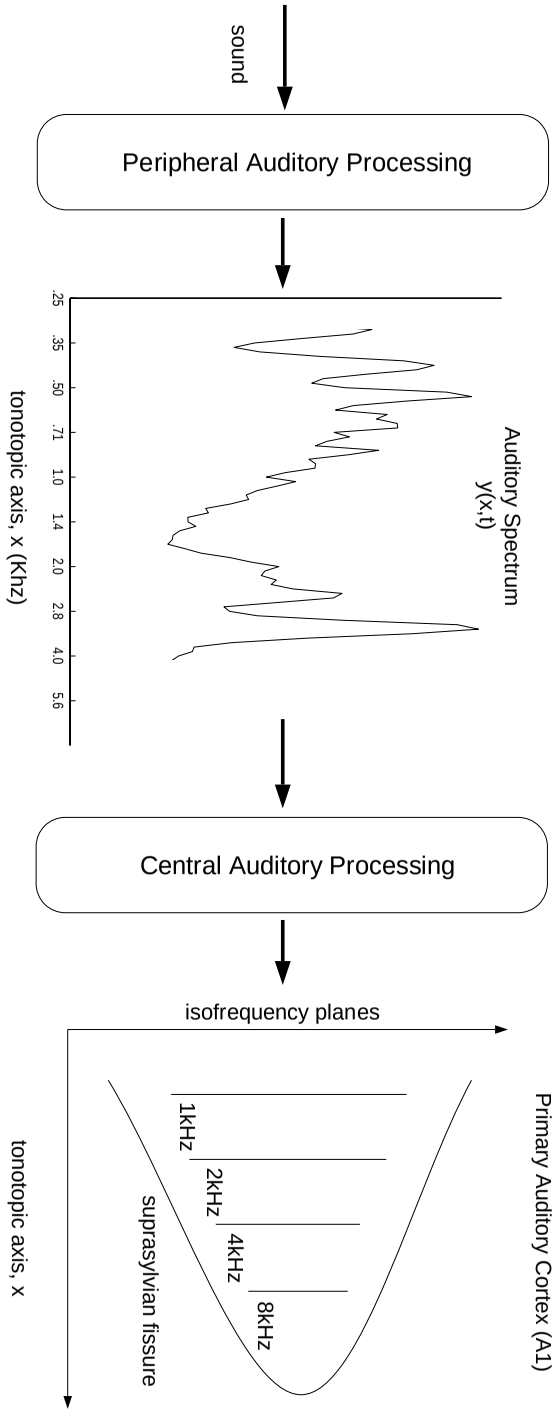

Figure 1.4: Schematic description of the auditory model (adapted from [26]).

dr1:	New England
dr2:	Northern
dr3:	North Midland
dr4:	South Midland
dr5:	Southern
dr6:	New York City
dr7:	Western
dr8:	Army Brat (moved around)

Table 1.1: Dialect regions in TIMIT database.

1.3 Database Description

All the experiments in this work were conducted using subsets of the TIMIT speech database. The TIMIT corpus was designed to provide speech data which are sampled at 16 KHz for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition systems. The database collects a total of 6300 sentences of approximately 2-4 seconds each, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. A speaker's dialect region is the geographical area of the U.S. where they lived during their childhood years. Table 1.1 shows the dialect regions classified in TIMIT database. The recognized dialect regions match geographical areas in U.S. except the Western region (dr7) in which dialect boundaries are not known with any confidence and dr8 where the speakers moved around a lot during their childhood. Table 1.2 shows the number of speakers including training and testing set for the 8 dialect regions, broken down by sex.

The text material in the TIMIT database can be categorized into 2 dialect

dr	#Male	#Female	Total
1	31	18	49
2	71	31	102
3	79	23	102
4	69	31	100
5	62	36	98
6	30	16	46
7	74	26	100
8	22	11	33
	438	192	630

Table 1.2: Dialect distribution of speakers in TIMIT database.

sentences (SA sentences), 450 phonetically-compact sentences (SX sentences) and 1890 phonetically-diverse sentences (SI sentences). The 2 SA sentences were meant to expose the dialectal variants of the speakers and were read by all 630 speakers. In addition, each speaker read 5 SX sentences which were designed to provide a good coverage of pairs of phones and 3 SI sentences which were selected from existing text sources to add diversity in sentence types and phonetic contexts. Table 1.3 summarizes the speech material in TIMIT database.

Additional information may be found in the printed documentation from National Institute of Standards and Technology (NIST# PB91-100354).

Type	#Sentences	#Speakers	Total	#Sentences/Speaker
Dialect (SA)	2	630	1260	2
Compact (SX)	450	7	3150	5
Diverse (SI)	1890	1	1890	3
Total	2342		6300	10

Table 1.3: Speech material in TIMIT database.

Chapter 2

Speaker Identification Using Cepstrum Features

2.1 Introduction

The traditional techniques for speaker identification can be categorized into three approaches. The earliest approach is to use long-term average acoustic feature, such as pitch, formant or spectrum representations. It is well known that the long-term average spectral features can represent a speaker's average vocal tract shape [13]. However, this approach discards some other useful speaker-dependent information and may require longer (>20s) utterances to get stable feature statistics.

The second approach is to model the speaker-dependent acoustic features extracted from the individual sounds. A probabilistic model, hidden Markov model (HMM), is often proposed to perform explicit segmentation of the speech into phonetic sound classes and model the temporal sequencing among the speech [8]. Intuitively, this sequential information may provide a better insight for speech recognition (text-dependent) task than for speaker identification

(speaker-dependent) task. In addition, other template based clustering algorithms such as vector quantization (VQ) can effectively represent every speaker by different codes from a codebook of spectral templates. However, due to the limited ability to model the possible variations caused by unconstrained speech, these temporal structure modeling techniques are inherently advantageous for text-dependent limited-vocabulary tasks.

The most recent approach to speaker recognition is the use of neural networks (NN's). Instead of training models for particular speakers, the NN's are trained to model the best decision function for a known speaker space. The major drawback for most NN's is that the overall network has to be retrained to find the new decision function when a new speaker is added to the system [2].

In this chapter, a speaker identification system with the structure shown in Figure 2.1 is evaluated for two different spectral features ,mel-cepstrum and *auditory cepstrum* that will be defined in Section 2.2.2. This system employed a so-called 'template matching' approach that compares an average feature derived from test data with a collection of stored average speakers' patterns (templates) which are built in training process. For text-independent speaker identification, ideally one has utterances of several seconds to ensure that a voice is modeled by features of a broad range of sounds, rather than by a particular sound. Then, test utterances are compared with trained templates by measuring the distance between them. In this study, a basic parametric model, uni-modal Gaussian model, is employed to measure the distribution distance between the test data and every speaker's template.

The mel-cepstrum feature has been well studied and successfully applied to speaker identification problems for years. Those studies have shown that the mel-

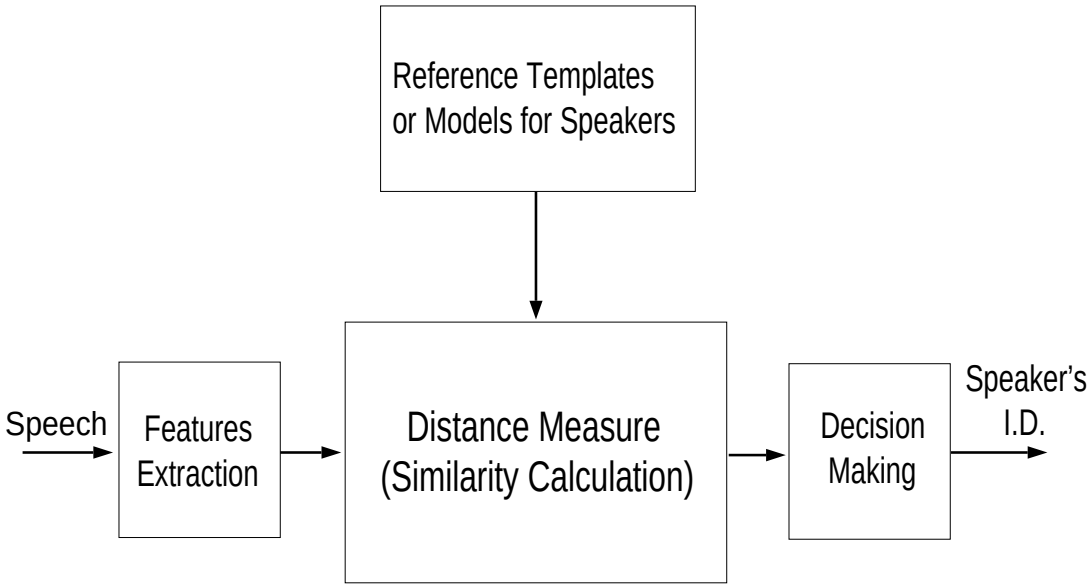

Figure 2.1: Block diagram of the proposed methodology for the speaker identification problem

cepstrum can effectively extract the vocal tract shape information of the speakers and yield good performance in distinguishing speakers [5, 14]. However, it is reasonable to expect the *auditory cepstrum*, which is derived from the *auditory spectrum*, will exhibit robust property for speaker recognition problems. In order to investigate the performance of the speaker identification system for these two feature spaces over a noisy environment, white noise is added to corrupt the speech signal and the correct identification rates corresponding to these two features are compared about increasing signal-to-noise ratio (SNR). However, the goal of this experiment is not to develop a robust algorithm for the feature spaces but to verify the extension of inherent robustness characteristic of the *auditory spectrum* which has been shown in some speech recognition applications [3].

2.2 Feature Spaces

Speech information primarily carried by the short-time spectrum is basic to many speech processing activities, including both speaker and speech recognition tasks. There have been a variety of traditional methods proposed to parameterize the short-term spectrum, such as fast Fourier transform (FFT), LPC model and Cepstral coefficient. In this study, the major motivation for comparing the two feature spaces built by *auditory cepstrum* and the mel-warped cepstrum is that both features come from the transform or model that is based on the non-linear human perception of the frequency of sounds.

In this study, the speech waveform is segmented into 16 ms frames that overlap by 8 ms and parameterized to a 14 dimensional feature vectors which establish the feature spaces for mel-cepstrum and *auditory cepstrum*.

2.2.1 Mel-Cepstrum

The cepstrum of a signal is computed by taking the inverse Fourier transform of the log magnitude of the signal spectrum.

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

The inverse Fourier transform is identical to Fourier transform within a multiplicative constant since $log|FFT|$ is real and symmetric. From the definition, the cepstrum can be considered as performing the frequency analysis on the magnitude of log spectrum. Hence, the lower order cepstral coefficients preserve the spectral envelope, the overall shape of the log spectrum, which contains important vocal tract shape information for speaker identification applications. The mel-warped cepstrum is obtained by transforming the linear frequency scale to

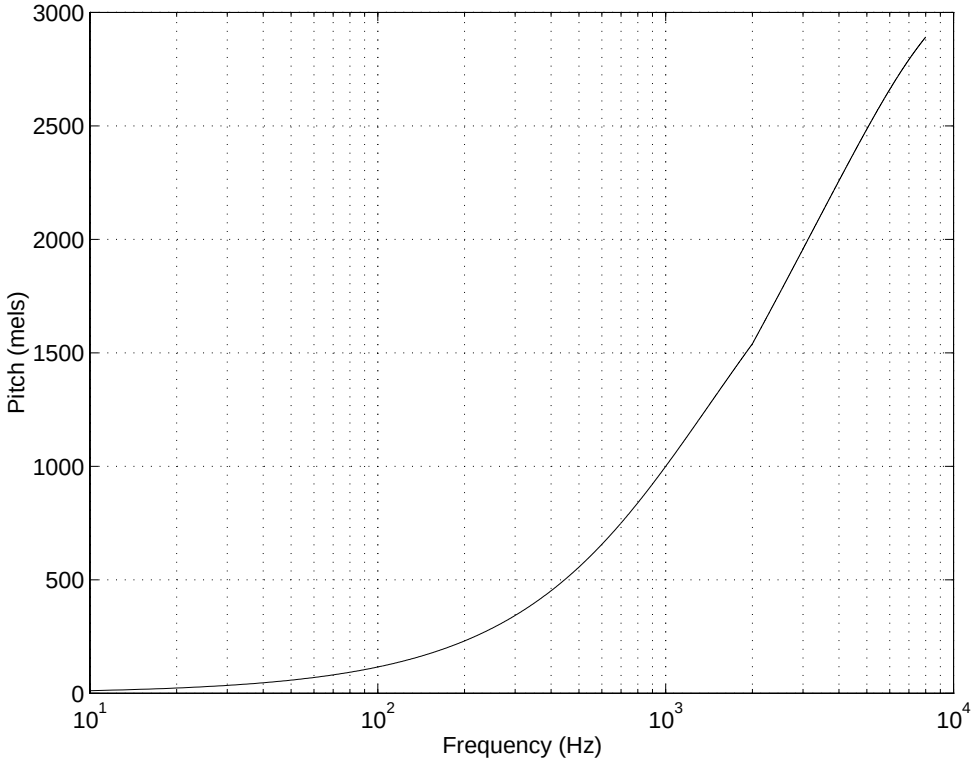

Figure 2.2: The subjective pitch in mel scale versus in logarithmic frequency scale.

mel-frequency scale which place less emphasis on high frequencies. The curve in Figure 2.2 shows the subjective pitch in mels as a function of the logarithmic frequency in Hz, and it indicates the human perception of the frequency content of sounds is linear with the logarithmic frequency beyond about 800 Hz [20]. This mel-frequency response can be thought of the spectral estimation of human auditory perception.

$$mel \Leftrightarrow cepstrum(frame) = FFT^{-1}[mel(log|FFT(frame)|)]$$

2.2.2 *Auditory Cepstrum*

As mentioned in Section 1.2, the *auditory spectrum* comes from the peripheral auditory model that can be divided into three stages: analysis, transduction and reduction. From the signal processing viewpoint, this early auditory system can be functionally described as an affine wavelet transform coupled with nonlinear compression and reduction processes. The characteristic of human hearing, linear along a logarithmic frequency above about 800 Hz, sets the transfer functions of the "cochlear filters" related to each other by a constant Q-factor translation or a dilation between the impulse responses of these filters. Therefore, an affine wavelet transform is fitted to interpret this spectral decomposition in the cochlea that converts a purely time-varying signal to a spatially distributed pattern of activity along the cochlea [16].

In practice, the spatial axis of the cochlea, which is also interpretable as the *scale* parameter axis, is discretized into finite number of channels. However, a tradeoff exists among numbers of channels, the frequency range covered by the model and the frequency resolution of the model [24]. In this work, 96 channels that cover frequency band from 250 Hz to 6.7 KHz with 20 channels per octave density are used to model the cochlear filter banks. The magnitude response of one cochlear filter with center frequency around 1 KHz is shown in Figure 2.3 (a). Part (b) shows the transfer functions of 20 cochlear filters appear approximately invariant except for a translation.

At the final output of the reduction stage, a representation that approximately reflects a short-time spectral profile of the signal is obtained and referred to as *auditory spectrum* of the speech signal. In a similar sense of defining cepstrum as the frequency analysis of the logarithm of the power spectrum,*auditory*

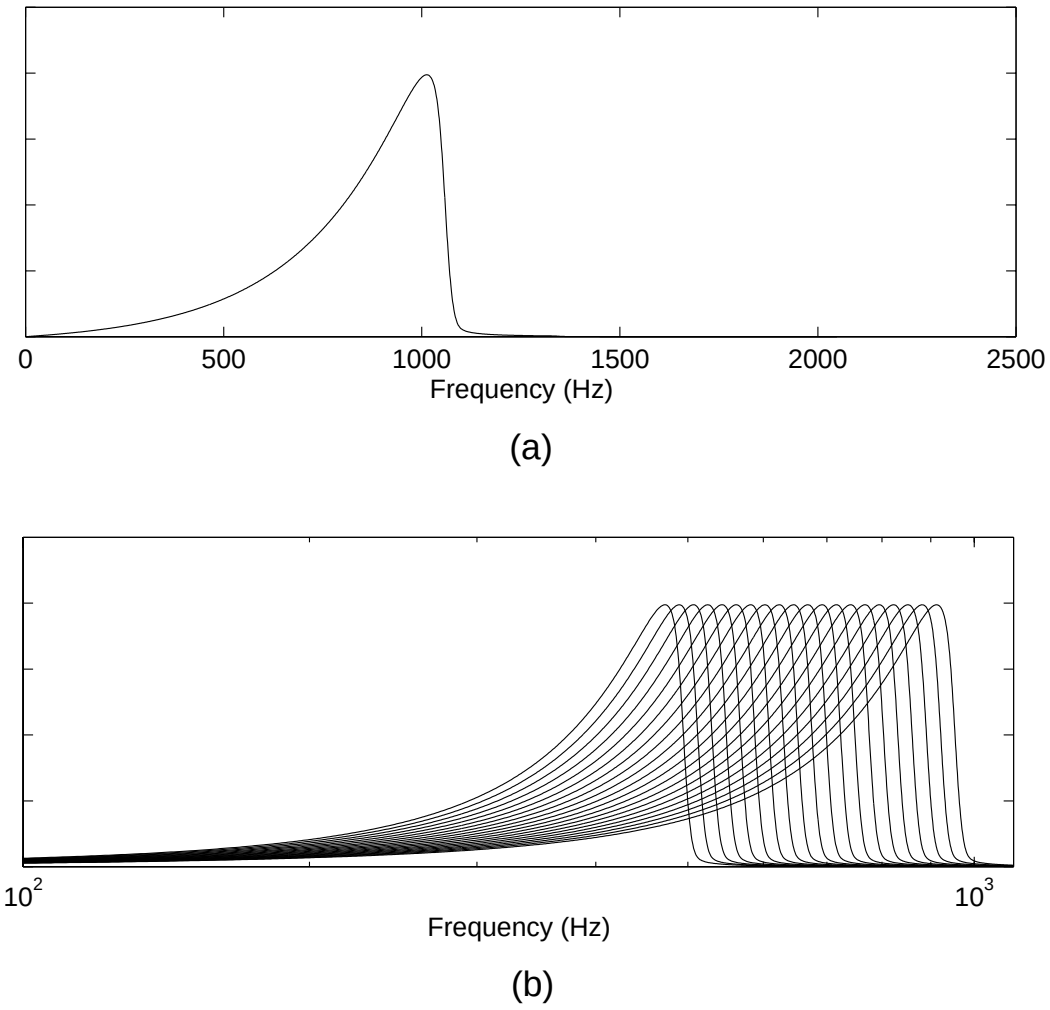

Figure 2.3: Magnitude responses of cochlear filters. (a) One cochlear filter with center frequency around 1 KHz. (b) 20 cochlear filters uniformly distributed in one octave.

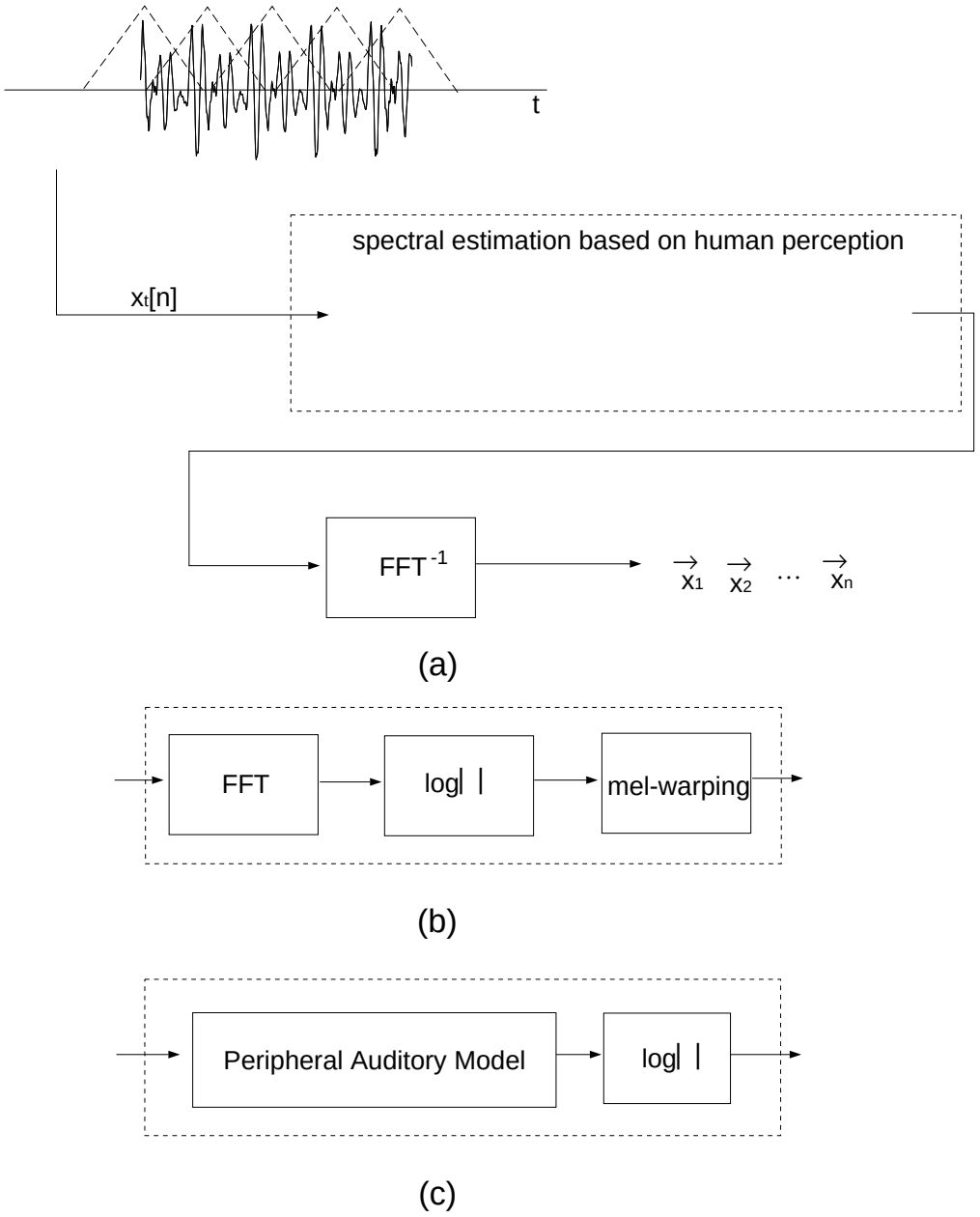

Figure 2.4: Schematic description of the cepstrum feature extraction module for speaker recognition application. (a) General steps for extracting cepstral coefficients. (b) For traditional mel-cepstrum features. (c) Modified steps for *auditory cepstrum* features

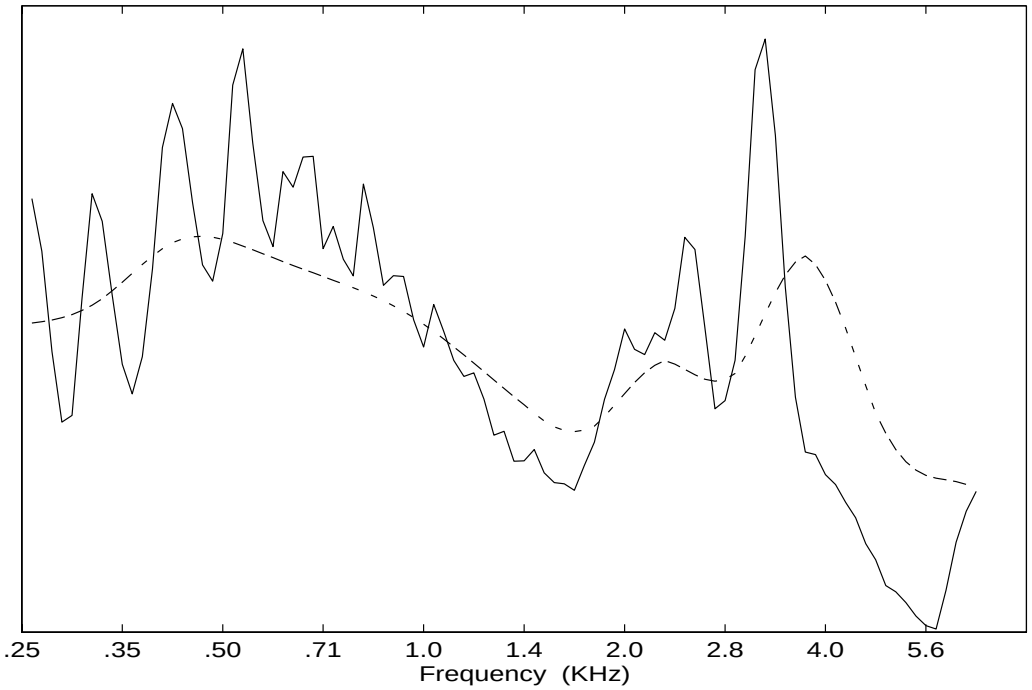

Figure 2.5: Original (solid line) and smoothed auditory spectrum (dashed line) for the utterance 'Come home right away.'.

cepstrum can be defined as the inverse Fourier transform of the logarithm of the *auditory spectrum*. The general procedures for extracting these two cepstral features are outlined in Figure 2.4 (a), and part (b),(c) describe the modified individual steps for mel-cepstrum and *auditory cepstrum*. The first few, 14 in this work, lower order cepstral coefficients of these two cepstra are retained as features and the performance of these two feature spaces for speaker identification task is studied in Section 2.4. Figure 2.5 shows the original average auditory spectrum (solid line) and the smoothed one (dashed line) from 14 auditory cepstral coefficients for the utterance 'Come home right away.'. Obviously, the smoothed one retains the shape information of the overall auditory spectrum.

2.3 Pattern Classification

Feature vectors produced by individual speakers are often assumed to be samples from a continuous probability distribution. The distributions of different speakers may overlap but share the feature space and have to be ideally distinguishable from each other so that identifying the speakers is achievable. In practice, it is usually assumed that the feature vectors are independent of one another to simplify matters, even the vectors are from consecutive frames which are correlated in reality. In this section, a well-known Bayes classifier is introduced and applied to the multivariate Gaussian distributed cepstral features for speaker identification task.

2.3.1 Bayes Classifier

A classifier for discriminating the speakers is built to find an optimal decision rule that minimizes the average risk (cost). The *minimum-risk decision rule* is implemented in terms of a class-conditional probability $p(\mathbf{x}|i)$ for the ith class, a corresponding *a priori* class probability $p(i)$ and loss L_{ij} assigned when pattern $\mathbf{x}$ is decided to be from class j but actually from class i, $i = 1, 2, \ldots, M$, where M is the number of classes. Since pattern $\mathbf{x}$ may belong to any of the M classes, the expected cost when assigns observation $\mathbf{x}$ to class j is given by

$$r_j(\mathbf{x}) = \sum_{i=1}^{M} L_{ij} p(i|\mathbf{x}) \tag{2.1}$$

where $p(i|\mathbf{x})$ means the probability that $\mathbf{x}$ comes from class i. This classifier has M possible categories to choose for each pattern $\mathbf{x}$. If it computes the quantities $r_1(\mathbf{x}), r_2(\mathbf{x}), \ldots, r_M(\mathbf{x})$, for each $\mathbf{x}$, and assigns each pattern $\mathbf{x}$ to the class which has the smallest loss, the total expected loss with respect to all decisions will

also be obviously minimized. The classifier which minimizes the total expected loss is called the *Bayes classifier*.

Considering Bayes' formula,

$$p(i|\mathbf{x}) = \frac{p(i)p(\mathbf{x}|i)}{p(\mathbf{x})} \tag{2.2}$$

one may express Eq. 2.1 in the following form:

$$r_j(\mathbf{x}) = \frac{1}{p(\mathbf{x})} \sum_{i=1}^{M} L_{ij} p(\mathbf{x}|i)p(i) \tag{2.3}$$

where $p(\mathbf{x}|i)$ is called the *likelihood function* of class i. The expression for the average loss can be reduced to

$$r_j(\mathbf{x}) = \sum_{i=1}^{M} L_{ij} p(\mathbf{x}|i)p(i) \tag{2.4}$$

by dropping the common factor $\frac{1}{p(\mathbf{X})}$ for all j in Eq. 2.3.

For simple binary test($M = 2$), if hypothesis 1 is chosen, then

$$r_1(\mathbf{x}) = L_{11} p(\mathbf{x}|1)p(1) + L_{21} p(\mathbf{x}|2)p(2) \tag{2.5}$$

and if hypothesis 2 is chosen,

$$r_2(\mathbf{x}) = L_{12} p(\mathbf{x}|1)p(1) + L_{22} p(\mathbf{x}|2)p(2) \tag{2.6}$$

As mentioned above, the classifier will assign a pattern $\mathbf{x}$ to the class with the lower value of r. Thus, $\mathbf{x}$ is assigned to class 1 if $r_1(\mathbf{x}) < r_2(\mathbf{x})$; in other words,

$$L_{11} p(\mathbf{x}|1)p(1) + L_{21} p(\mathbf{x}|2)p(2) < L_{12} p(\mathbf{x}|1)p(1) + L_{22} p(\mathbf{x}|2)p(2) \tag{2.7}$$

or equivalently,

$$(L_{21} \Leftrightarrow L_{22}) p(\mathbf{x}|2)p(2) < (L_{12} \Leftrightarrow L_{11}) p(\mathbf{x}|1)p(1) \tag{2.8}$$

Eq. 2.8 can be rearranged as

$$\frac{p(\mathbf{x}|1)}{p(\mathbf{x}|2)} > \frac{p(2)}{p(1)} \frac{(L_{21} \Leftrightarrow L_{22})}{(L_{12} \Leftrightarrow L_{11})} \tag{2.9}$$

The left-hand side term in Eq. 2.9 is defined as the *likelihood ratio*:

$$l_{12}(\mathbf{x}) = \frac{p(\mathbf{x}|1)}{p(\mathbf{x}|2)} \tag{2.10}$$

which is the ratio of two likelihood functions. Hence, the Bayes decision rule for $M = 2$ is as follows:

1. Assign $\mathbf{x}$ to class 1 if $l_{12}(\mathbf{x}) > \theta_{12}$.
2. Assign $\mathbf{x}$ to class 2 if $l_{12}(\mathbf{x}) < \theta_{12}$.
3. Make an arbitrary decision if $l_{12}(\mathbf{x}) = \theta_{12}$.

where θ_{12} is called the threshold value which is given by

$$\theta_{12} = \frac{p(2)}{p(1)} \frac{(L_{21} \Leftrightarrow L_{22})}{(L_{12} \Leftrightarrow L_{11})} \tag{2.11}$$

If a loss of 1 is assigned for an incorrect decision and a loss of zero is assigned for a correct decision, that is

$$L_{ij} = 1, for\ i \neq j$$

$$L_{ij} = 0, for\ i = j$$

the minimum-risk decision rule reduces to the *minimum-probability-of-error decision rule*. Under this rule, Eq. 2.5 and Eq. 2.6 reduce to

$$r_1(\mathbf{x}) = p(\mathbf{x}|2)p(2) \tag{2.12}$$

and

$$r_2(\mathbf{x}) = p(\mathbf{x}|1)p(1) \tag{2.13}$$

and the criterion in Eq. 2.9 becomes

$$\frac{p(\mathbf{x}|1)}{p(\mathbf{x}|2)} > \frac{p(2)}{p(1)} \tag{2.14}$$

Therefore, the decision rule assigns $\mathbf{x}$ to class 1 if

$$p(1)p(\mathbf{x}|1) > p(2)p(\mathbf{x}|2) \tag{2.15}$$

the so-called *maximum a posteriori* (MAP) compute.

It is easy to extend Eq. 2.15 to M-ary hypothesis testing problem [10]. For a minimum average probability of error decision for pattern $\mathbf{x}$, decide category j if and only if

$$p(j)p(\mathbf{x}|j) = \max_{1 \leq i \leq M} \{p(i)p(\mathbf{x}|i)\} \tag{2.16}$$

A schematic description of a Bayes classifier for M-ary hypothesis testing problem is shown in Figure 2.6.

2.3.2 Gaussian Model

In practice, it is always assumed that the probability density functions $p(\mathbf{x}|i)$ are multivariate Gaussian (normal) to simplify analysis. Because of its analytical tractability, the multivariate Gaussian density function has received considerable attention and represented an appropriate model for many important practical applications. Consider M pattern classes governed by the multivariate Gaussian models with parameters mean vector μ_j and covariance matrix Σ_j for $j = 1, 2, \ldots, M$

$$p(\mathbf{x}|j) = \frac{1}{(2\pi)^{1/2}|\Sigma_j|^{1/2}} \exp[\Leftrightarrow\frac{1}{2}(\mathbf{x} \Leftrightarrow\mu_j)'\Sigma_j^{-1}(\mathbf{x} \Leftrightarrow\mu_j)] \tag{2.17}$$

where μ_j and Σ_j are defined as

$$\mu_j = E_j[\mathbf{x}]$$

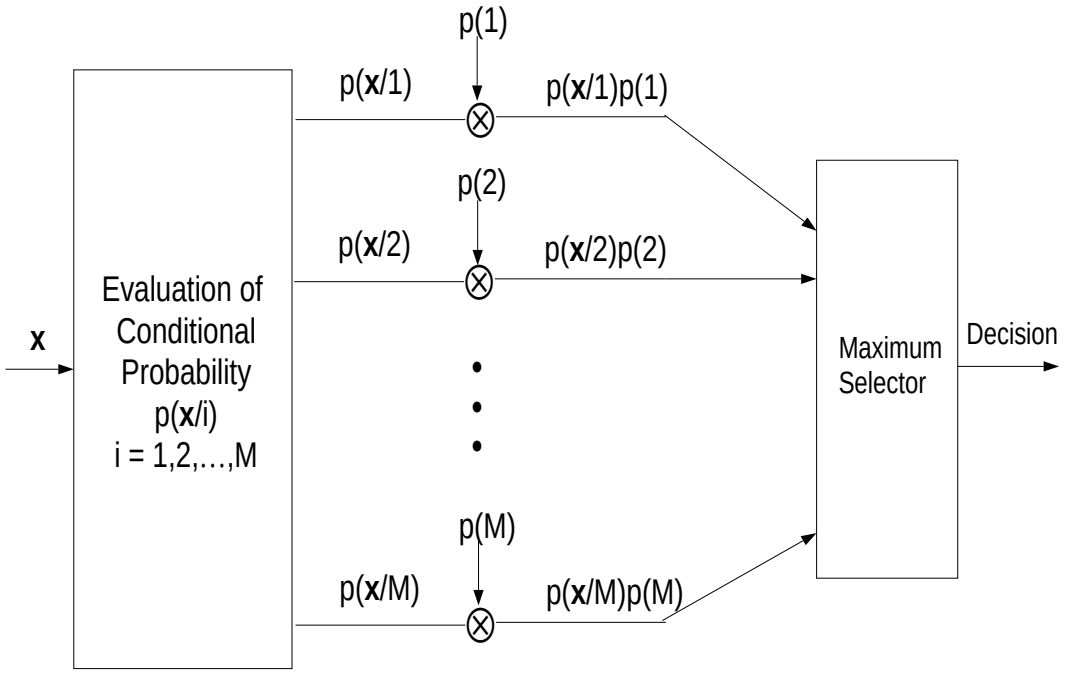

Figure 2.6: A Bayes classifier for M-ary hypothesis testing problem.

and

$$\Sigma_j = E_j[(\mathbf{x} \Leftrightarrow \mu_j)(\mathbf{x} \Leftrightarrow \mu_j)']$$

The likelihood of a test utterance consisting of n *independent* feature vectors, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ for Gaussian model (μ_j, Σ_j) is given by

$$\begin{aligned} L(\mathbf{X}; j) = \quad p(\mathbf{X}|j) \quad &= \prod_{i=1}^{n} \{ \frac{1}{(2\pi)^{1/2} |\Sigma_j|^{1/2}} \exp[\Leftrightarrow \frac{1}{2} (\mathbf{x}_i \Leftrightarrow \mu_j)' \Sigma_j^{-1} (\mathbf{x}_i \Leftrightarrow \mu_j)] \} \\ &= |2\pi\Sigma_j|^{-n/2} \exp[\Leftrightarrow \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i \Leftrightarrow \mu_j)' \Sigma_j^{-1} (\mathbf{x}_i \Leftrightarrow \mu_j)] \end{aligned} \quad (2.18)$$

where $|2\pi\Sigma_j|$ means the determinant of $2\pi\Sigma_j$. The log likelihoods are much more convenient for computation,

$$l(\mathbf{X}; j) = \log[L(\mathbf{X}; j)] = \Leftrightarrow \frac{n}{2} \log |2\pi\Sigma_j| \Leftrightarrow \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i \Leftrightarrow \mu_j)' \Sigma_j^{-1} (\mathbf{x}_i \Leftrightarrow \mu_j) \quad (2.19)$$

Equivalently,

$$l(\mathbf{X}; j) = \Leftrightarrow \frac{1}{2} \log |2\pi\Sigma_j| \Leftrightarrow \frac{n}{2} tr(\Sigma_j^{-1} S) \Leftrightarrow \frac{1}{2} (\bar{\mathbf{x}} \Leftrightarrow \mu_j)' \Sigma_j^{-1} (\bar{\mathbf{x}} \Leftrightarrow \mu_j) \quad (2.20)$$

where S and $\bar{\mathbf{x}}$ are the covariance and mean of the test utterance.

The multivariate Gaussian distribution is completely characterized by its mean μ and covariance matrix Σ such that make the applications of class discrimination with low computation complexity. Figure 2.7 exposes the motivation of applying the Gaussian model to pattern recognition tasks. The large ellipses in Figure 2.7 denote the contours of constant density of two speakers' Gaussian models. These contours are determined by the eigenvectors and eigenvalues of the covariance matrices Σ_1, Σ_2 [9] which are estimated from the training vectors. The small ellipsoidal clusters represent each training utterance of two speakers. The ultimate goal is to employ the log likelihood function in Eq. 2.20 to measure the 'distance' between test utterance which is parameterized by $(\bar{\mathbf{x}}, S)$ and each speaker model which is parameterized by (μ_j, Σ_j) for $j = 1, 2, \ldots, M$.

2.4 Experimental Evaluation

This section presents the experimental evaluation of the Gaussian speaker model on mel-cepstrum and *auditory cepstrum* features for text-independent speaker identification. The performance of the Gaussian model is examined under the *equally likely* condition, i.e., the *a priori* probability of every speaker is assumed to be equal not biased. Under this *equally likely* assumption, i.e.,$p(j) = 1/M$ for $j = 1, 2, \ldots, M$, Eq. 2.16 simplifies to

$$p(\mathbf{X}|j) = \max_{1 \le i \le M} \{p(\mathbf{X}|i)\} \quad iff \quad d(\mathbf{X}) = j \tag{2.21}$$

where $d(\cdot)$ means a decision function, or equivalently,

$$Speaker \;\; Index = \arg \max_{1 \le i \le M} \{l(\mathbf{X}; i)\} \tag{2.22}$$

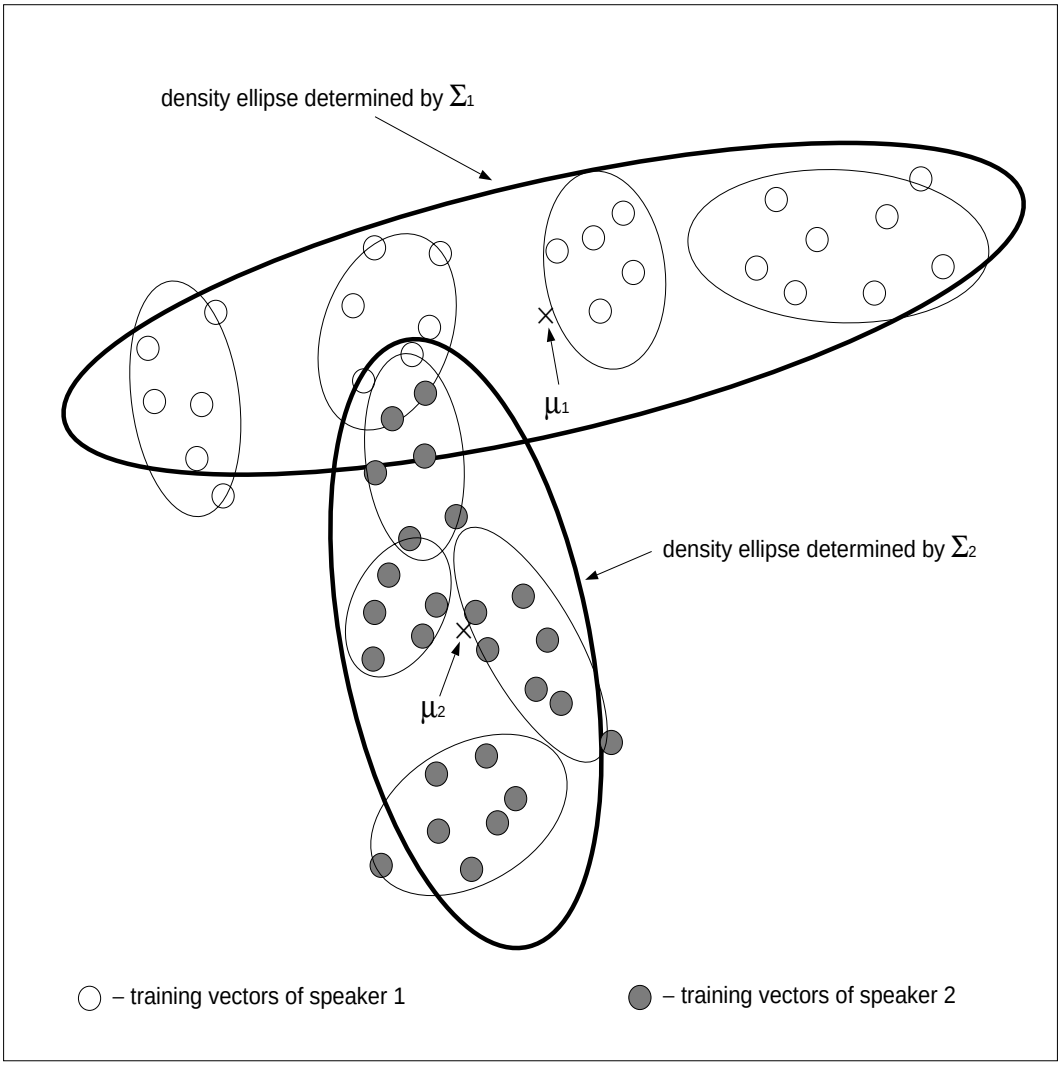

Figure 2.7: Gaussian models for 2 speakers. The means determine the location and covariance matrices estimate the shape of Gaussian distribution. The large and small ellipses show the contours of speakers' Gaussian models and each training utterance respectively.

All the speech data involved in this experiment are spoken by 23 female speakers from training division in dialect region 2 of the TIMIT database. Seven out of ten utterances for each speaker are used to build the Gaussian speaker model during training session and the other three utterances are used to test the model. In this work, the effect of noise on identification performance is investigated for different signal-to-noise ratio.

2.4.1 Model Training

The idea of training Gaussian models is to estimate the sufficient statistics (μ_i, Σ_i) for $i = 1, 2, \ldots, 23$ from the corresponding seven utterances (total of approximately 20 sec) of each speaker. Meanwhile, the sufficient statistics $(\bar{\mathbf{x}}_j, S_j)$ for $j = 1, 2, \ldots, 161$ of all training utterances are recorded to evaluate trained speaker models. Figure 2.8 shows the relative locations of the first two principal components of the mean vectors in the cepstral feature spaces for all 23 female speakers. To give these two dimensional figures, the principal component analysis is applied to the original 14 dimensional cepstrum vectors to compress the data into first two principal components. Intuitively, these locations should be far away from one another to effectively represent each speaker.

The models for auditory cepstrum and mel-cepstrum are separately demonstrated on the upper and lower part of this figure. Each number from 1 to 23 represents each speaker. Since cepstral features are assumed to be Gaussian distributed, the Euclidean distance, distance between two points in Figure 2.8 without normalization by individual covariance matrix, is not the same as the distance measured by Bayes classifier. However, one can still catch some idea and insight by observing Figure 2.8. For example, speaker 9 and 14 are somewhat

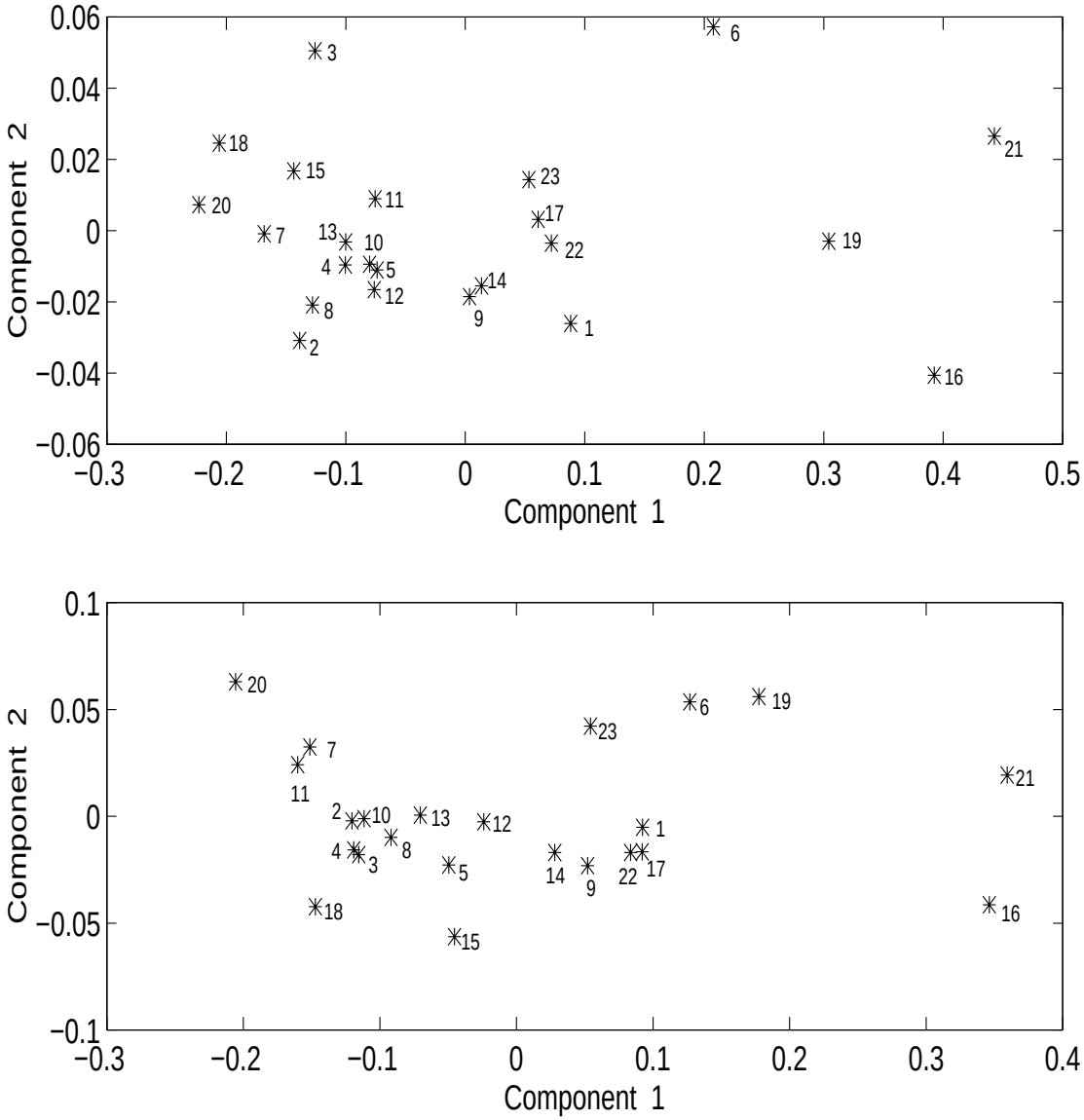

Figure 2.8: First two principal components of average auditory cepstral features (shown in upper part) and average mel-cepstral features (shown in lower part) for each speaker. Each number represents each speaker.

similar in their vocal tract shapes when talking, so is speaker 17 and 22. On the other hand, speaker 16, 19 and 21 that are highly separated on the 'speaker map' are believed in being identified without ambiguity. These deductions are consistent in both auditory cepstrum and mel-cepstrum spaces.

To evaluate the Gaussian models for all speakers, total 161 training utterances (7 for each speaker) are employed to test the models by using the Bayes maximum likelihood classifier stated in Section 2.3. The correct identification rate that is defined as

$$\%\ \ identification\ \ rate = \frac{\#\ of\ correctly\ identified\ utterances}{total\ \#\ of\ utterances} \times 100 \quad (2.23)$$

for both *auditory cepstrum* and mel-cepstrum is 100%. The confusion matrices for both cepstral features are the same and the matrix for *auditory cepstrum* is given in Figure 2.9. Each row of the confusion matrix represents the speaker who is speaking and the columns are the responses of the classifier to utterances spoken by each speaker. In other words, each element (i, j) in this matrix represents the number of utterances that are spoken by speaker i but recognized as by speaker j. Clearly, the sum of each row of the matrix is equal to the total utterances spoken by that speaker.

2.4.2 Robustness Performance

The ability of each cepstral representation to discriminate speakers under different noise level are measured by using the remaining three utterances for each speaker. In this work, Gaussian white noise is used as additive noise to simulate different noise level for comparison. The AWGN (additive white Gaussian noise)

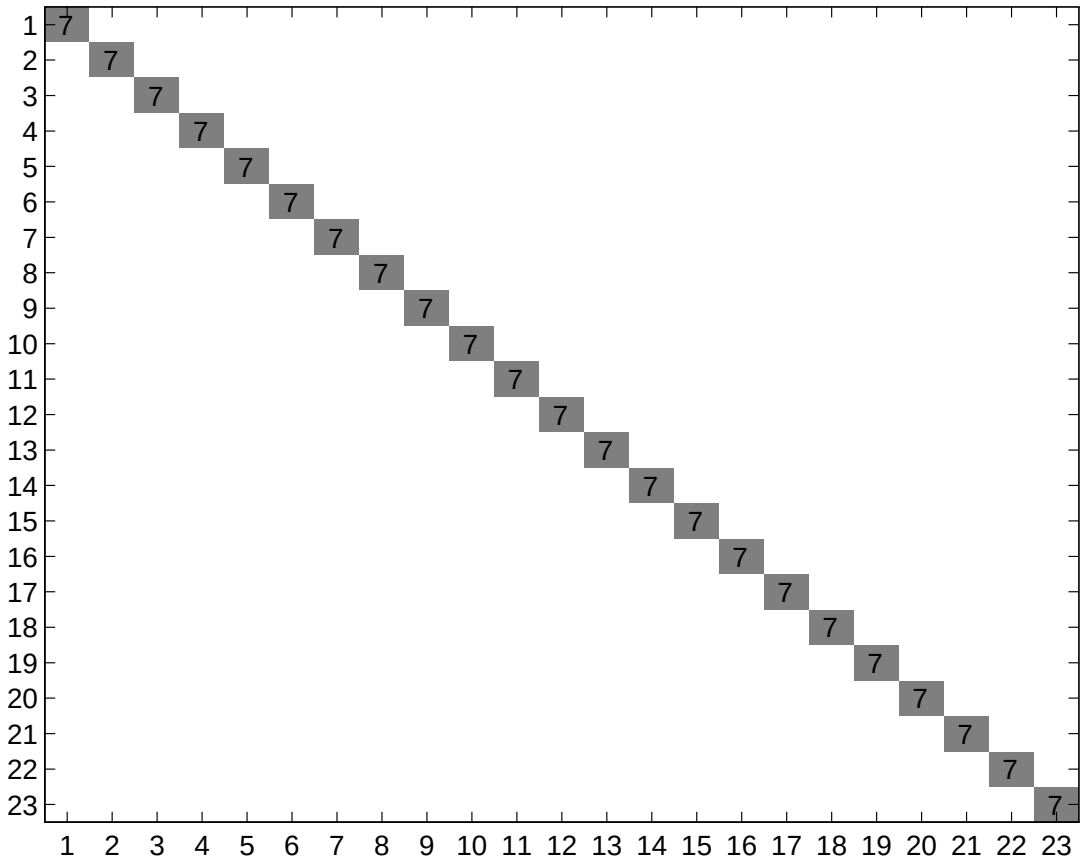

Figure 2.9: Confusion matrix for auditory cepstrum by using the training data for testing. The elements in this matrix indicate the number of utterances.

is added into the clear speech signal using the following formula [7]:

$$SNR = 10\log(\frac{\sum_{n=1}^{T} S(n)^2}{\sum_{n=1}^{T} N(n)^2}) \quad [dB] \tag{2.24}$$

where T is the number of samples in speech and noise data, $S(n)$ and $N(n)$ are sampled clean speech and noise data. The signal-to-noise ratios investigated here are 0dB $\Leftrightarrow$24dB in step of 3dB. The confusion matrices for auditory cepstrum and mel-cepstrum under 12dB $\Leftrightarrow$24dB SNR are respectively displayed in Figure 2.10 $\Leftrightarrow$Figure 2.15. In these figures, all the upper matrices are for auditory cepstrum and the lower ones are for mel-cepstrum. In addition, different gray levels in these matrices are designed to represent different number of utterances.

Since the diagonal elements of the confusion matrices represent the number of correctly identified utterances, one can conclude that the more dominant the diagonal of the matrix, the higher the recognition rate of the identification system. By observing the first two confusion matrices obtained under 12dB and 15 dB SNR for both cepstral features in Figure 2.10 and Figure 2.11, one can see that the upper confusion matrices for auditory cepstrum begin to diagonalize while the lower ones for mel-cepstrum still stay the similar format. It is quite interesting that most of the utterances are recognized as spoken by speaker 16 for both cepstral features in high noise level (0-15dB). A simple inspection on the individual spectrum reconstructed from cepstral coefficients gives some intuition in explaining the fact. Figure 2.12 shows the spectral templates that are reconstructed from mel-cepstrum for each speaker. Each average template is plotted along tonotopic frequency axis and the speaker index is given in each sub-plot. It is easy to observe that speaker 16 and 5 possess more smoothed spectra which are more similar to the spectrum of the white Gaussian noise than other spectra. Additionally, inspection on the spectrum reconstructed from the

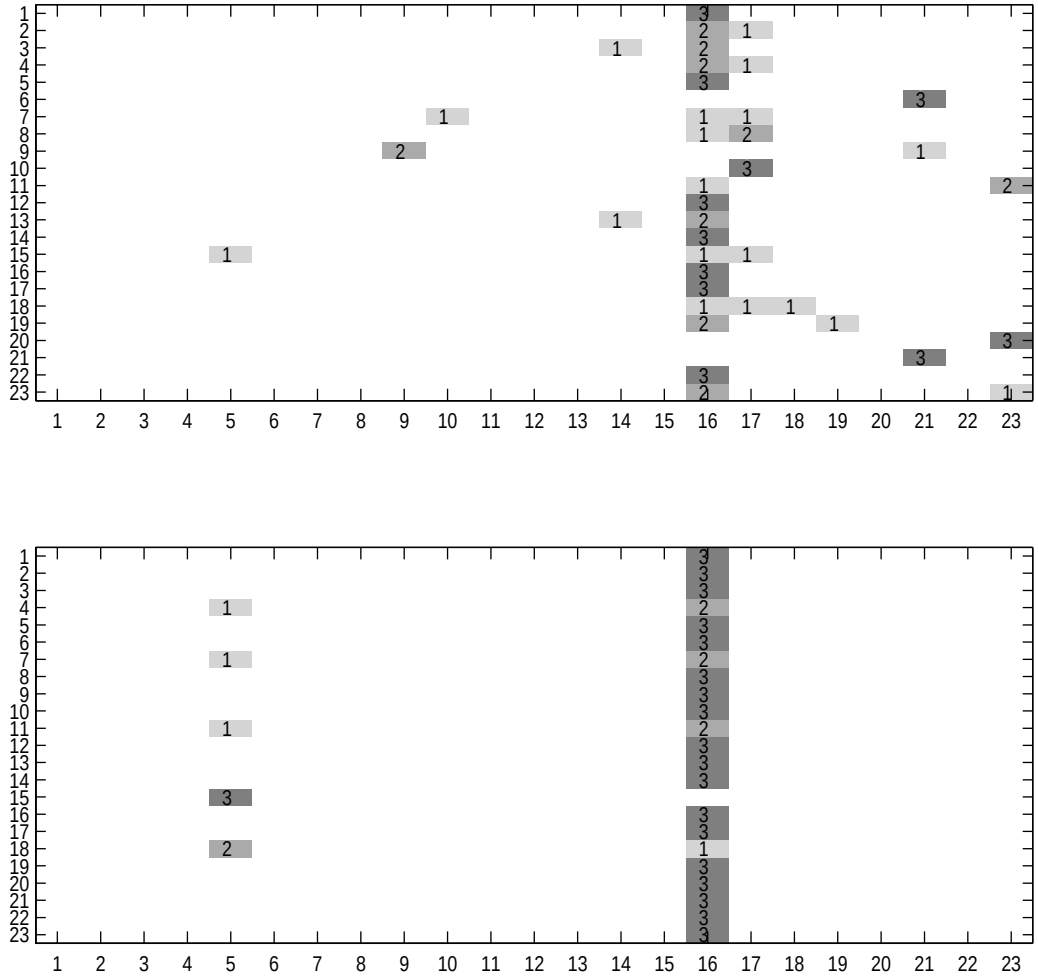

Figure 2.10: Confusion matrix for auditory cepstrum (upper part) and mel-cepstrum (lower part) under 12 dB SNR.

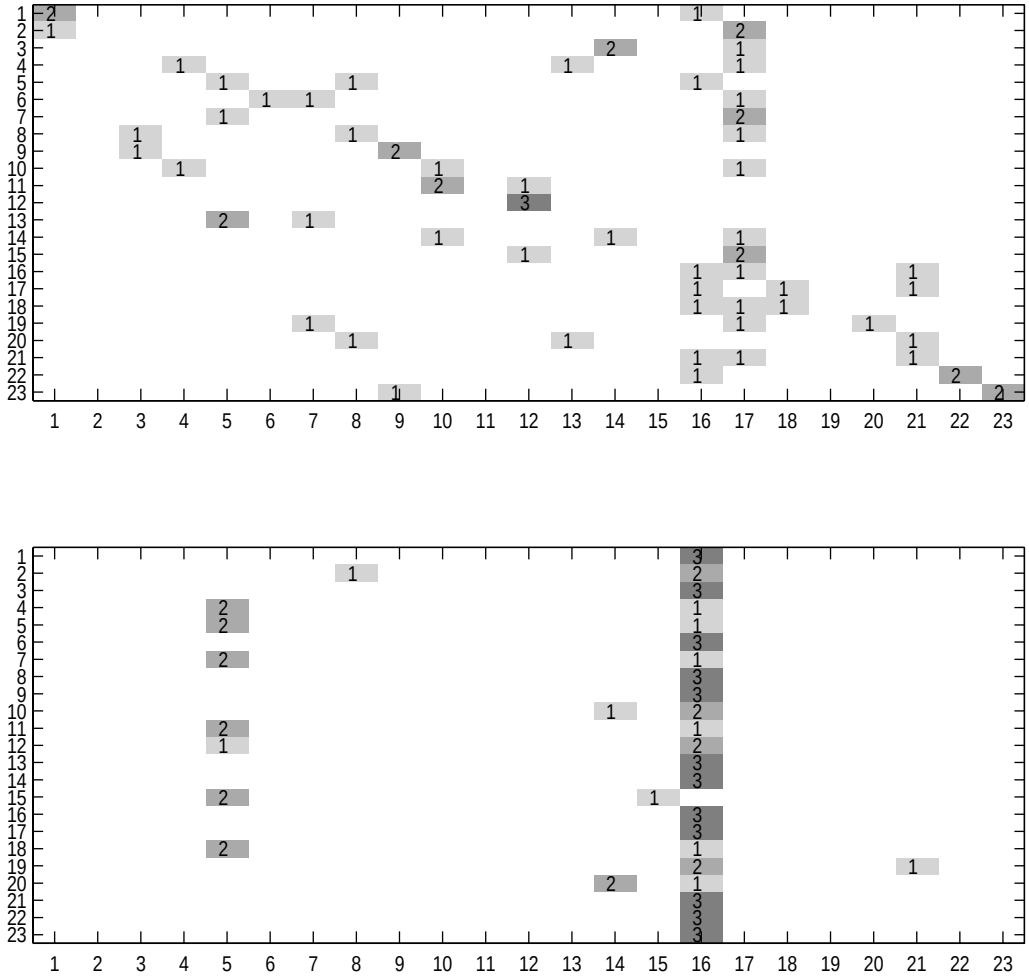

Figure 2.11: Confusion matrix for auditory cepstrum (upper part) and mel-cepstrum (lower part) under 15 dB SNR.

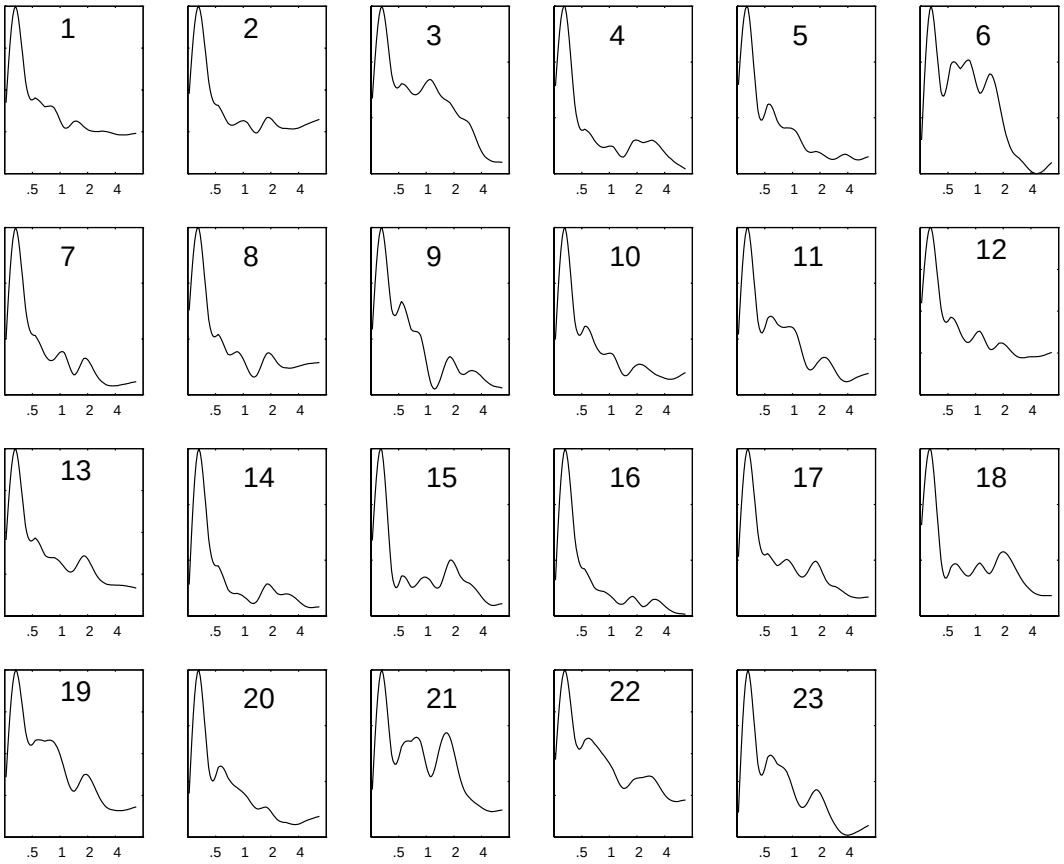

Figure 2.12: Reconstructed spectra from 14 average mel-cepstral coefficients for 23 female speakers. Each spectrum is plotted along tonotopic frequency axis which is given in KHz.

auditory cepstrum for each speaker yields the similar observations. To clarify the occurrence, however, one has to investigate the mean vector ,the eigenvalues and the eigenvectors of the covariance matrix of cepstral feature for each speaker.

By following Figure 2.13, Figure 2.14 to Figure 2.15, one can see the confusion matrices for mel-cepstrum features become diagonal dominant around 21dB which is 6dB higher than the SNR needed for auditory cepstrum to achieve almost the same identification rate. In other words, auditory cepstrum will yield

SNR	Auditory cepstrum	Mel-cepstrum
12dB	11/69	3/69
15dB	20/69	6/69
18dB	29/69	9/69
21dB	33/69	24/69
24dB	40/69	39/69

Table 2.1: Identification rate for auditory cepstrum and mel-cepstrum of 69 test utterances.

higher identification rate than mel-cepstrum under the same SNR condition.

In addition, the confusion matrix $\mathbf{C}$ also provides a simple way in computing the identification rate by $\frac{tr(\mathbf{C})}{3\times 23}$; where the denominator 3×23 is the total number of test utterances in this experiment. Moreover, the probability of two types of errors,*miss* and *false alarm*, for each speaker $j, j = 1, 2, \ldots, 23$ can also be obtained respectively from the row and column which correspond to speaker j in confusion matrix by following two formulas:

$$P_m(j) = pr(d \neq j | H = j) = \frac{3 \Leftrightarrow \mathbf{C}_{jj}}{3}$$

$$P_{fa}(j) = pr(d = j | H \neq j) = \frac{\sum_{i;i\neq j}^{23} \mathbf{C}_{ij}}{3 \times 22}$$

where H means hypothesis, d is decision function and $pr(\cdot)$ means probability measure. These two measurements are essential to be investigated for determining the threshold for every speaker in speaker verification problem which is not the considered case in in this work.

The identification performance for auditory cepstrum and mel-cepstrum about 12dB $\Leftrightarrow$ 24dB SNR is presented in Table 2.1. The maximum difference of the identification rate between these two cepstral features is about 30% under the

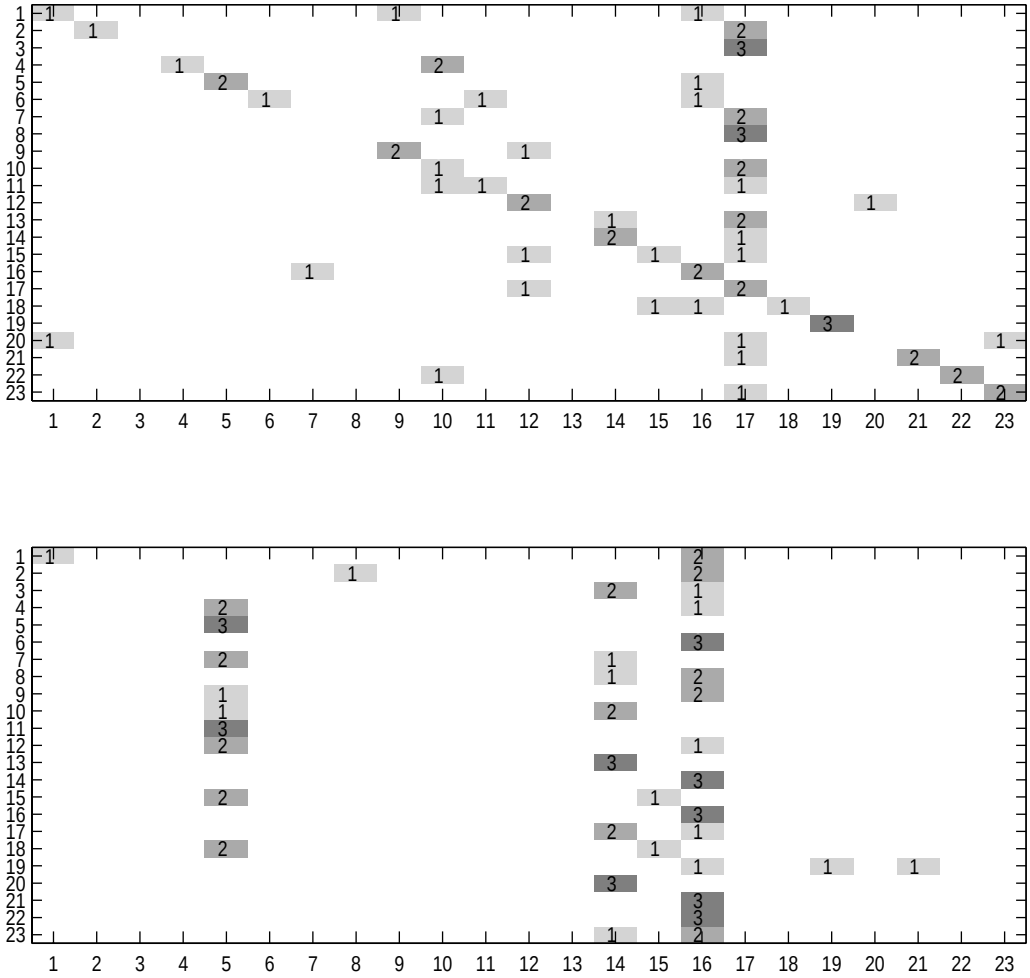

Figure 2.13: Confusion matrix for auditory cepstrum (upper part) and mel-cepstrum (lower part) under 18 dB SNR.

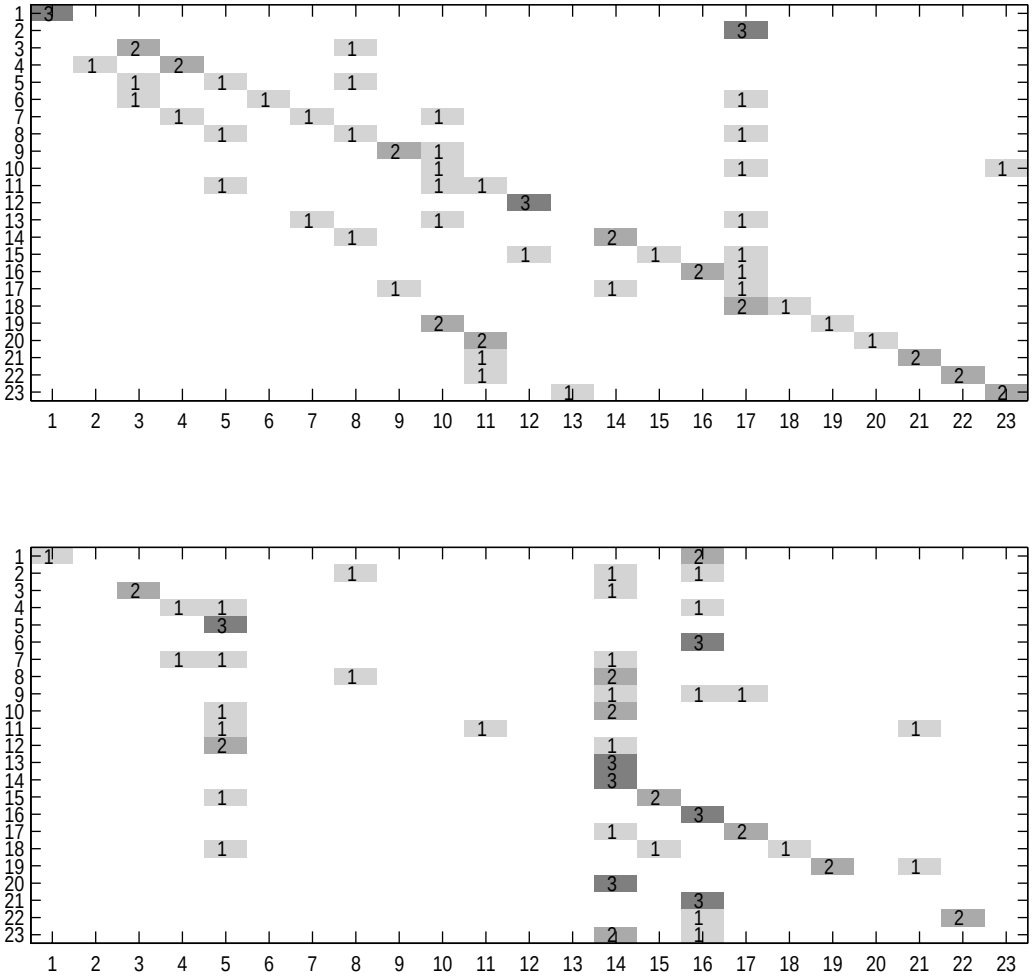

Figure 2.14: Confusion matrix for auditory cepstrum (upper part) and mel-cepstrum (lower part) under 21 dB SNR.

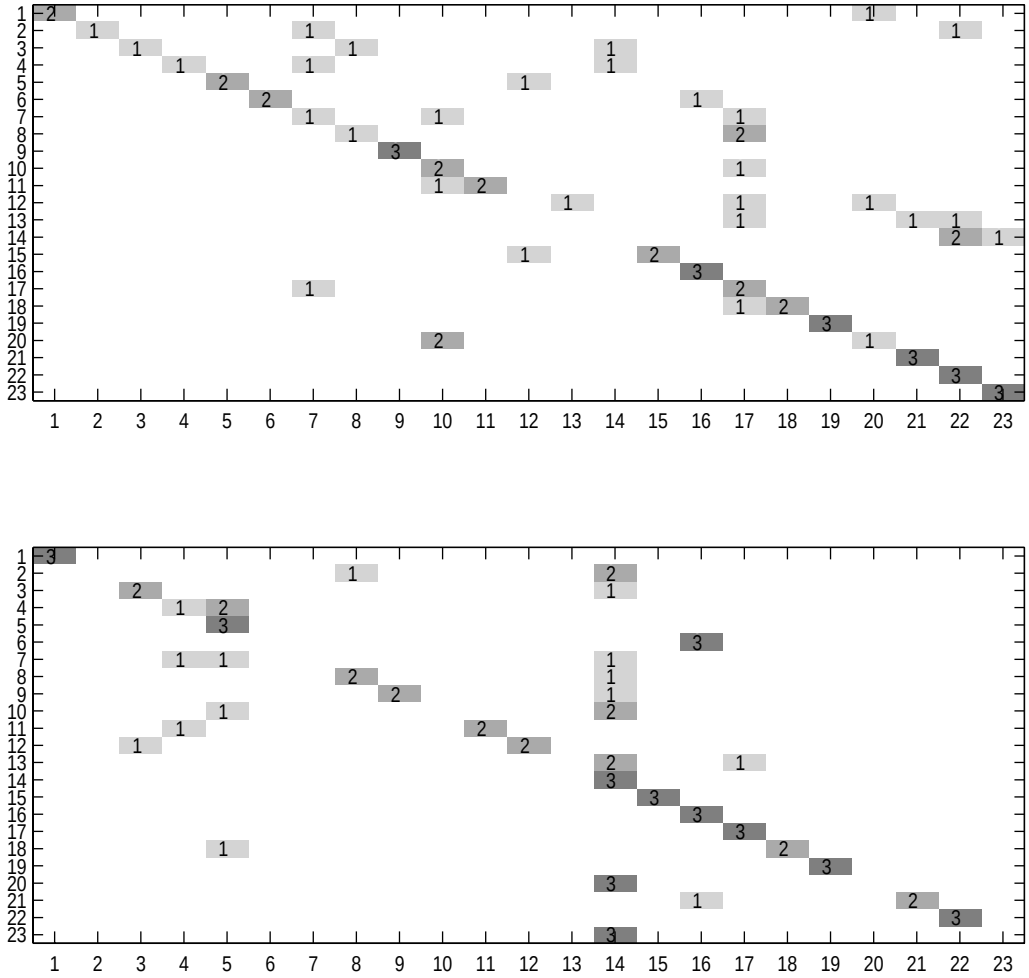

Figure 2.15: Confusion matrix for auditory cepstrum (upper part) and mel-cepstrum (lower part) under 24 dB SNR.

18dB SNR condition. In other words, under 18dB SNR environment which is reasonable assumption for telephone line, the auditory cepstrum yields much better performance than mel-cepstrum for the speaker identification application.

2.5 Summaries and Remarks

In this chapter, a speaker identification experiment is evaluated for the *auditory cepstrum* features which is defined as the inverse Fourier transform of the auditory spectrum and the well-known mel-cepstrum features. A simple uni-modal Gaussian model with Bayes maximum likelihood algorithm is used to perform the speaker identification under different signal-to-noise ratio to examine the robustness characteristics of these two cepstral features.

According to the experimental results shown in Figure 2.16, it is believed that the auditory cepstrum is more robust than mel-cepstrum with SNR ranging from 0dB to 24dB . This conclusion is not surprising since the auditory cepstrum convey the same information as the *auditory spectrum* which is based on human perception model and has been shown to have robust properties in a previous study [3].

As mentioned in Section 2.2.1, the cepstrum based frameworks are often employed for spectral shape analysis. The lower order cepstral coefficients describe the global trend of the spectral shape and the higher order coefficients preserve the 'local' information in the spectrum. Therefore, the higher order coefficients have to be considered to obtain the 'local' spectral characteristics like formant frequencies and harmonics which are advantage in improving the identification rate.

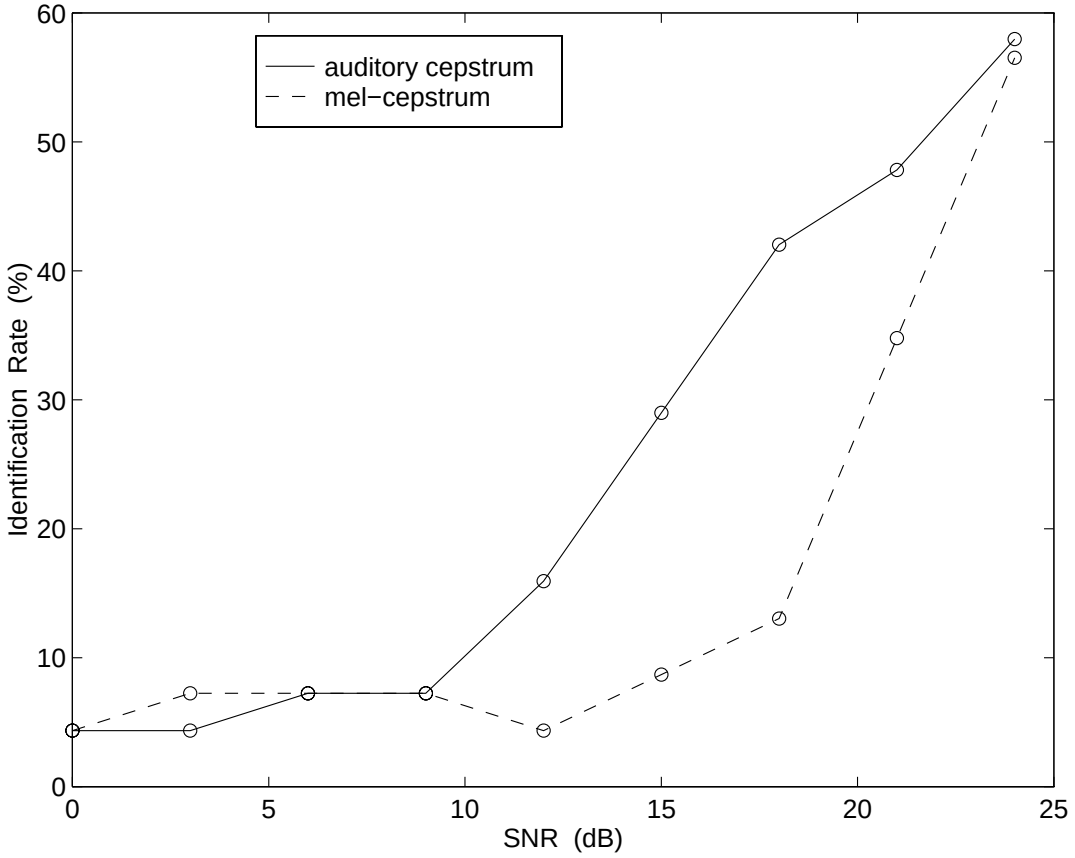

Figure 2.16: Speaker identification rate for auditory cepstrum and mel-cepstrum with respect to 0dB-24dB signal-to-noise ratio.

In recent years, many researchers proposed noise robustness algorithm like Gaussian mixture model for the recognizer level to improve the performance for speech recognition or speaker identification systems [4, 15, 17]. It is believed that a robust feature as *auditory spectrum* that matches more closely with human perception will cost much less than a robust classifier.

Chapter 3

Speaker Identification Using Cortical Representation

3.1 Introduction

A cortical model based on principles discovered in the primary auditory cortex (AI) is formulated for spectral profile analysis at higher auditory stages [27]. Briefly speaking, the feature extracted by the auditory cortex is organized along tonotopic, scale and phase dimensions which correspond to the selectively responsive properties of auditory cells for best tuning frequency, bandwidth and symmetry, respectively. This selectively responsive property of AI cells to different scales and local shapes of the input spectral profile suggests that the function of AI cells can be effectively described as performing an affine complex wavelet transform to the input spectrum. This multiscale cortical representation of the input spectral profile can then be used in various recognition and identification tasks as discussed below.

The multiscale transformation generates a three-dimensional image (called *cortical representation*) for the input spectral profiles (auditory spectrum) in the

tonotopic and scale plane. In this chapter, a simple correlator approach ,which is often used for scene matching application in digital image processing, is applied to measure the distance between test and template cortical representations. To examine the fidelity of cortical representation with this template matching method, the speaker verification task is performed for a particular speaker first. Both the LPC spectrum and auditory spectrum features are investigated with the correlator technique for the purpose of comparison. As in Chapter 2, white noise is added to corrupt the cortical patterns and the noise-robustness performance of this representation is examined for speaker identification application. In addition, the noise-robustness of the phase and magnitude features of cortical representations are demonstrated in Section 3.4.2 and the performance of the more robust phase features are inspected for each scale. Finally, the robustness of the cortical phase representation with the correlator technique is compared to the performance stated in Chapter 2 of the two cepstral features with the Bayes classifier technique.

3.2 Cortical Representation

The cochlea of the inner ear analyzes the complex sound into a tonotopic ordered array of channels tuned to different characteristic frequencies and creates the spectral profile of the sound signal. Such profile is referred to as *auditory spectrum*. This auditory spectrum is further processed in the primary auditory cortex to separate features corresponding to different percepts such as pitch and timbre.

Some relevant physiological results suggest that cortical cells respond to the

input spectral profile by tuning to the same frequency along the so-called isofrequency planes which is perpendicular to the tonotopic axis (Refer to Figure 1.4.). These studies also indicate the bandwidth and symmetry features of the input stimulus at a certain frequency are extracted by the cortical cells along these isofrequency planes. In other words, the cortical cells selectively respond to different scales and local shapes of the input spectral profile. This knowledge provides the insight to describe the neurons function as a bank of filters tuned to different scales, phases and frequencies, which transform the input profile into its ripple domain. It is stated in [23] that a family of functions varying systematically in symmetry can be composed by sinusoidally interpolating a symmetric function $h_s(x)$ and its Hilbert transform $\hat{h}_s(x)$ $(= \frac{1}{\pi}\int_R \frac{h_s(z)}{x-z}dz\ )$:

$$\omega_s(x,\phi) \ = \ h_s(x)\cos\phi \ + \ \hat{h}_s(x)\sin\phi \tag{3.1}$$

The parameter ϕ is referred to as the *symmetry index* to indicate the symmetry of $\omega_s(x,\phi)$, for example, $\omega_s(x,0) = h_s(x)$ is symmetric and $\omega_s(x,\pm\frac{\pi}{2}) = \hat{h}_s(x)$ is antisymmetric.

Assuming linearity of the cortical model, the response to a input spectral envelope $y(x)$ is computed as

$$r_s(x,\phi) \ = \ y(x) *_x \omega_s(x,\phi) \ = \ \int_R y(z)\omega_s(x \Leftrightarrow z,\phi)dz \tag{3.2}$$

Substitute Eq. 3.1 into Eq. 3.2 and express the response $r_s(x,\phi)$ in Fourier domain, one can get

$$r_s(x,\phi) \ = \ a_s(x)\cos(\phi \Leftrightarrow \psi_s(x)) \tag{3.3}$$

where

$$a_s(x) = |< Y(\Omega)e^{j2\pi\Omega x},\ H_s(\Omega)(1 \Leftrightarrow j\ sgn(\Omega)) >| \tag{3.4}$$

$$\psi_s(x) = \arctan \frac{< \Im\{Y e^{j2\pi\Omega x}\},\ H_s(\Omega) sgn(\Omega) >}{< \Re\{Y e^{j2\pi\Omega x}\},\ H_s(\Omega) >} \quad , \tag{3.5}$$

$\Re\{Y\}$,$\Im\{Y\}$ represent the real and imaginary part of Y respectively, and Ω (ripple frequency) is Fourier domain of x (tonotopic axis) [26]. Eq. 3.3 indicates that the cortical response $r_s(x, \phi)$ is always a sinusoid in ϕ when given x, s. It implies that the three-dimensional response can be specified by two two-dimensional functions $a_s(x)$ and $\psi_s(x)$.

By investigating the analytical signal

$$\omega_s(x) \ = \ h_s(x) + j\hat{h}_s(x) \ = \ \omega_s(x, 0) + j\omega_s(x, \pm\frac{\pi}{2}) \quad ,$$

one can easily get the corresponding complex cortical response from Eq. 3.3

$$r_s(x) \ = \ y(x) *_x \omega_s(x) \ = \ a_s(x) e^{j\psi_s(x)} \tag{3.6}$$

From this viewpoint, the cortical processing can be characterized by applying a complex wavelet transform with impulse response $\omega_s(x) \ = \ h_s(x) + j\hat{h}_s(x)$ to the input spectrum $y(x)$. In other words, the cortical processing can be considered as carrying out a windowed frequency analysis to the spatial pattern (auditory spectrum) along the spatial axis.

Figure 3.1 (a) shows the magnitude impulse responses of cortical bandpass filters which tuned around individual characteristic ripple frequencies. Clearly, the magnitude impulse responses related to one another by a dilation operation. In fact, the cortical and cochlear processing is quite similar that one can view the spectral profile (input of the cortical processing along the tonotopic axis) as an acoustic signal (input of the cochlear processing along the time axis), and the cortical constant Q (ripple) filters as the cochlear filters. The analogy between cortical and cochlear processing can be summarized as following:

$$\textit{cochlear processing}: \quad \textit{time axis} \quad \Leftrightarrow\!\rightarrow \textit{tonotopic } (\log f) \textit{ axis}$$

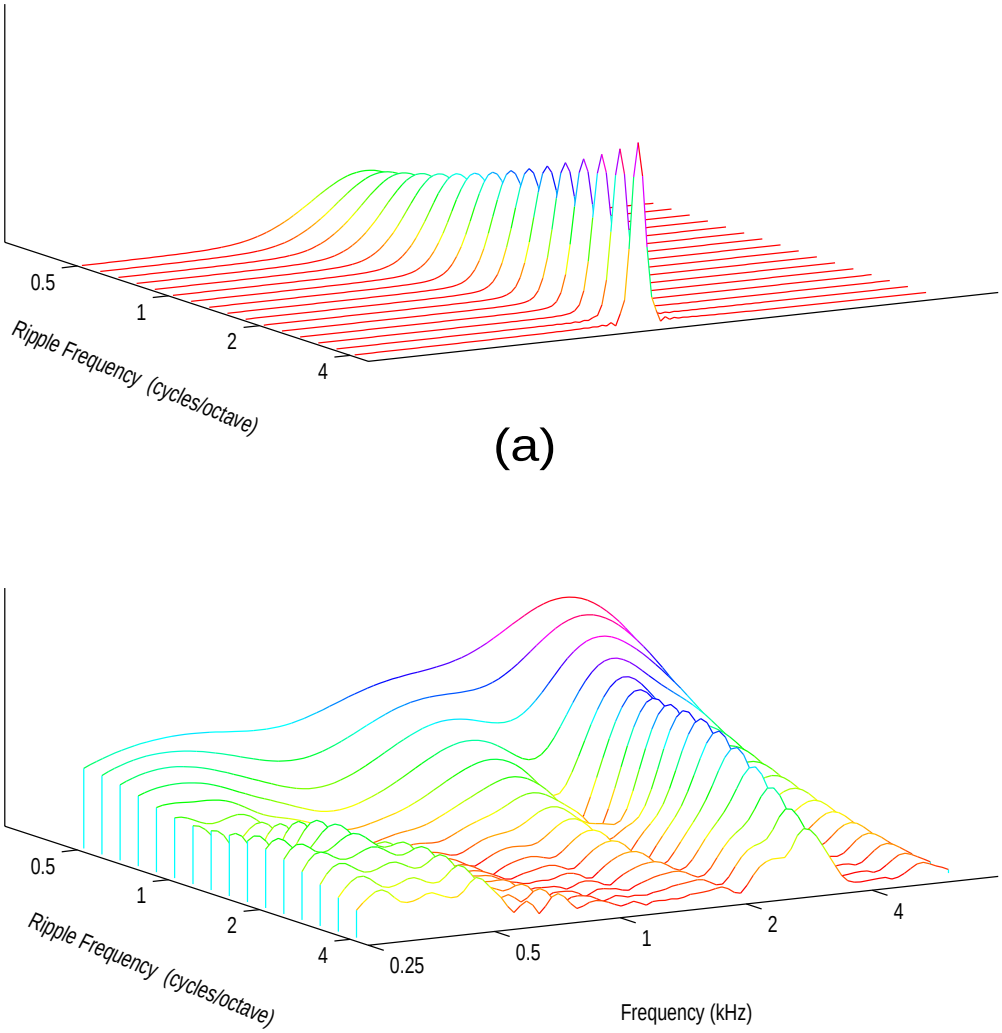

Figure 3.1: The magnitude impulse responses and analysis output of cortical filter banks. (a) The magnitude impulse responses of cortical filters with respect to different ripple frequencies. (b) The multiresolution output of the cortical filter banks for the utterance 'Come home right away.'.

$$cortical \;\; processing: \;\; tonotopic \;\; axis \;\; \Leftrightarrow\!\!\!\rightarrow \; scale \; (\log \Omega) \; axis$$

where f and Ω denote the acoustic frequency and spatial frequency, respectively.

The magnitude of the cortical representation of the long-term average auditory spectrum of the utterance 'Come home right away.' (Figure 2.5) are depicted in Figure 3.1 part (b). Note, the higher resolution the output representation of the cortical filters tuned to higher ripple frequencies. In addition, Eq. 3.3 indicates that not only the magnitude response $a_s(x)$ but the phase response $\psi_s(x)$ of the cortical processing is also necessary for preserving all the information of the auditory spectrum. Figure 3.2 illustrates the computed values of these two functions for utterance 'Come home right away.'. The phase $\psi_s(x)$ is represented by colors in the following manner: $\Leftrightarrow\!\!\!\!\!3\pi/4 \sim \Leftrightarrow\!\!\!\!\!\pi/4$ is red; $\Leftrightarrow\!\!\!\!\!\pi/4 \sim \pi/4$ is yellow; $\pi/4 \sim 3\pi/4$ is blue; other is purple. Therefore, at a given scale, the yellow color roughly indicates the peaks of the spectral profile resolved at that scale. In addition, the magnitude $a_s(x)$ is denoted by the intensity of the color. In this cortical representation, only the coarse outlines of the auditory spectrum (solid line) is resolved at the lowest scales, whereas the finer structure is represented at the higher scales (as indicated by the increasing number of the yellow bands towards the higher ripple frequencies of the scale axis).

In the following computer simulations in this chapter, the complex cortical filters are generated by dilating and sinusoidally interpolating a mother function $\omega_s(x)$ as described in this section. The scale (spatial frequency) axis covers the range from 0.5 cycle/octave to 4.6 cycle/octave with 5 channels per octave resolution (shown in Figure 3.1 (a)),i.e., with a dilation factor equivalent to 0.2/octave or 14.87%. This resolution is higher than estimated (20%) from the psychoacoustic experiments [22]. Meanwhile, the tonotopic axis is discretized

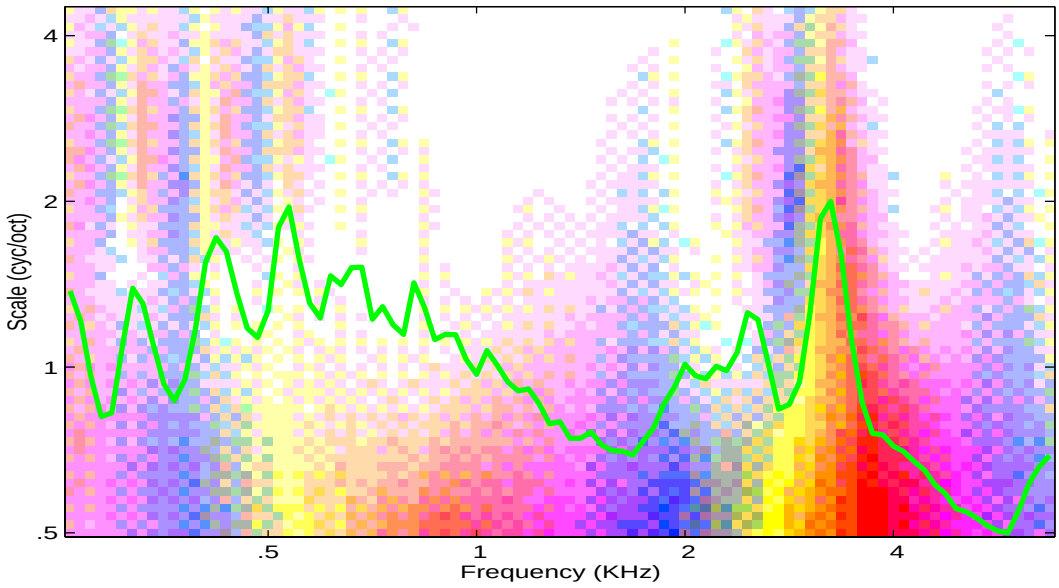

Figure 3.2: The cortical representation of auditory spectrum (solid line) for utterance 'Come home right away.'. The tonotopic axis is given in KHz. The scale axis increases in resolution from bottom to top.

with the same resolution 20 channels per octave as in the peripheral model in Chapter 2, which covers the frequency range from 250 Hz to 6.7 KHz. Finally, the two-dimensional cortical representation that effectively encodes the local bandwidth and asymmetry of the auditory spectrum around each frequency is employed for the speaker verification and identification application.

To demonstrate the effectiveness of the cortical representation to the speaker identification problem, Figure 3.3 illustrates the cortical representations of the long-term average auditory spectra (solid line) of 12 female and male speakers from training division in dialect region 1 (dr1) of the TIMIT database. These patterns are stable and distinctive and generally reflect vocal tract shape resonances of each speaker. This insight manifests the feasibility of the utility of these representations to speaker identification problem.

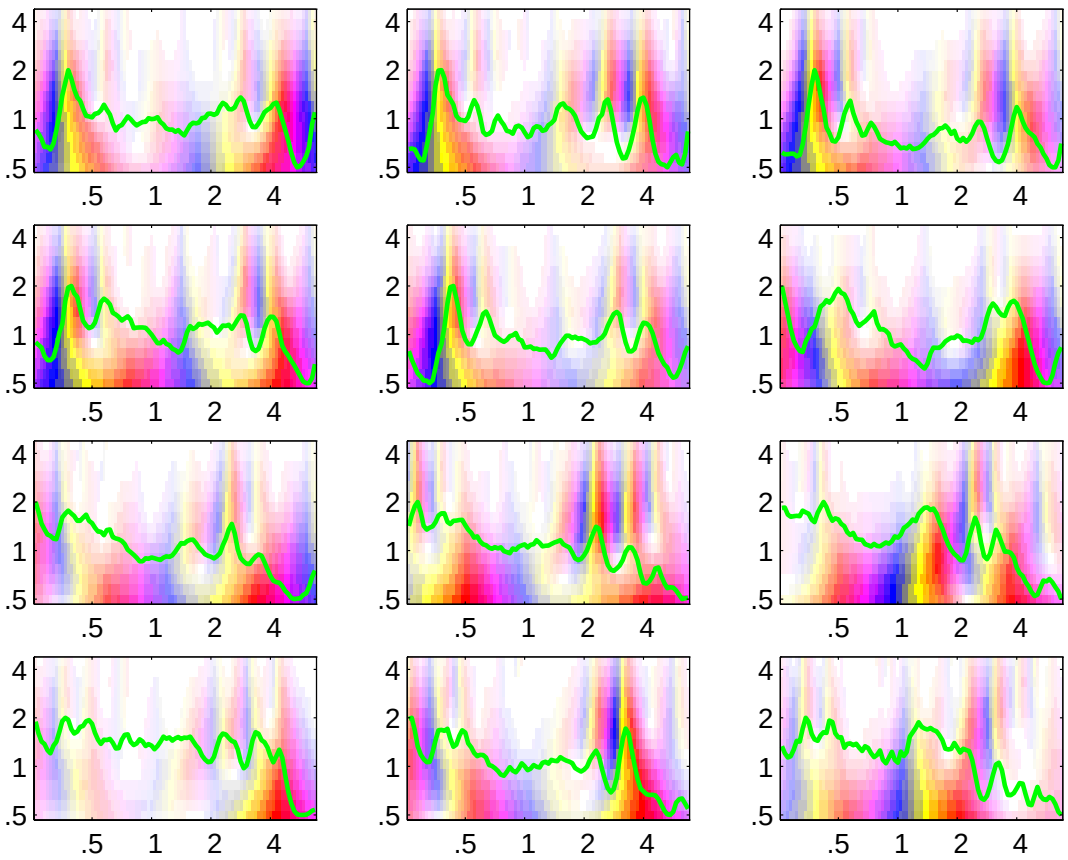

Figure 3.3: Examples of the cortical representation of 6 female (top 2 rows) and 6 male speakers. Each representation is derived from the average auditory spectrum (solid line) of 8 sentences (5 SX sentences and 3 SI sentences). Like in Figure 3.2, the tonotopic axis is given in KHz and the scale axis is given in ripple frequency (cyc/oct).

3.3 Pattern Matching

The cortical representation of the long-term average auditory spectrum is believed to contain speaker-dependent information. Taking advantage of the fidelity of the cortical map drives the speaker identification problem into the image registration problem. Therefore, a *correlator* which is often used as a template matching approach for traditional scene matching problem [6] is employed here to distinguish speakers by measuring the similarity among images (cortical representations).

The correlation[1] of two continuous functions $f(x)$ and $g(x)$, denoted as $f(x) \circ g(x)$, is defined

$$f(x) \circ g(x) \;=\; \int_{-\infty}^{\infty} f^*(\alpha) g(x+\alpha) d\alpha \tag{3.7}$$

where $*$ means complex conjugate. This definition can be easily extended to two-dimensional correlation of continuous functions $f(x,y)$ and $g(x,y)$

$$f(x,y) \circ g(x,y) \;=\; \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f^*(\alpha,\beta) g(x+\alpha, y+\beta) d\alpha d\beta \tag{3.8}$$

To the equivalent discrete case, Eq. 3.7 and Eq. 3.8 can be respectively modified to

$$f(x) \circ g(x) \;=\; \sum_{m=0}^{M-1} f^*(m) g(x+m) \tag{3.9}$$

for $x = 0, 1, 2, \ldots, M \Leftrightarrow 1$ and

$$f(x,y) \circ g(x,y) \;=\; \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} f^*(m,n) g(x+m, y+n) \tag{3.10}$$

for $x = 0, 1, 2, \ldots, M \Leftrightarrow 1$ and $y = 0, 1, 2, \ldots, N \Leftrightarrow 1$.

[1] If $f(x)$ and $g(x)$ are the same function, Eq. 3.7 is called the *autocorrelation* function; otherwise, it is called *cross correlation* function.

Like *convolution* operation, the correlation has similar relationship between the spatial and frequency domains called correlation theorem where

$$f(x,y) \circ g(x,y) \Longleftrightarrow F^*(u,v)G(u,v)$$

and

$$f^*(x,y)g(x,y) \Longleftrightarrow F(u,v) \circ G(u,v)$$

The above notation $f \Leftrightarrow F$ means that f and F constitute a Fourier transform pair. However, it is much more intuitive to deal with signal in spatial domain than in frequency domain.

Considering template $T(x,y)$ and test image $I(x,y)$ of the same size $P \times Q$, the correlation between these two functions can be simply stated as

$$c(s,t) \;=\; \sum_x \sum_y T^*(x,y)I(x \Leftrightarrow s, y \Leftrightarrow t) \tag{3.11}$$

where $s = \Leftrightarrow(P \Leftrightarrow 1), \ldots, \Leftrightarrow 1, 0, 1, \ldots, P \Leftrightarrow 1$ and $t = \Leftrightarrow(Q \Leftrightarrow 1), \ldots, \Leftrightarrow 1, 0, 1, \ldots, Q \Leftrightarrow 1$, and the summation is taken over the overlapped image region. Obviously, the accuracy is seriously degraded for values of s and t near the boundary. In addition, the correlation function $c(s,t)$ has the disadvantage of being sensitive to changes in the amplitude of $T(x,y)$ or $I(x,y)$. To overcome this drawback, it is necessary to normalize the correlation function into *correlation coefficient* $R(s,t)$, which is

$$R(s,t) \;=\; \frac{\sum_x \sum_y [T^*(x,y) \Leftrightarrow \mu_T^*][I(x \Leftrightarrow s, y \Leftrightarrow t) \Leftrightarrow \mu_I]}{\sqrt{\sum_x \sum_y [T^*(x,y) \Leftrightarrow \mu_T^*]^2 \sum_x \sum_y [I(x \Leftrightarrow s, y \Leftrightarrow t) \Leftrightarrow \mu_I]^2}} \tag{3.12}$$

where $s = \Leftrightarrow(P \Leftrightarrow 1), \ldots, \Leftrightarrow 1, 0, 1, \ldots, P \Leftrightarrow 1$ and $t = \Leftrightarrow(Q \Leftrightarrow 1), \ldots, \Leftrightarrow 1, 0, 1, \ldots, Q \Leftrightarrow 1$, μ_T and μ_I are the average values of the pixels in overlapped areas of $T(x,y)$ and $I(x,y)$ and the summations are taken over the coordinates common to both

T and I. Furthermore, the Cauchy-Schwartz inequality indicates that $\Leftrightarrow 1 \leq R(s,t) \leq 1$. However, the major disadvantage of this correlator approach is the computation complexity. From Eq. 3.12, one can see a great amount of computation must be performed since the search range for s,t are usually large in an actual image. In other words, no decision can be made until the correlation array $R(s,t)$ is computed for all possible s,t with this technique.

The traditional template matching application in image processing is to find the closest match between an unknown image and a set of known images. The correlator approach is to compute the correlation coefficients between the unknown image I and each of the known images T_j. Finally, the known image with the largest correlation coefficient will be selected as the closest match for the unknown image. In the next section, this above concept is applied to the cortical representation for speaker identification application.

3.4 Experimental Evaluation

The experimental results of applying the multiscale cortical representation, which conveys all information about the auditory spectrum and can be thought as a two-dimensional image, to speaker identification problem is presented in this section. The cortical representations are tested as 16×95 (16 channels and 95 channels respectively distributed on scale axis and frequency axis) images in all following experiments. First, the fidelity of the simple correlator method is examined for a speaker verification task. In this investigation, three features, LPC spectrum, auditory spectrum and cortical representation, are employed to test the performance with the correlator approach. After its performance inspected,

the correlator is used to identify the speakers by matching the cortical representations of test utterances and long-term average template for each speaker. Finally, the robustness of the cortical representation is also investigated under different SNR conditions for identification task. In addition, the noise robustness of the cortical representation is compared with that of LPC spectrum at the end of this chapter. The main reason of selecting LPC spectrum for comparison is that the all-pole model of LPC is well known to provide a good approximation to the vocal tract spectral envelope.

3.4.1 Speaker Verification

The speaker verification experiment is the detection of a given target speaker. Given a test utterance, a target speaker identity will be assigned as a test hypothesis, and the task is to determine whether this hypothesis is true or false (binary test). In brief, the purpose of this experiment is just to verify the fidelity of the correlator method when used for the cortical representation. With this in mind, one reasonable assumption that $R(0,0)$ is the maximum correlation coefficient in the correlation array $R(s,t)$ (Eq. 3.12) is made to reduce the computation in this work. In other words, $R(0,0)$ is *assumed* to be the correlation coefficient between the test utterance I and speaker template T with the following formula:

$$R(0,0) = \frac{\sum_x \sum_y [T^*(x,y) \Leftrightarrow \mu_T^*][I(x,y) \Leftrightarrow \mu_I]}{\sqrt{\sum_x \sum_y [T^*(x,y) \Leftrightarrow \mu_T^*]^2 \sum_x \sum_y [I(x,y) \Leftrightarrow \mu_I]^2}} \tag{3.13}$$

Furthermore, to remove the text-dependent information, only 8 sentences (5 phonetically-compact sentences plus 3 phonetically-diverse sentences) per speaker are used as test utterances and the first female speaker of dialect region 1 in TIMIT database is chosen as the target speaker.

The traditional LPC spectrum is formulated by the following equation:

$$H(e^{j\omega}) \;=\; \frac{G}{1 \Leftrightarrow \sum_{k=1}^{p} \alpha_k e^{-j\omega k}} \tag{3.14}$$

where G is called gain parameter and $\{\alpha_k\}$ are linear prediction coefficients which satisfy the minimum mean-squared prediction error criterion for a p^{th} order linear prediction model. This linear prediction model provides a reliable and accurate method for estimating the parameters that characterize the linear speech synthesis model and the LPC spectrum has been successfully applied to wide range of speech problems [13]. In this experiment, the LPC spectrum is calculated for a 16^{th} order LPC processor with a first-order preemphasis system $H(z) = 1 \Leftrightarrow 0.9z^{-1}$. The speech waveform is also segmented into 16 ms (256 sample points at 16 KHz sampling rate) frames with 8 ms shift between frames. In addition, the Hamming window (a "typical" window used in LPC-based speech recognition system) which has the form

$$w(n) \;=\; 0.54 \Leftrightarrow 0.46 \cos(\frac{2\pi n}{N \Leftrightarrow 1}), \qquad 0 \leq n \leq N \Leftrightarrow 1 \tag{3.15}$$

is used to window each frame to minimize the signal discontinuities at the beginning and end of each frame.

Figure 3.4 shows the log magnitude of the long-term average LPC spectrum with the log magnitude of the long-term average FFT spectrum superimposed for the utterance 'Come home right away.' by a male speaker. It can be observed that the LPC spectrum clearly retains the formant but no pitch information. Figure 3.5 exhibits the long-term average auditory spectrum (solid line) and the LPC spectrum (dashed line). For demonstration purpose, the LPC spectrum is plotted versus mel-frequency not linear frequency. This figure demonstrates that both overall spectral profiles have the same trend except the auditory spec-

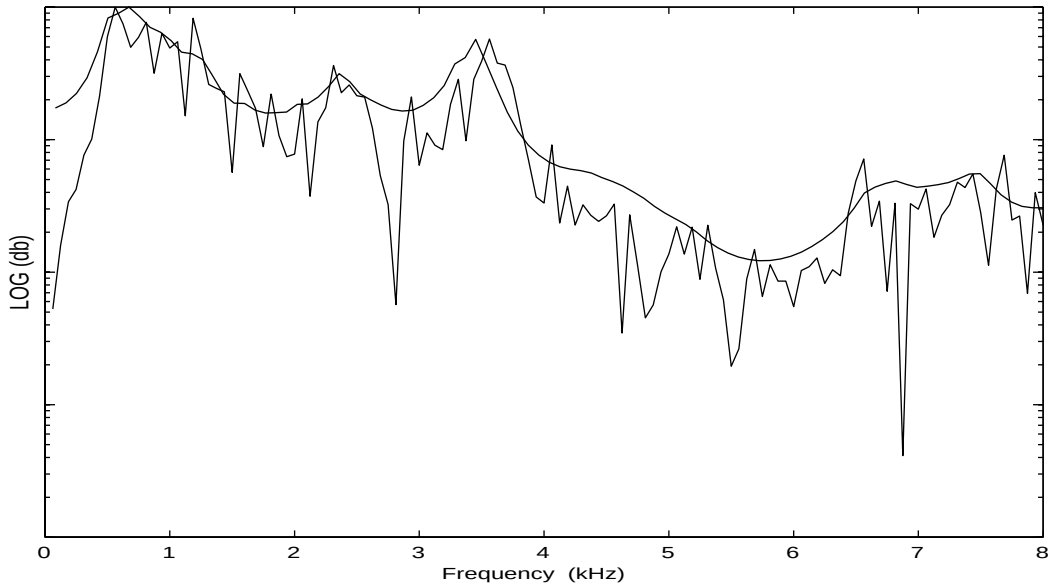

Figure 3.4: The log magnitude of the long-term average LPC spectrum with FFT spectrum superimposed for the utterance 'Come home right away.'.

trum provides much more detailed information such as harmonic peaks in lower frequency and more suppression in amplitude gain in lower frequency than LPC spectrum due to the characteristics of the peripheral auditory model. Intuitively, this suppression operates as a highpass filter of the spectral pattern, enhancing the relative expression of nearby peaks while reducing the overall slow variations or tilts in the spectrum [24].

Eq. 3.12 computes the correlation coefficient for two-dimensional images. It can be simplified to one-dimensional LPC spectrum or auditory spectrum which comes from a $L^2(\Re)$ signal space. The correlation coefficient between a test spectrum f and template g is evaluated by

$$r(s) = \frac{\sum_x [g(x) \Leftrightarrow \mu_g][f(x \Leftrightarrow s) \Leftrightarrow \mu_f]}{\sqrt{\sum_x [g(x) \Leftrightarrow \mu_g]^2 \sum_x [f(x \Leftrightarrow s) \Leftrightarrow \mu_f]^2}} \tag{3.16}$$

In addition, the assumption that $r(0)$ will yield the maximum value simplifies

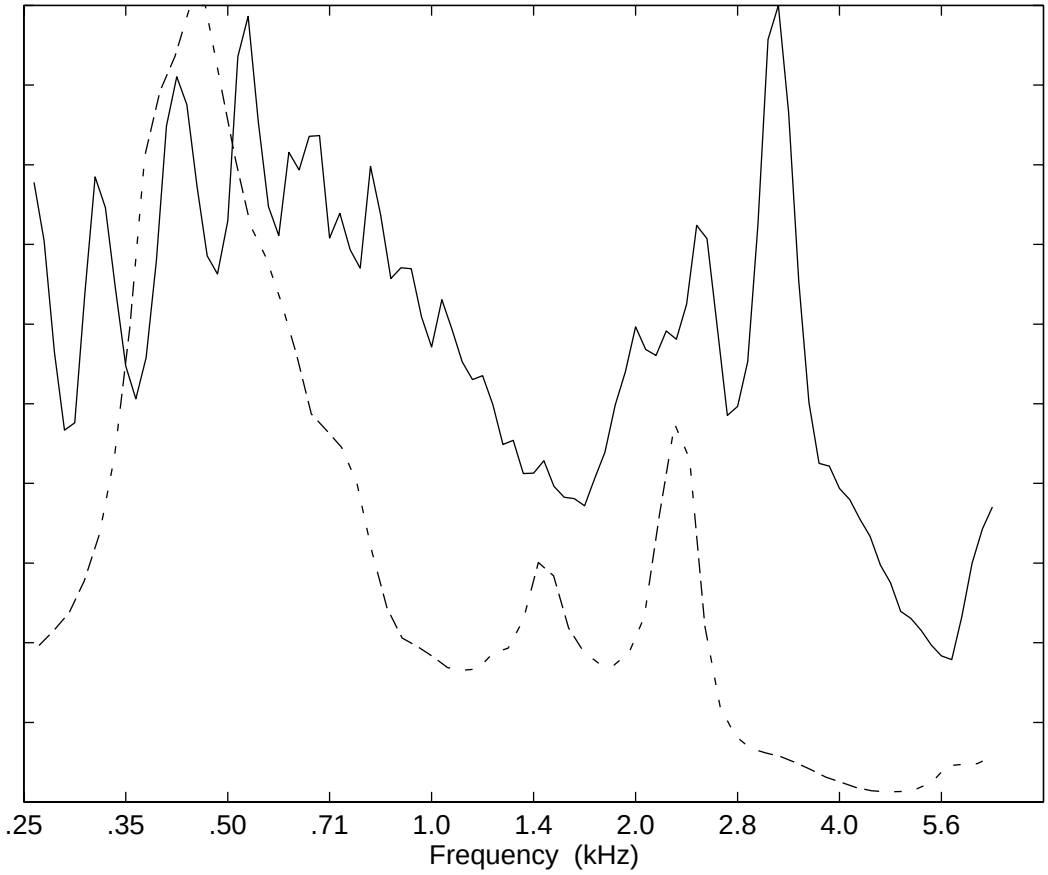

Figure 3.5: The magnitude of the long-term average LPC spectrum (dashed line) with the long-term average auditory spectrum (solid line) for the utterance 'Come home right away.'. The LPC spectrum is plotted in mel-frequency for demonstration.

Eq. 3.16 into

$$r(0) = \frac{< f, g >}{\|f\| \cdot \|g\|} \tag{3.17}$$

where the inner product and the induced norm on $L^2(\Re)$ is defined for discrete signal as

$$< f, g > = \sum_x f(x)g(x)dx$$

and

$$\|f\| = \sqrt{< f, f >}$$

The speaker verification task is evaluated for the first female speaker in dialect region 1. The long-term average templates for cortical representation, auditory spectrum and LPC spectrum are trained from 8 sentences spoken by this target speaker. By Eq. 3.13 and Eq. 3.17, one can calculate the correlation coefficients between the template and any individual test utterance for these three representations. Figure 3.6 shows the distribution of the computed correlation coefficients corresponding to these three representations.The dotted bars represent the distribution for her own 8 training sentences while the solid bars represent the evaluation for 104 testing utterances spoken by other 13 female speakers in the same dialect region 1. The performance of the cortical representation, auditory spectrum and LPC spectrum is depicted from top to bottom respectively. The separation in cortical representation feature between those utterances for target speaker and non-target speakers is more obvious than in other two spectral features.

However, shown in Figure 3.6, there is one training sentence not able to match the template as well as other seven training sentences for all three features. This particular utterance which lasts about 1.18 sec (18927 sample points at 16 KHz sampling frequency) is the shortest one spoken by the target speaker

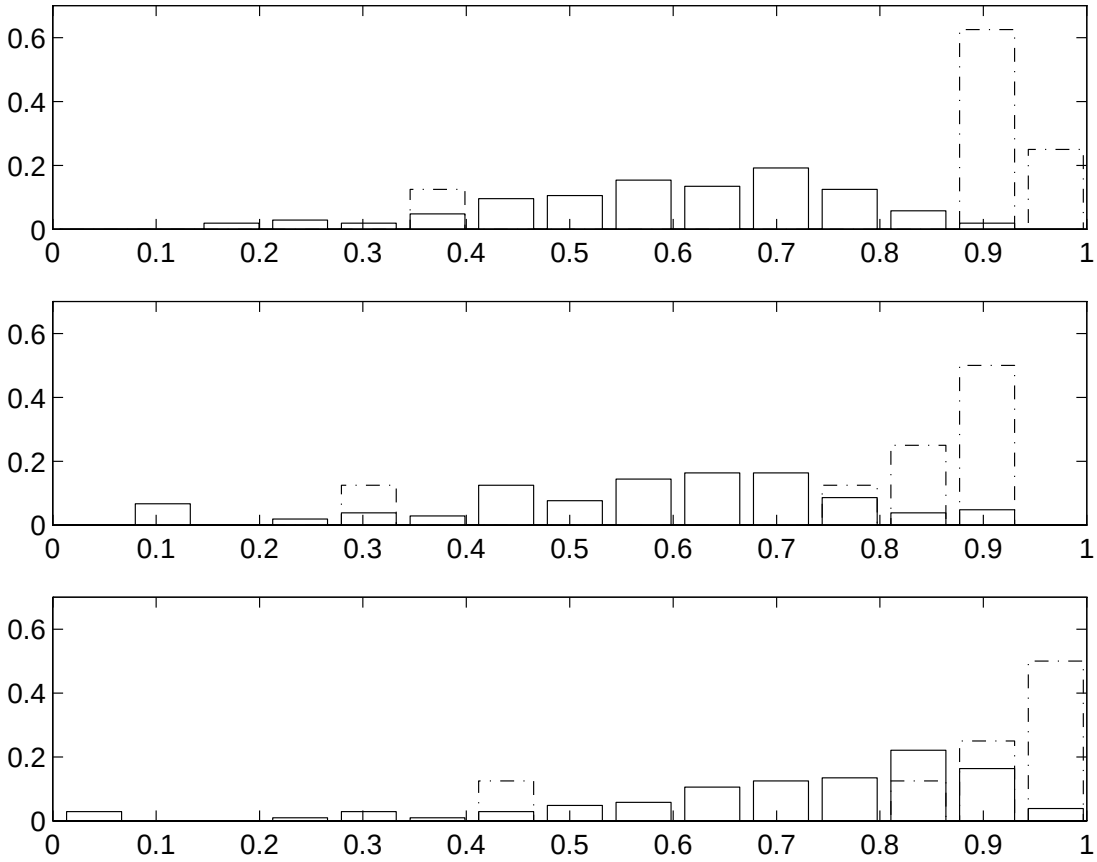

Figure 3.6: The distribution of the correlation coefficients between the templates and test utterances spoken by female speakers in dr1. The evaluation for cortical representation, auditory spectrum and LPC spectrum is respectively plotted from top to bottom.

and is believed not able to provide stable features to represent the target speaker. Therefore, it is not adequate to be included in the training set. After this particular utterance discarded from the training and testing sets, the performance of the correlator is re-investigated based on a modified template that results from the remaining seven training sentences of the target speaker. Figure 3.7 demonstrate the new experimental evaluations corresponding to old results shown in Figure 3.6.

This speaker verification task can be summarized as a binary test problem which is

$$d = \begin{cases} 1 & \text{if } matching \;\; coefficient \geq \eta \\ 0 & \text{otherwise} \end{cases}$$

where d is the decision function and η is called decision parameter. According to the data shown in Figure 3.7, one can get the relation between *probability of miss* and *probability of false alarm* by gradually changing decision parameter η from 0 to 1 and the outcomes are demonstrated in Figure 3.8 for these three features. The miss and false alarm probabilities are defined as

$$P_m = pr(d = 0|H = 1)$$

$$P_{fa} = pr(d = 1|H = 0)$$

where hypothesis H means the test utterance belonging to the target speaker and pr means the probability measure.

The curve describing the relation between P_m and P_{fa} is useful in determining the minimum of *detection cost function (DCF)* which is a function of P_m and P_{fa} and often used to measure the performance of the speaker verification system. The DCF can be described as

$$DCF \;\; = \;\; C_m \times P_m \times P_t + C_{fa} \times P_{fa} \times (1 \Leftrightarrow P_t)$$

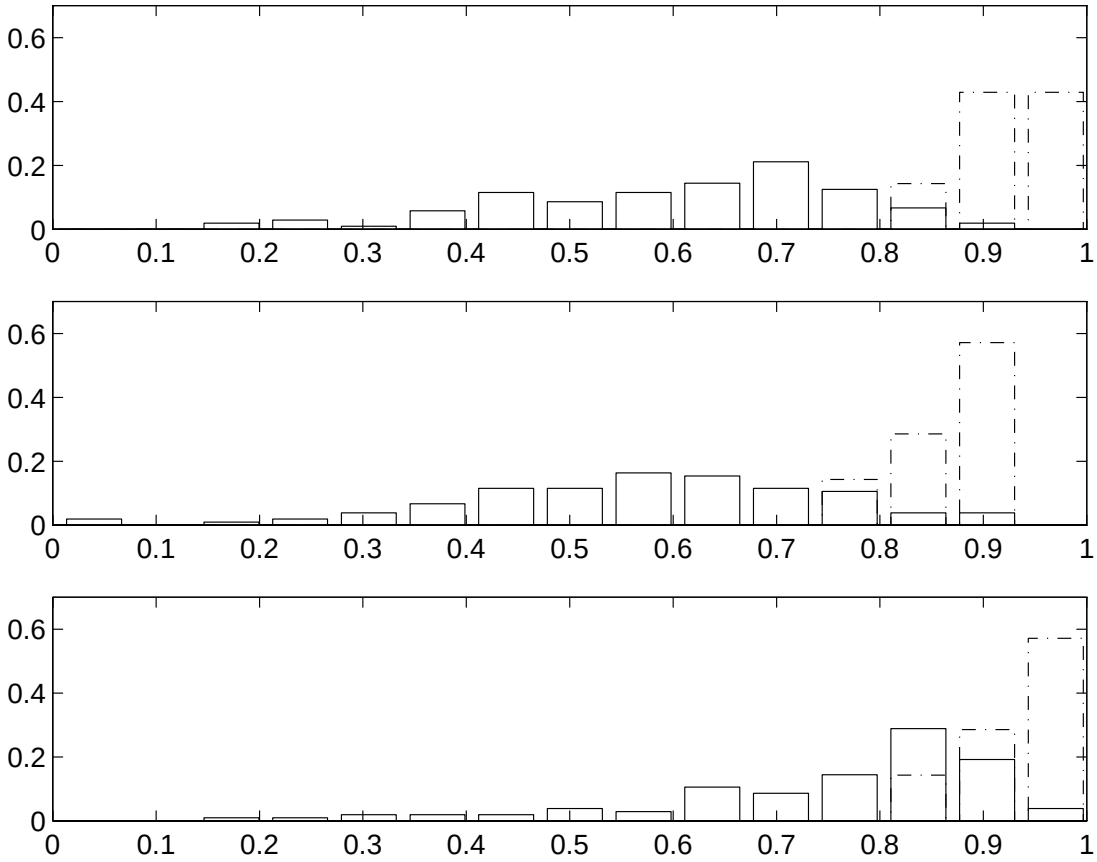

Figure 3.7: The distribution of the correlation coefficients between the *modified* templates and test utterances spoken by female speakers in dr1. The evaluation for cortical representation, auditory spectrum and LPC spectrum is respectively plotted from top to bottom.

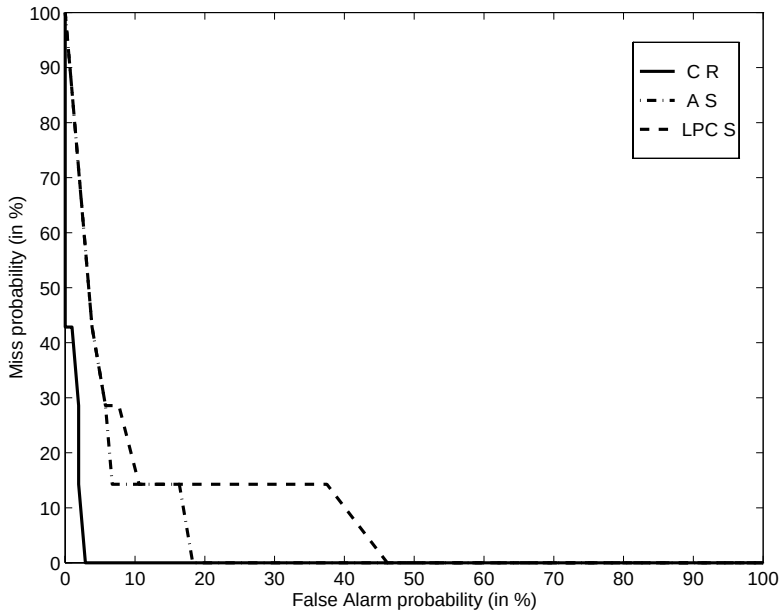

Figure 3.8: Probability of miss versus probability of false alarm for data shown in Figure 3.7.

where C_m and C_{fa} indicate the cost of miss error and false alarm error, P_t means the *priori* probability of the target speaker. In Figure 3.8, if the decision parameters η for these three representations are selected to make $P_m = 0$, the cortical representation yields much better performance (much lower minimum false alarm probability) than the other two spectral representations. In fact, given any constant P_m, the cortical representation carries a much lower P_{fa} than others. The minimum false alarm probability corresponding to zero miss probability for these three representations is summarized in Table 3.1 (Remember there are 104 testing utterances belonging to other female speakers).

It is very easy to distinguish male from female speakers by the pitch information in one sentence. Since the long-term average cortical representation and auditory spectrum preserves the pitch information of the speaker, they should

	False Alarm Probability (P_{fa})
Cortical Representation	3/104
Auditory Spectrum	12/104
LPC Spectrum	37/104

Table 3.1: Minimum false alarm probability corresponding to zero miss probability for the data shown in Figure 3.7.

ideally yield the perfect performance in separating male and female speakers. The experimental result for testing 192 utterances spoken by 24 male speakers from the same dialect region 1 is shown in Figure 3.9. As predicted, the cortical representation and auditory spectrum completely separate all sentences that belong to the female target speaker (dashed bar) from those that belong to male test speakers (solid bars). The LPC spectrum can not perform this task because of the lack of pitch information.

In the next experiment, the correlator is tested for the sentences spoken by female speakers but coming from *different* dialect region. Figure 3.10 demonstrates the evaluation for testing 184 sentences belonging to 23 female speakers from dialect region 2 of TIMIT database. As in some previous figures, the solid bars depict the distribution of the test sentences spoken by test speakers while the dashed bars indicate the distribution of the training sentences spoken by the target speaker. In addition, the minimum false alarm probability with zero miss probability for these three representations is summarized in Table 3.2. It is evident from these data that the cortical representation contains stable and unique cues for discriminating speakers coming from different dialect regions.

In summary, the cortical representation provides a much better performance than auditory spectrum and LPC spectrum for the text-independent speaker

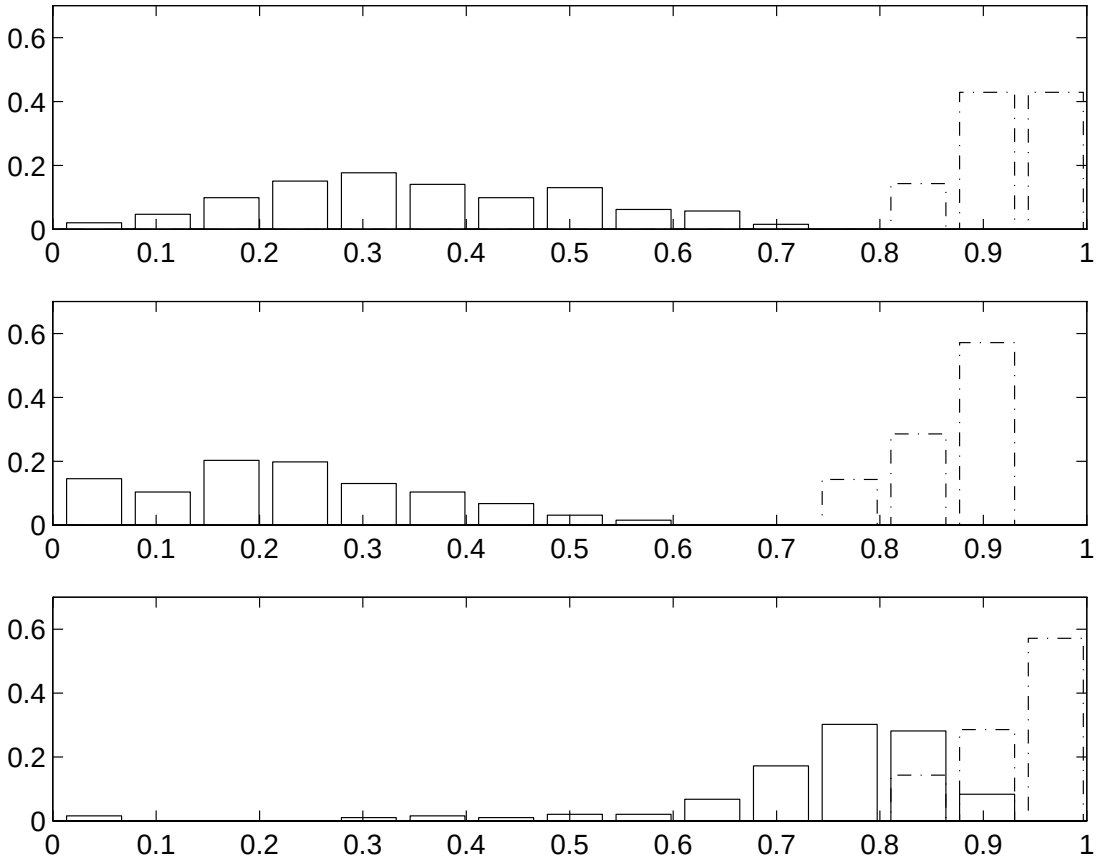

Figure 3.9: The distribution of the correlation coefficients between the *modified* templates and test utterances spoken by male speakers in dr1. The evaluation for cortical representation, auditory spectrum and LPC spectrum is respectively plotted from top to bottom.

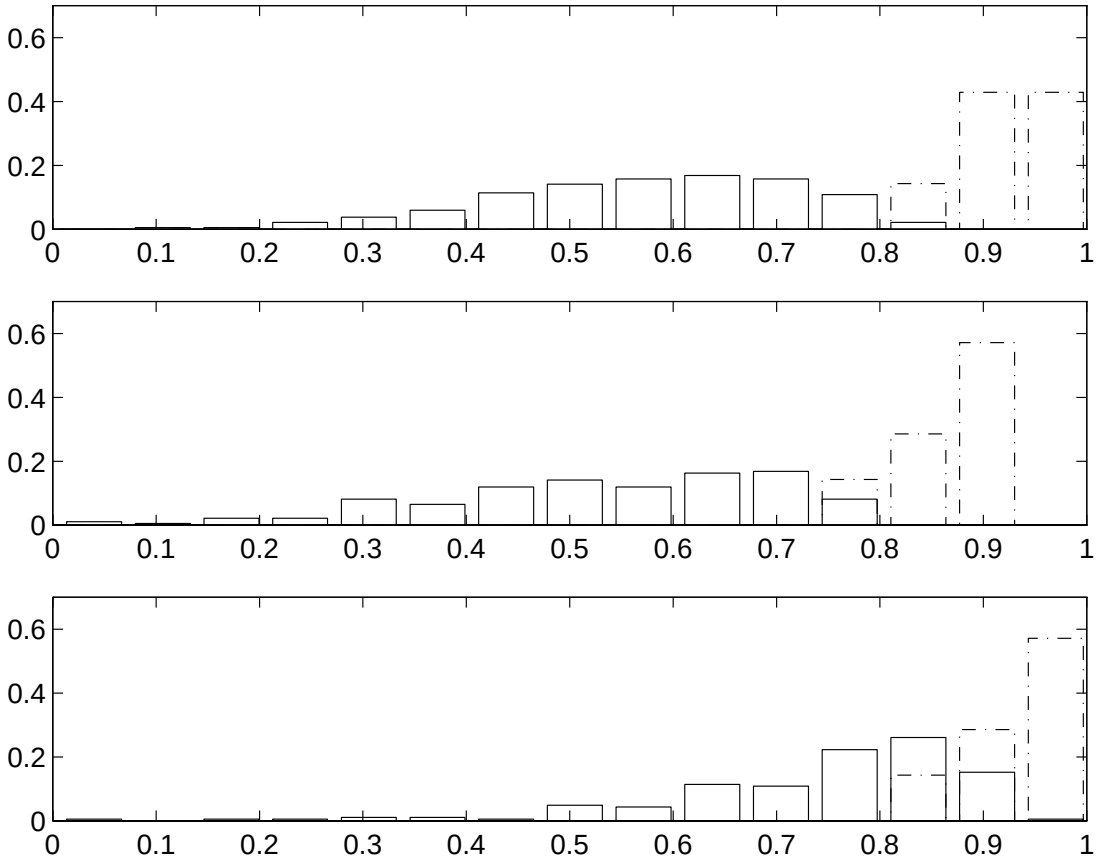

Figure 3.10: The distribution of the correlation coefficients between the *modified* templates and test utterances spoken by female speakers in dr2. The evaluation for cortical representation, auditory spectrum and LPC spectrum is respectively plotted from top to bottom.

	False Alarm Probability (P_{fa})
Cortical Representation	0/184
Auditory Spectrum	11/184
LPC Spectrum	58/184

Table 3.2: Minimum false alarm probability corresponding to zero miss probability for the data shown in Figure 3.10.

	False Alarm Probability (P_{fa})
Cortical Representation	3/480
Auditory Spectrum	23/480
LPC Spectrum	135/480

Table 3.3: Overall minimum false alarm probability corresponding to zero miss probability for speaker verification experiments (480 test utterances).

verification (binary test) task with the use of correlation coefficients to measure the degree of matching between speakers. Combining the results from the above three experiments (tests for 13 female speakers in dr1, 24 male speakers in dr1 and 23 female speakers in dr2), the overall minimum false alarm probability with zero miss probability is stated in Table 3.3. This high accuracy of cortical representation suggests applying the correlator technique to the speaker identification (M-ary test) problem.

3.4.2 Speaker Identification

The correlator method, whose fidelity has been examined in previous section, is applied to the text-independent speaker identification problem with cortical representations. As in Section 2.4, this experiment is evaluated for 23 female

Range of t	Identification Rate
0	53/69
-1 $\sim$ 0	54/69
-1 $\sim$ 1	55/69
-1 $\sim$ 2	55/69
-1 $\sim$ 3	55/69
0 $\sim$ 1	54/69
-1 $\sim$ 1	55/69
-2 $\sim$ 1	55/69
-3 $\sim$ 1	55/69

Table 3.4: Identification rate for various ranges of frequency t with zero spatial frequency s.

speakers from dialect region 2 of the TIMIT database. For each speaker, the same seven utterances that have been used to establish the Gaussian model in Chapter 2 are employed to build the template and the other three utterances constitute the testing space. The basic idea of this correlator approach to identify the speaker can be summarized as

$$Speaker \quad Index = \arg \max_{1 \leq j \leq M} \{R_j\} \tag{3.18}$$

where R_j is the correlation coefficient between test utterance I and template T_j and M is the total number of speakers.

The assumption in speaker verification task that $R_j(0,0)$ is equal to the maximum value of $R_j(s,t)$ for all possible shifts of spatial frequency s and acoustic frequency t is abandoned in this experiment. However, it is still intuitive to *assume* that the maximum value of $R_j(s,t)$ will occur on a small enough range of

s and t that the test feature I and template T_j are almost on the same position. This leads to a series of pre-experiments to determine the minimum needed range of s and t to reduce redundant computation. All cortical representation features used in these pre-experiments are derived from the clean ($SNR = \infty$) speech signals. First, the correlator is operated on various ranges of t while the spatial frequency variable s is set to zero. Table 3.4 demonstrates the identification rate which is defined as

$$identification\ \ rate = \frac{\#\ of\ correctly\ identified\ utterances}{total\ \#\ of\ utterances}$$

with respect to different small ranges of t. The top portion of this table implies 1 is the efficient upper bound for t with fixed lower bound $\Leftrightarrow$1. Combining the similar results from lower part of this table, the efficient range of frequency variable t is finally selected as $\Leftrightarrow 1 \leq t \leq 1$ to reduce computation. Similar procedures are performed to determine the efficient range of spatial frequency s with constant (zero) frequency t. Table 3.5 indicates the maximum correlation coefficients between test utterances and speaker templates occur on the condition $s = 0$. This result matches the intuition that there is just a little correlation between different resolutions of auditory spectrum (which are represented by elements of cortical representation on different spatial frequency s). Finally, according to the implications of Table 3.4 and Table 3.5, the efficient ranges of spatial frequency s and acoustic frequency t are selected as $s = 0$ and $\Leftrightarrow 1 \leq t \leq 1$ to reduce the computation for the following speaker identification experiments in noisy environments. As depicted in Section 2.4, the noisy environments are simulated by adding white noise with different signal-to-noise ratios into the clean speech signal and the inspected signal-to-noise ratios are still 0dB $\Leftrightarrow$24dB in step of 3dB.

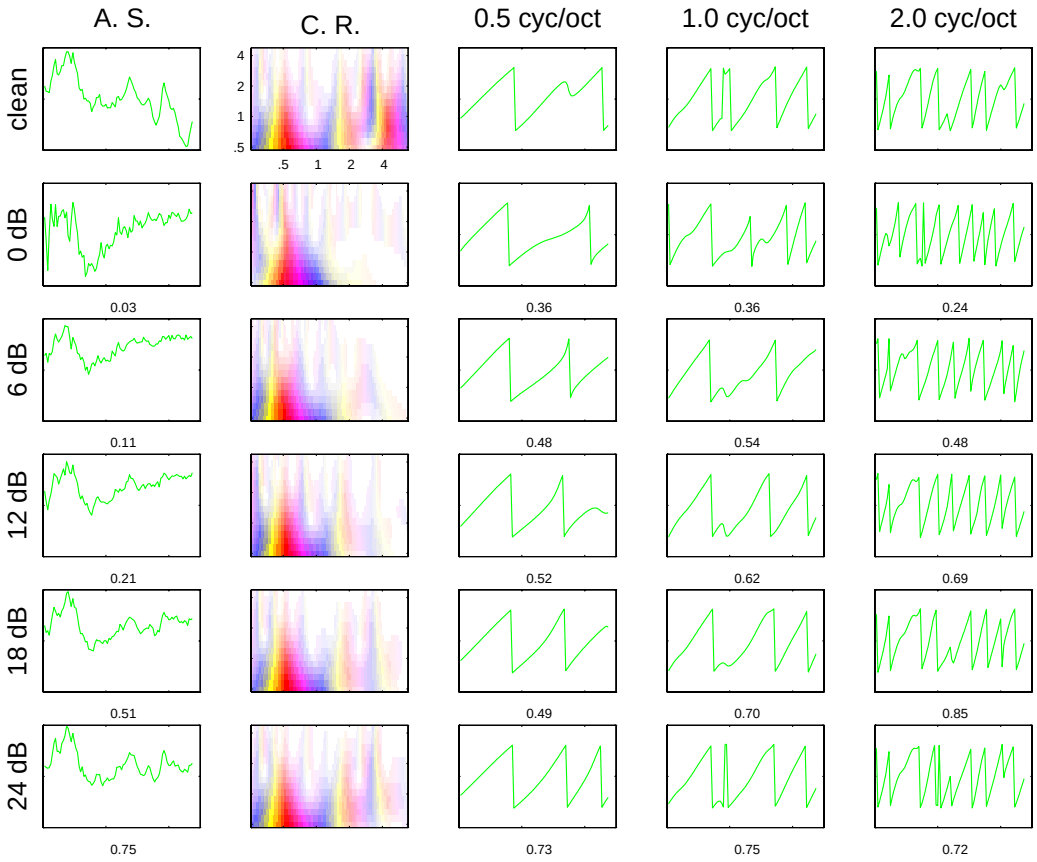

Figure 3.11: The auditory spectrum and the different scaled (ripple frequency = 0.5, 1.0, 2.0 cyc/oct) phase responses of the cortical processing under various noisy conditions for the test utterance 'Those answers will be straightforward if you think them through carefully first.'. The number below each panel indicates the correlation coefficient between itself and the same-scaled clean signal (top row).

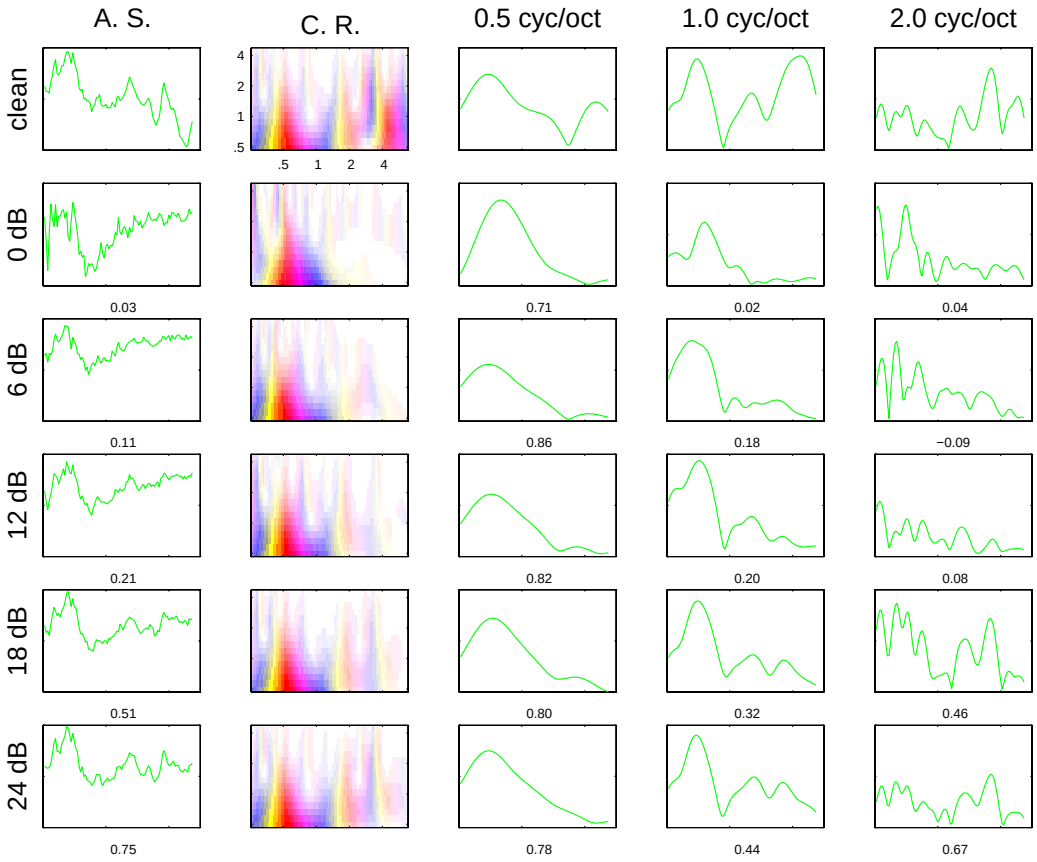

Figure 3.12: The auditory spectrum and the different scaled (ripple frequency = 0.5, 1.0, 2.0 cyc/oct) magnitude responses of the cortical processing under various noisy conditions for the test utterance 'Those answers will be straightforward if you think them through carefully first.'. The number below each panel indicates the correlation coefficient between itself and the same-scaled clean signal (top row).

Range of s	Identification Rate
0	52/69
-1 $\sim$ 0	52/69
-1 $\sim$ 1	52/69
0 $\sim$ 1	52/69
-1 $\sim$ 1	52/69
-2 $\sim$ 1	52/69

Table 3.5: Identification rate for various ranges of spatial frequency s with zero frequency t.

As mentioned in Section 3.2, the cortical representation consists of the magnitude response $a_s(x)$ and phase response $\psi_s(x)$. In many signal processing/system applications, the phase response has been shown more robust than the magnitude response. This observation seems to hold equally well for cortical representations as illustrated in Figure 3.11 and Figure 3.12. In each figure, the first two columns represent the auditory spectrum (A.S.) and cortical representation (C.R.) under clean and various SNR environments for test utterance 'Those answers will be straightforward if you think them through carefully first.'. The remaining three columns depict the responses at three different Ω (ripple frequency),i.e., three different resolutions of the auditory spectrum. The number below each panel is the correlation coefficient between the corrupted response and the clean one (Top row). The phase response is shown in Figure 3.11 and the magnitude response is shown in Figure 3.12. Observing the coefficient below each panel, it is clear that the phase response is much more stable and consistent (as indicated by increasing correlation towards the bottom panels) than the magnitude response

SNR	CR (Phase)	AS	LPCS
0dB	14/69	12/69	3/69
3dB	16/69	13/69	5/69
6dB	18/69	10/69	6/69
9dB	19/69	11/69	7/69
12dB	23/69	11/69	8/69
15dB	24/69	12/69	11/69
18dB	32/69	17/69	12/69
21dB	33/69	22/69	16/69
24dB	38/69	31/69	20/69

Table 3.6: Identification rate for cortical phase representation, auditory spectrum and LPC spectrum of 69 test utterances with respect to different noise levels.

at all resolutions.

Due to this observation, the phase features of the cortical representation are used in following experiments to identify speakers under different SNR conditions with correlator technique. The robust performance of this cortical phase representation is illustrated by identification rate in Table 3.6 and compared with those of auditory spectrum (AS) and LPC spectrum (LPCS). It clearly demonstrates the superior robustness of cortical phase responses. Meanwhile, the phase performance is examined for each single scale under different noisy conditions. The results shown in Figure 3.13 tell that the phase in lower scales (around 1.5 cyc/oct) carry important information to identify speakers under high noisy background (0dB - 6dB). The other important band is around 2.5 cyc/oct ,which is believed to contain the pitch information, and dominates for higher SNR conditions. Figure 3.14 demonstrates the identification performance

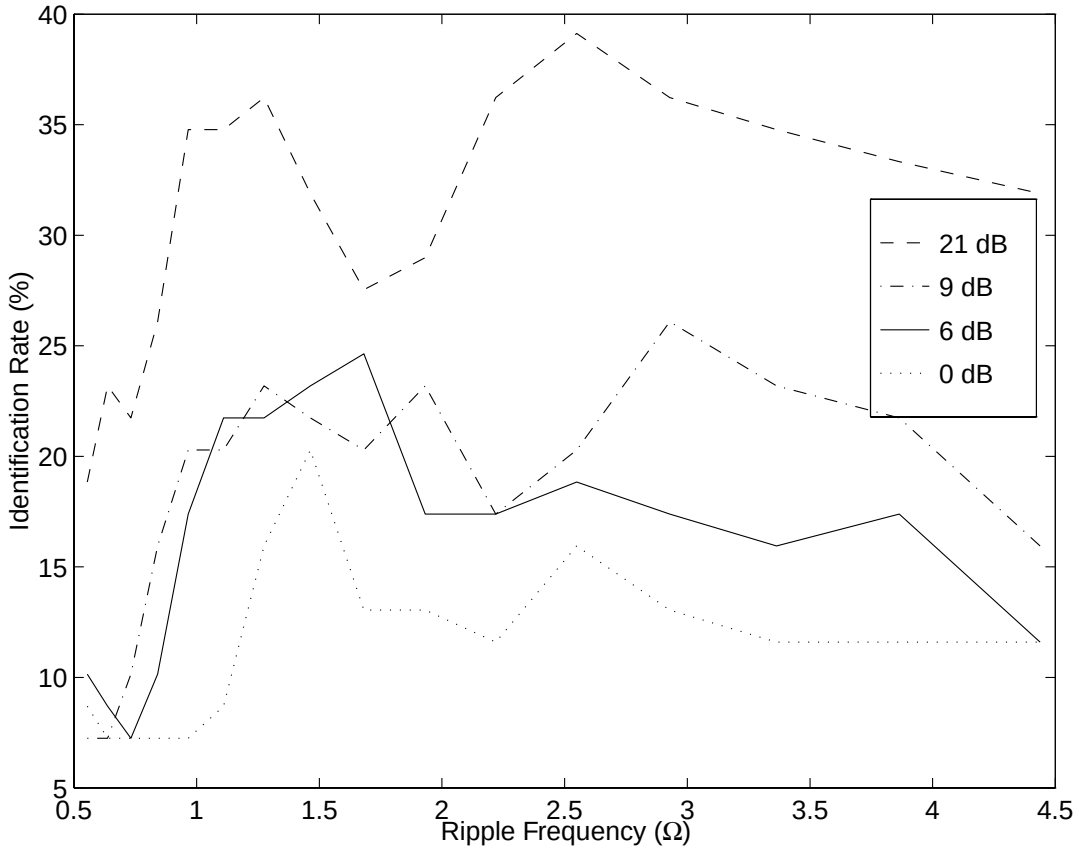

Figure 3.13: Identification rate for each single scale cortical phase representation around different SNR conditions.

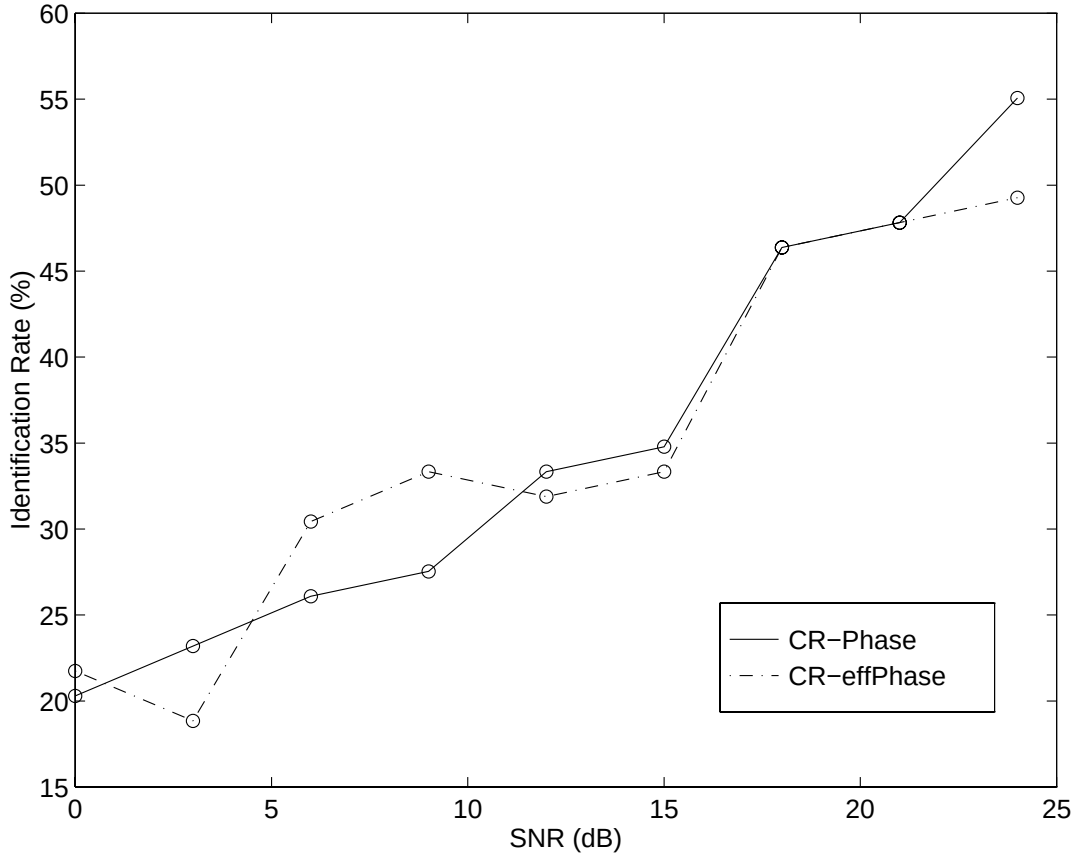

Figure 3.14: Identification rate of cortical phase representation in all scales (CR-Phase) and two subbands (CR-effPhase). These two important subbands for cortical processing are believed around $1.1 \sim 1.5$, $2.2 \sim 3.3$ cyc/oct.

for cortical phase representation in all scales and in two subbands ($1.1 \sim 1.5$, $2.2 \sim 3.3$ cyc/oct). It shows that these two subbands convey most of the robust information in identifying speakers in this case.

3.5 Summaries and Remarks

In this chapter, the double wavelet transformed cortical representation is applied to the speaker identification problem. Viewing this complex representation as an

image provides the rationale of transferring the speaker identification problem into a conventional image template matching problem. The ordinary correlator approach, which has been successfully used in scene matching applications and medical image processing for years, is employed to calculate the matching coefficients between cortical representations of test utterance and all speaker templates as measurements of similarity. Speaker verification experiments are performed for clean speech to test the fidelity of the correlator approach. However, the major drawback of the correlator approach is its computation complexity. To overcome such disadvantage, series of pre-experiments by using clean utterances for speaker identification are performed to determine the most efficient computation steps. Finally, this efficient correlator is applied to cortical representation features to distinguish speakers under noisy conditions.

Due to the experimental evaluations associated with speaker verification task in Section 3.4.1, the long-term average LPC spectrum seems not capable of effectively identifying speakers in spite that the short-term LPC spectrum is most widely used for text-dependent speech recognition system. All acoustic signals involved in this paper are extracted from TIMIT database which consists of phonetic not conversational speech. The fact that people sometimes change their pitch in a conversation but seldom in a 2-3 seconds utterance makes the long-term average pitch information a vital cue to characterize different speakers in this study. Therefore, it is not surprising that the long-term average cortical representation or auditory spectrum yields much better performance than LPC spectrum in this speaker identification application. In addition, the phase response is shown to be more robust than the magnitude response of the cortical processing and this occurrence happens to be usually observed in many pro-

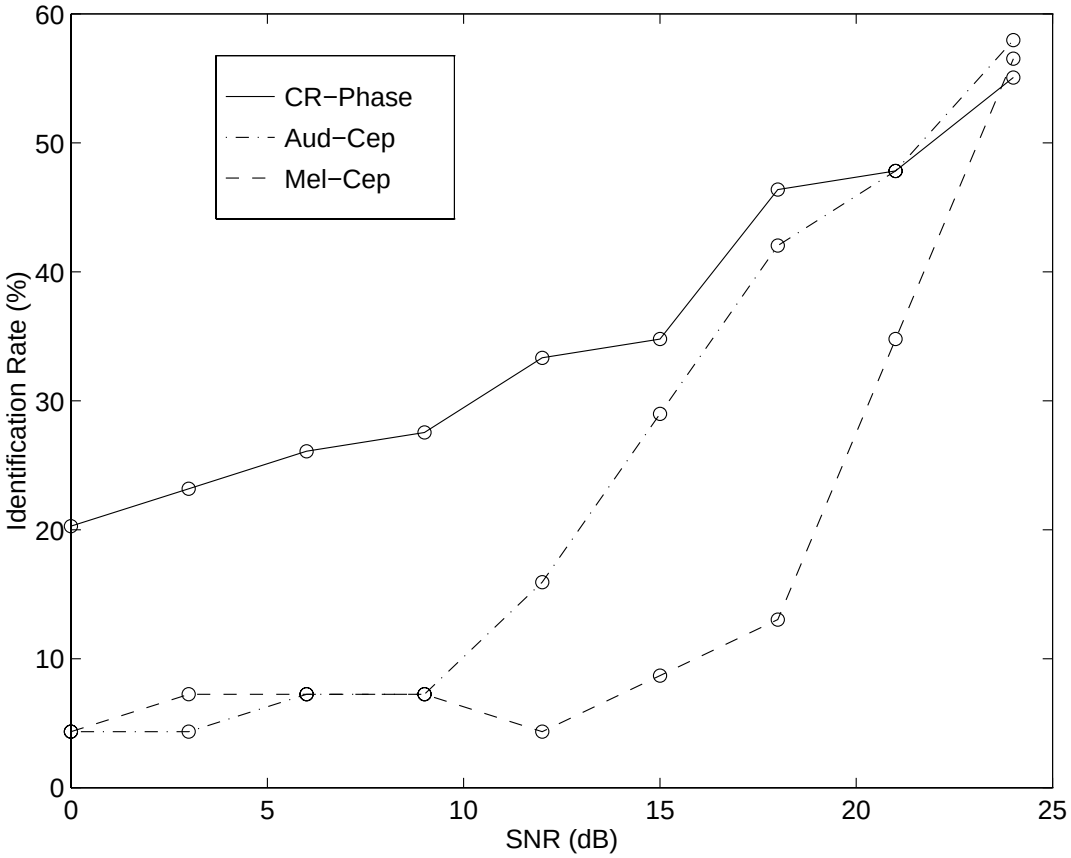

Figure 3.15: Speaker identification rate for cortical phase representation (CR-Phase) with correlator technique. This experimental result is compared with those stated in Chapter 2 for auditory cepstrum (Aud-Cep) and mel-cepstrum (Mel-Cep) with Bayes classifier technique. (Refer to Figure 2.16.)

cesses. However, the significance of pitch in characterizing speakers for general conversational speech is not investigated in this paper.

Figure 3.15 shows the phase feature of cortical representation with correlator method is much more robust than those two cepstral representations (auditory cepstrum and mel-cepstrum) with Bayes classifier in speaker identification application. As demonstrated in Chapter 2, the 14^{th} order auditory cepstrum only conveys the general trend of the auditory spectrum such that it can be seen as a particular resolution of the auditory spectral profile (Refer to Figure 2.5.). Therefore, the cortical representation which provides a multiresolution transformation of the auditory spectrum certainly contains much more 'local' characteristics about the spectral profile than cepstral representations so as to yield improvements on the identification rate and the extent of robustness.

Implied by Figure 3.13, different scales of the cortical representation should be put on different weights to analyze the speech. The correlator technique that, however, puts the same weight on each resolution while computing the matching coefficient is not appropriate in measuring "cortical distance". To find the psychophysical meaning of each scale and design a biology-based distance function for the cortical representation should be an innovative component of further researches.

Chapter 4

Conclusions and Future Studies

4.1 Conclusions

A speaker identification system often starts with a feature extraction stage to transform the acoustic signals into a compact representation which, hopefully, contains the information for effectively distinguishing among speakers. In conventional approaches, FFT or LPC based power spectrum or cepstrum which is motivated by human audition is the most accepted features. However, it seems quite natural to examine the speaker identification problem from a psychophysical point of view which is intuitively believed to provide advantages in noise-robustness and perceptual relevance for speech processing applications.

Instead of proposing a technique to improve the performance of the conventional speaker identification system, this work is applying to the identification problem with the auditory features that are inspired by signal processing strategies discovered in early and cortical stages of the auditory system and evaluating the benefits by employing such auditory representations. The peripheral cochlea and primary cortex models used here for generating auditory representations

have been analyzed and demonstrated successfully in several signal processing contexts [25, 26, 27, 28]. In brief, the peripheral model estimates the spectral profiles by passing the acoustic signals into a wavelet filter bank coupled with nonlinear compression and reduction stages. After that, a complex wavelet transform based cortical model decomposes the estimated spectrum, called the *auditory spectrum*, at different levels of resolution and produces a multiscale representation that is referred as the *cortical representation* in this work.

To justify the attributes of robustness of these auditory features for speaker identification problem, the system performance is investigated under acoustic environments of different computer simulated noise-level for both auditory spectrum and cortical representation. The Gaussian distributed *auditory cepstrum* which, in principle, conveys the same amount of information as the auditory spectrum is employed to test the identification system (a Bayes classifier) and shows more robust results than the traditionally well-studied mel-cepstrum feature. Furthermore, the two-dimensional multiscale cortical representation yields better performance than LPC spectrum when using the correlation coefficient as a similarity measure. In addition, the phase responses show better robustness than the amplitude responses in the cortical transform and possess even better performance than the auditory cepstrum with a probabilistic approach (Bayes classifier). Since the double wavelet transform (cortical representation) is somewhat ,in spirit, analogous to the double Fourier transform (cepstrum) except the wavelet transform retains much 'local' information and the Fourier transform focuses on the 'global' shape, it is not surprising that the multiresolution cortical representation yields better results than the single resolution cepstrum feature. To sum up, this work demonstrates the superior robustness of the

auditory-based representations than the traditional vocal-based representations for a realistic speaker identification application.

4.2 Future Studies

These realistic speaker identification experiments are conducted to support the feasibility of the auditory approach for the speech signal processing. Accordingly, a considerable amount of techniques applied to the traditional speech recognition problem may be also adapted for the auditory processing to improve the recognition algorithms.

Vector quantization (VQ) has been traditionally used as a compression algorithm for speech processing application [12]. It is motivated by Shannon's rate-distortion theory that a better trade-off between the amount of compression and distortion can be achieved by directly coding the vectors instead of simply coding the scalar components. However, due to the similar goals of compression and classification, not only a compressor but a classifier can VQ be viewed as. For example, the compression can be viewed as a form of classification since it assigns a template or codeword to groups of input speech in a manner that provides a good approximation to the input. However, these two performance measures (compression vs. classification) are competing and bring out the *multi-objective* design problem. Modifying the cost function by incorporating the Bayes risk (cost on misclassification) and minimizing it is one approach for solving this problem.

Given a VQ encoder/decoder pair γ ,δ , the average distortion can be ex-

pressed as

$$D(\gamma, \delta) \;=\; E[\rho(\mathbf{f}, \delta(\gamma(\mathbf{f})))]$$

where $\mathbf{f}$ is the feature vector and ρ is the distortion (distance) function. In general, a quadratic function $\rho(\mathbf{f}, \delta(\gamma(\mathbf{f}))) = \|\mathbf{f} \Leftrightarrow \delta(\gamma(\mathbf{f}))\|^2$ is used as the distortion measure. On the other hand, given a decision rule d, the Bayes risk of the classifier can be measured as [11]

$$J_B(\gamma, d) \;=\; \sum_i \sum_j p(d(\gamma(\mathbf{f})) = j | \mathbf{f} \in i) p(\mathbf{f} \in i) L_{ij}$$

where L_{ij} is the relative cost assigned to the decision that $d(\gamma(\mathbf{f})) = j$ while the $\mathbf{f}$ actually comes from class i. One important observation is that the decoder δ does not affect the Bayes risk J_B. To implement the idea of solving the multi-objective problem, the ordinary distortion and classification error have to be considered simultaneously:

$$J_\lambda(\gamma, \delta, d) = D(\gamma, \delta) + \lambda J_B(\gamma, d)$$

Clearly, λ can be thought as a measure of the relative emphasis to put on the compression and classification ($\lambda \to 0$ corresponding to regular VQ, while $\lambda \to \infty$ corresponding to Bayes classification). The design of the VQ encoder/decoder pair proceeds by employing a descent algorithm to minimize $J_\lambda(\gamma, \delta, d)$ with respect to γ, δ and d.

Due to its inherent hierarchical structure, a hierarchical tree-structured VQ (TSVQ) algorithm is potentially helpful in analyzing the wavelet-based multiscale cortical representation. This algorithm has been successfully applied to various engineering tasks such as the automatic target recognition (ATR) system based on the analysis of different range resolution radar returns [1]. Certainly,

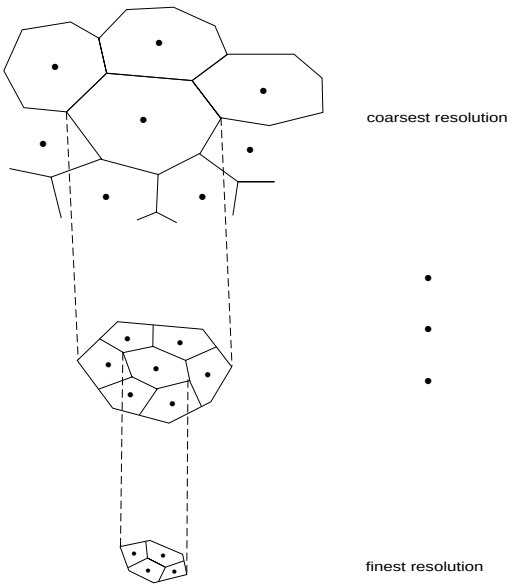

Figure 4.1: TSVQ cells based on different resolution data (adapted from [1]).

the procedures for optimizing the multi-objective VQ can be extended for the TSVQ algorithm. The basic idea of TSVQ algorithm is to partition the signal space, at each resolution, into different clusters or cells. As illustrated in Figure 4.1, the coarsest approximation of the data vector is used to provide partial classification and finer details are added progressively until satisfactory performance such as a requisite stopping level for J_λ is met.

Applying the TSVQ algorithm to the multiscale cortical representations of the spectral profile, one can generate a speaker model for identification process by clustering higher level features of these profiles. This would proceed by first computing the multiscale representation of a large number of speakers under different conditions then partitioning the signal space, at each scale, into different cells. An important insight is that the clustering process at the same resolution is believed based on features that "belong together". Therefore, the result of this clustering is a hierarchical organization of the signal space into cells reflect-

ing different speakers with varying degrees of resolution. The interpretation of these cells likely depends on the exact nature of the signal database. For instance, clusters at different scales may signify vocal tract configurations which reflect phonemic classes, male/female or dialect region distinctions. Investigating the significance and connecting with psychophysical meaning for each resolution should provide a biology-based speaker identification method that potentially offers advantages in noise-robustness, perceptual relevance and identification accuracy.

Bibliography

[1] J. S. Baras and S. I. Wolk. Hierarchical wavelet representations of ship radar returns. Technical Report TR93-100, Institute for Systems Research, University of Maryland, 1993.

[2] Y. Bennani. Text-independent talker identification system combining connectionist and conventional models. *IEEE, Neural Networks for Signal Processing*, pages 131–138, 1992.

[3] W. Byrne, J. Robinson, and S. Shamma. The auditory processing and recognition of speech. *Proc. DARPA Workshop on Speech Recognition*, pages 325–331, November 1989.

[4] C. M. del Álamo, F. C. Gil, C. de la Torre Munilla, and L. H. Gómez. Discriminative training of gmm for speaker identification. In *Proc. ICASSP*, pages 89–92, 1996.

[5] H. Gish and M. Schmidt. Text-independent speaker identification. *IEEE Signal Processing Magazine*, pages 18–32, October 1994.

[6] R. C. Gonzalez and R. E. Woods. *Digital Image Processing.* Addison-Wesley Publishing Company, Inc., 1992.

[7] S. Hayakawa and F. Itakura. The influence of noise on the speaker recognition performance using the higher frequency band. In *Proc. ICASSP*, pages 321–324, 1995.

[8] B. hwang Juang and K. K. Paliwal. Hidden markov models with first-order equalization for noisy speech recognition. *IEEE Transactions on Signal Processing*, 40(9):2136–2143, September 1992.

[9] R. A. Johnson and D. W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice-Hall, Inc., 1992.

[10] E. A. Patrick. *Fundamentals of Pattern Recognition*, chapter 3. Prentice-Hall, Inc., 1972.

[11] K. O. Perlmutter, S. M. Perlmutter, R. M. Gray, R. A. Olshen, and K. L. Oehler. Bayes risk weighted vector quantization with posterior estimation for image compression and classification. *IEEE Transactions on Image Processing*, 5(2):347–360, February 1996.

[12] L. Rabiner and B. hwang Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

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

[14] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17:91–108, 1995.

[15] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, 3(1):72–83, January 1995.

[16] O. Rioul and M. Vetterli. Wavelets and signal processing. *IEEE Signal Processing Magazine*, pages 14–38, October 1991.

[17] M. Schmidt, H. Gish, and A. Mielke. Covariance estimation methods for channel robust text-independent speaker identification. In *Proc. ICASSP*, pages 333–336, 1995.

[18] S. Shamma and H. Versnel. Ripple analysis in the ferret primary auditory cortex. ii. prediction of unit responses to auditory spectral profiles. *J. Auditory Neuroscience*, 1995.

[19] S. Shamma, H. Versnel, and N. Kowalski. Ripple analysis in the ferret primary auditory cortex. i. response characteristics of single units to sinusoidally ripple spectra. *J. Auditory Neuroscience*, 1995.

[20] S. S. Stevens and J. Volkmann. The relation of pitch to frequency: A revised scale. *The American of Journal of Psychology*, 53:329–353, July 1940.

[21] H. Versnel, S. Shamma, and N. Kowalski. Ripple analysis in the ferret primary auditory cortex. iii. topographic and columnar distribution of ripple response parameters. *J. Auditory Neuroscience*, 1995.

[22] S. Vranić-Sowers. *Modeling the perception of profile changes*. PhD thesis, Department of Electrical Engineering, University of Maryland, 1993.

[23] K. Wang. *Neural computations in the auditory system for acoustical information processing*. PhD thesis, Department of Electrical Engineering, University of Maryland, August 1994.

[24] K. Wang and S. A. Shamma. Zero-crossings and noise suppression in auditory wavelet transformations. Technical Report TR92-94, Institute for Systems Research, University of Maryland, 1992.

[25] K. Wang and S. A. Shamma. Self-normalization and noise-robustness in early auditory representations. *IEEE Transactions on Speech and Audio Processing*, 2(3):421–435, 1994.

[26] K. Wang and S. A. Shamma. Auditory analysis of spectro-temporal information in acoustic signals. *IEEE Engineering in Medicine and Biology*, pages 186–194, March/April 1995.

[27] K. Wang and S. A. Shamma. Spectral shape analysis in the central auditory system. *IEEE Transactions on Speech and Audio Processing*, 3(5):382–395, 1995.

[28] X. Yang, K. Wang, and S. A. Shamma. Auditory representations of acoustic signals. *IEEE Transactions on Information Theory, Special Issue on Wavelet Transforms and Multiresolution Signal Analysis*, 38(2):824–839, March 1992.