

國立交通大學

電信工程研究所

碩士論文

時域-頻域上的聽覺頻譜平滑化之  
強健性語者辨識

**Spectro-temporal Smoothed Auditory  
Spectra for Robust Speaker Recognition**

研究生：林廷翰

Student: Ting-Han Lin

指導教授：冀泰石 博士

Advisor: Dr. Tai-Shih Chi

中華民國九十九年七月

時域-頻域上的聽覺頻譜平滑化之

強健性語者辨識

**Spectro-temporal Smoothed Auditory Spectra for  
Robust Speaker Identification**

研究生：林廷翰

Student: Ting-han Lin

指導教授：冀泰石 博士

Advisor: Dr. Tai-shih Chi



A Thesis

Submitted to Institute of Communication Engineering  
College of Electrical and Computer Engineering

National Chiao-Tung University

In Partial Fulfillment of the Requirements

for the Degree of

Master of Science in

Communication Engineering

July 2010

Hsin-Chu, Taiwan, Republic of China

中華民國九十九年七月


# 時頻上的平滑聽覺頻譜之語者辨識

學生：林廷翰

指導教授：冀泰石 博士

國立交通大學電信工程研究所

## 中文摘要



傳統使用的語者辨識系統，辨識率很容易受到加成性雜訊及摺積性雜訊干擾，這是由於傳統上使用的特徵參數只有表達出語句最低層的線索，而較高層的線索被證實出對雜訊較具有抗雜性。本篇論文利用聽覺模型抽取出的語音特徵參數，和時域-頻域調變特性來處理並補捉較高層的線索，最後應用於雜訊下的語者辨識。本論文使用文句不限定及封閉集合語者辨識系統，使用 TIMIT 和 GRID 語料庫進行測試，而實驗結果顯示所提出的參數在各個 SNR 環境下，辨識率比傳統的 MFCC 參數大大提升；而時域-頻域調變濾波器與最近提出的 ANTCC 相比，在低 SNR 下有優異的表現。

# Spectro-temporal Smoothed Auditory Spectra for Robust Speaker Recognition

Student: Ting-han Lin

Advisor: Dr. Tai-shih Chi

Institute of Communication Engineering  
National Chiao-Tung University



The performance of conventional speaker recognition systems is severely compromised by interference, such as additive or convolutional noises. High-level information of the speaker is considered more robust cues for recognizing speakers. This paper proposes an auditory-model based spectral features, auditory cepstral coefficients (ACCs), and a spectro-temporal modulation filtering (STMF) process to capture high-level information for robust speaker recognition. Text-independent closed-set speaker recognition experiments are conducted on TIMIT and GRID corpora to evaluate the robustness of ACCs and benefits of the STMF process. Experimental results show ACCs' significant improvement over conventional MFCCs in all SNR conditions. The superior performance of STMF to newly developed ANTCCs is also demonstrated in low SNR conditions.

# 誌 謝

能夠完成這篇論文，首先感謝兩年來諄諄教誨的冀泰石教授，老師就像一個寶庫，我念碩士的這兩年，從老師身上學到許多的知識寶藏，老師不僅在學業研究上耐心指導，也會在做事處事、個性態度上給我們建議，把我們每個學生從裡到外磨成一塊玉，讓我能脫胎換骨，很開心能夠成為您的學生☺。

感謝實驗室的學長們大師、阿郎及 NICK 在課業上及研究上的指點；同屆的藍雲、勝哥、大樹及禮偉一起研究奮鬥，一同玩樂與磨練；學弟妹華山、雞排、文中及靖雯給實驗室帶來新的歡樂；IT LAB 很 nice 的小玄子、谷嶸及 nino；語音 LAB 的學長們江振宇、楊智合及黃信德，很親切地幫了許多忙。也謝謝許多的同學們及朋友們的幫助。

另外謝謝女朋友 pico，在我最煩躁乏味時陪伴我，用開心的笑容給我鼓舞為我打氣。最後感謝我的父母一直默默的在背後當我的後盾，才能順利的走到今天的成就，爸媽我愛你們，也感謝小妹的陪伴，讓宅宅的老哥開心地過這兩年。

謝謝大家陪伴我這兩年來，生命因你們更添光亮。

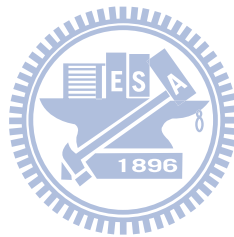
廷翰  
2010 年夏

# Contents

<b>Chinese Abstract</b> .....	i
<b>English Abstract</b> .....	ii
<b>Acknowledgement</b> .....	iii
<b>Contents</b> .....	iv
<b>List of Figures</b> .....	vi
<b>List of Tables</b> .....	vii
<b>Chapter 1 Introduction</b> .....	1
1.1 Introduction .....	1
1.2 Motivation.....	3
1.3 Outline of this thesis.....	6
<b>Chapter 2 Speaker Recognition Systems</b> .....	7
2.1 Introduction to Speaker Recognition Systems.....	7
2.2 Gaussian Mixture Models .....	9
2.3 Maximum A Posteriori Adapted Gaussian Mixture Models .....	14
<b>Chapter 3 Auditory Model and Features</b> .....	17
3.1 The Motivated Use of Auditory Model .....	17
3.2 Cochlear Module and Auditory cepstral coefficients.....	19
3.3 Cortical Module and Spectro-temporal Modulation Filtering .....	23
<b>Chapter 4 Evaluation</b> .....	30
4.1 Database and Evaluation Measurements.....	30
4.2 Results.....	32



4.2.1 Results in GMM.....	32
4.2.2 Results in MAP-GMM.....	38
4.3 Discussions .....	44
<b>Chapter 5 Conclusion and Future Works.....</b>	<b>46</b>
<b>REFERENCES.....</b>	<b>47</b>



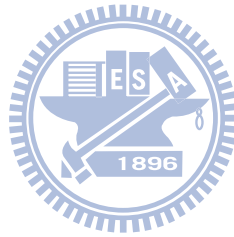
## List of Figures

FIGURE 1-1 Structure of Speaker Recognition System.....	2
FIGURE 1-2 A summary of features from viewpoint of their physical interpretation. ....	3
FIGURE 2-1 Speaker recognition systems. ....	8
FIGURE 2-2 LBG algorithms.....	11
FIGURE 3-1 Hearing pathway.....	18
FIGURE 3-2 The anatomy of the ear. ....	19
FIGURE 3-3 The basilar membrane diagram and the characteristic frequency at the basilar membrane. ....	20
FIGURE 3-4 The firing rate of auditory neuron. ....	20
FIGURE 3-5 Stages of the early cochlear. ....	21
FIGURE 3-6 An example of moving ripple stimulus. ....	24
FIGURE 3-7 The response for 8 modeled neurons in the cortex.....	25
FIGURE 3-8 Rate-scale representation from the A1 module. ....	26
FIGURE 3-9 Auditory spectrograms and Rate-scale representations of clean speech and white noise. ....	27
FIGURE 3-10 Noise suppression by STMF. ....	29
FIGURE 4-1 The 6 characters in GRID corpus.....	31
FIGURE 4-2 Average 0~15dB recognition rates (in %) for GRID corpus.....	32
FIGURE 4-3 Average recognition rates (in %) of 70 people in TIMIT corpus. ....	35
FIGURE 4-4 Average recognition rates (in %) of GRID corpus. ....	37
FIGURE 4-5 Different mixture UBM.....	39
FIGURE 4-6 Different training sentence. ....	39
FIGURE 4-7 Average recognition rates with various adaptations (in %) ....	40
FIGURE 4-8 Average recognition rates (in %) of 70 people. ....	43
FIGURE 4-9 Average recognition rates (in %) of GRID corpus. ....	43



## List of Tables

Table 1 Correct recognition rates (in %) with different STMF ( $\delta$ , $\alpha$ ) parameters under various SNRs of the pink noise.....	33
Table 2 Correct recognition rates (in %) of 70 people in TIMIT corpus. ....	34
Table 3 Correct recognition rates (in %) of GRID corpus. ....	36
Table 4 Correct recognition rates (in %) of 70 people in TIMIT corpus. (UBM stands for MAP-GMM).....	41
Table 5 Correct recognition rates (in %) of GRID corpus. ....	42



# Chapter 1 Introduction

## 1.1 Introduction

In modern lives, identification authentication technologies are commonly used in entrance security systems as well as in portable electronic devices by entering the password or having the magnetic/ID cards scanned to confirm the personal identification. With progresses of the science and technology, any unique characteristic of the human body, such as fingerprints, retinas, facial figures, the voices and so on, is studied and used in advanced identification authentication systems. These biometric verification/recognition technologies are very powerful against intruders who steal passwords or ID cards from authenticated users.

In daily lives, people usually have no problem in recognizing the caller only from his/her voice through communication channels. This is a perfect example of a naive speaker recognition task that people perform everyday. Speaker recognition algorithms have been developed over the last few decades. The basic block diagram of the speaker recognition system is shown in Figure 1-1. Such system consists of three main modules: the feature extraction, speaker models and the recognizer. Basically, feature parameters extracted from input speech are compared with stored models, which are built during training processes, of registered speakers. The recognition decision is made according to certain similarity measures.

Since the characteristics of voice and speaking style are different among people, speaker recognition systems are constructed to recognize speakers by using these features. Conventional approaches adopt short-time spectral features such as Mel-frequency cepstral coefficients (MFCCs), linear predictive cepstral coefficients (LPCCs) and perceptual linear predictive (PLP) coefficients to model each speaker.

These features basically capture smoothed spectral profiles, which reflect vocal tract information of speakers, and usually yield high recognition rates in clean or matched test conditions [1]. However, the recognition performance is often significantly degraded in mismatched testing conditions where speech is deteriorated by either convolutional or additive noises. Therefore, the robustness of speaker recognition systems has drawn a lot attention from researchers. Just like in Nokia's slogan: technology always stems from human nature, the most natural way for humans to verify the identity of a person is by using their sensory inputs, such as from vision (face recognition) and/or hearing (speaker recognition). Our purpose is to develop a robust speaker recognition system, which mimics what humans do, and hopefully to make the world more convenient.

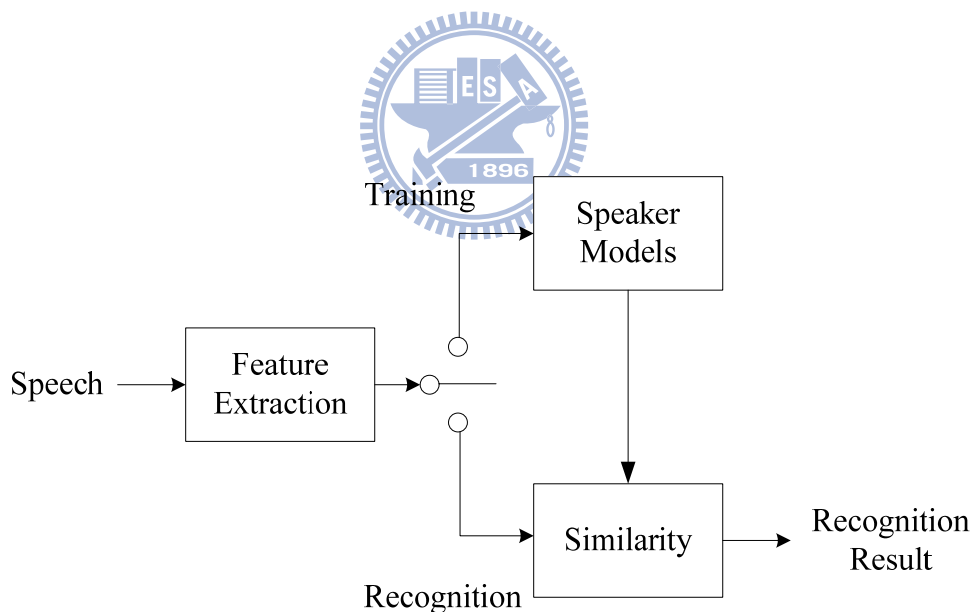


FIGURE 1-1 Structure of Speaker Recognition System

## 1.2 Motivation

Speech is roughly characterized into low-level information and high-level information from the viewpoint of its physical exposition, as shown in Figure 1-2. The low-level information corresponds to characteristics of individual's vocal system, while the high-level information corresponds to characteristics of individual's vocabulary from his cultural or schooling experiences. It is clear that conventional MFCCs or LPCCs only catch the low-level vocal-tract information. On the other hand, the language-dependent speaking rate is considered as a high-level feature. High-level features are believed to be more robust, but less discriminative among speakers in clean environments [2]. It has been shown in [3] that the speaker recognition accuracy can be improved by fusing high- and low-level features.

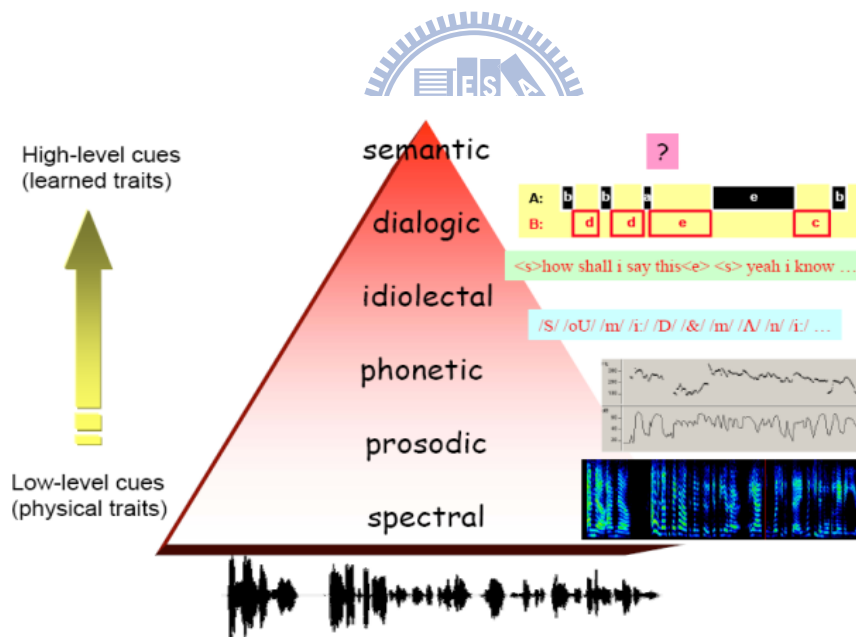


FIGURE 1-2 A summary of features from viewpoint of their physical interpretation. (ACLCLP-vol. 15, no. 5)

Automatic speaker recognition is a tough problem due to the mismatch between handsets and/or channels (convolutional noises) and environmental noises (additive noises), which are two of the most prominent factors to the recognition rate [4]. It has been shown that almost perfect recognition is achievable for clean and well-matched speech. Therefore, researchers have focused on the problems of transducer mismatches and robustness over past years.

To deal with the handset/channel mismatch, linear and nonlinear compensation techniques have been proposed, with applications to feature, model and score domains [5]. The feature compensation is to remove the handset/channel effect on the feature, such as the cepstral mean subtraction (CMS) [6], RASTA [7], discriminative feature design [8], feature mapping [9], and various feature transformation methods such as feature warping [10] and short-time Gaussianization [11]. The score-domain compensation aims to remove handset-dependent biases from the likelihood ratio scores. The most prevalent compensations include the H-norm [12], Z-norm [13], and T-norm [14]. The model-domain compensation is to adapt model parameters to match different handsets/channels. It involves modifying the speaker model parameters instead of the feature vectors. Examples of the model-domain compensation methods include the speaker model synthesis [15], maximum a posteriori (MAP) [16], combination of MLLR and PDBNN [17].

To tackle the environmental noise robustness problem, many speech enhancement techniques have been proposed, for example, spectral subtraction [18], Kalman filtering [19]. They are all based on forming a statistical estimate for removing additive noise. However, noise estimates are never perfect, which may result in removing not only the noise but also speaker-dependent components of the original speech. Other techniques focus on noise compensation, such as model-domain spectral subtraction [20], missing feature theory [5] and parallel model

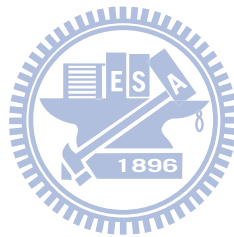
combination [21].

Lippmann demonstrated that human hearing is very robust against any noises in any recognition tests [22]. Presumably, human hearing analyzes every aspect of sounds (low- and high-level) to reduce the noise impact upon the accuracy of recognition tasks. Therefore, it is natural to include psycho-acoustical and neuro-physiological findings about human hearing in the development of speech processing systems to enhance their performance. Recently, several novel features based on human hearing have been proposed and utilized in speaker recognition systems [23]–[25].

In this study, we investigate auditory spectral features and combining with spectro-temporal features for speaker recognition tasks in additive noise environments. The spectral features are referred to as the auditory cepstral coefficients (ACCs), which are derived from the auditory spectrum in a way similar to MFCCs. In addition, high-level constraints are enforced by a spectro-temporal modulation filtering (STMF) process embedded in the auditory model. The two-stage auditory model produces a two-dimensional auditory spectrogram for any input speech, and then analyzes spectro-temporal amplitude modulations of the auditory spectrogram [26]. While low-level spectral features are well preserved in the auditory spectrogram, certain high-level features, such as the speaking rate, are embedded in the spectro-temporal modulations of the speech. Therefore, adopting spectro-temporal modulation features extracted by the auditory model shall enhance the robustness of a speaker recognizer, hopefully, like people have.

## 1.3 Outline of this thesis

The remainder of this thesis is organized as follows. Chapter 2 describes the auditory perceptual model and speaker recognition systems. Our proposed method would be presented in Chapter 3. Simulation results by using proposed features in noisy conditions are demonstrated in Chapter 4. We end in Chapter 5 with conclusions and future works.



# Chapter 2 Speaker Recognition Systems

In this chapter, we briefly review the speaker recognition systems. The most commonly used approaches in speaker recognition systems are the Gaussian Mixture Model (GMM) [1] as well as the Maximum A posteriori Adapted Gaussian Mixture Model (MAP-GMM) [27]. We introduce these two models in section 2.2 and 2.3.

## 2.1 Introduction to Speaker recognition systems

As human beings, we are able to recognize someone just by hearing his or her voice. Usually, a few seconds of speech are sufficient for humans to identify a familiar voice. Similarly, the automatic speaker recognition systems employ computational algorithms to recognize humans by their voices.

Speaker recognition systems can be divided into two different tasks: speaker identification and speaker verification tasks, as shown in Figure 2-1. Speaker verification, or speaker authentication, is the computational task of deciding whether a speech utterance is delivered by a claimed speaker or not. More formally, it is the task of deciding, given a speech signal  $x$  and a hypothesized speaker  $S$ , whether  $x$  was spoken by  $S$ . This is referred to as the one-to-one decision. On the other hand, there is no a priori identity claim in the speaker identification task. Speech from an unknown speaker is compared against trained speech models of  $N$  known speakers, and the best matching speaker is reported as the recognition decision. This is referred to as one-to- $N$  decision.

Speaker recognition tasks can be further categorized into text-dependent and text-independent tasks. The difference between these two tasks is whether using the same utterance in training and testing procedures. In the text-dependent task, the



utterance is known to the recognition systems beforehand. Undoubtedly, the text-independent task is more flexible. Furthermore, speaker identification task can include closed- and open-set tasks. In the former case, the identification system chooses the best matching speaker from trained models - no matter how poor this match is. In the latter case, a predefined tolerance level is considered to prevent the wrong recognition.

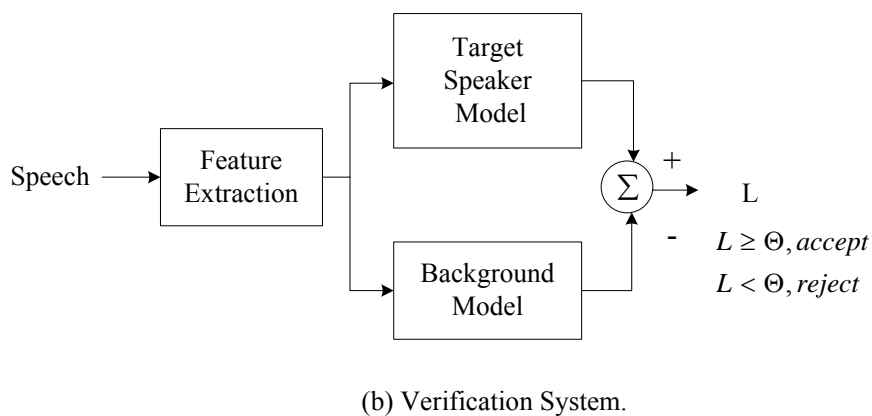
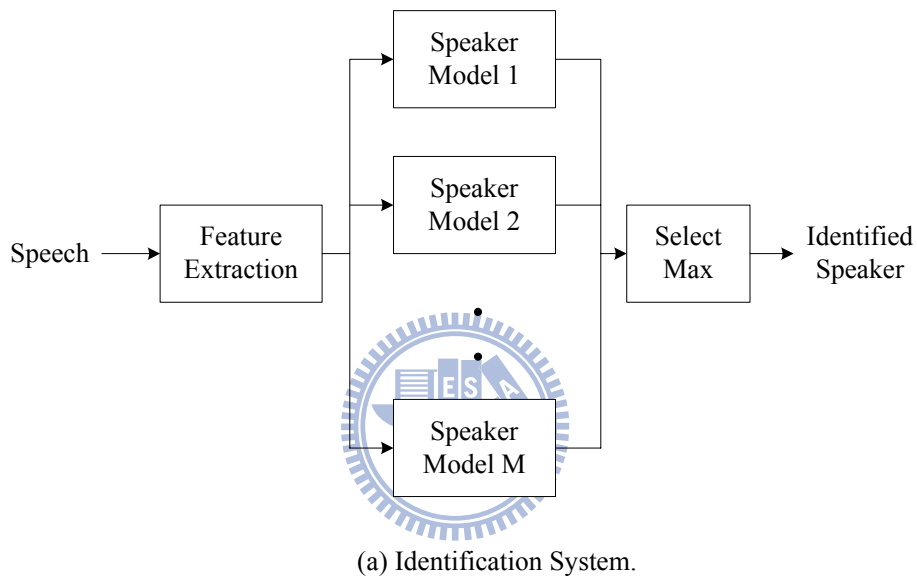


FIGURE 2-1 Speaker recognition systems.

In this thesis, we tackle the text-independent closed-set speaker identification problem. The state-of-the-art closed-set speaker modeling methods are the Gaussian Mixture Model (GMM) and the Maximum A posteriori Adapted Gaussian Mixture Model (MAP-GMM) (which is also named Gaussian Mixture Model Universal Background Model (GMM-UBM)). In the following section, we introduce two commonly used statistical modeling methods for estimating parameters of the GMM.

## 2.2 Gaussian Mixture Model

The Gaussian Mixture Model (GMM) is a stochastic model, which is composed of a finite number of mixtures of multivariate Gaussian components, to fit an observed probability density function (PDF). Since the GMM can fit arbitrary shapes of PDF of features from speaker's voice and its training is simple, fast and giving good performances, it has become the default reference method in any speaker recognition systems.

A GMM, denoted by  $\lambda$ , is characterized by its probability density function:

$$p(\bar{x}|\lambda) = \sum_{i=1}^M w_i N_i(\bar{x}|\bar{\mu}_i, \Sigma_i) \quad (2-1)$$

where  $\bar{x}$  is D-dimensional feature vector; M is the number of Gaussian mixtures;  $w_i$  is the prior probability (mixtures weight) of the i-th mixture with constrain  $\sum_{i=1}^M w_i = 1$ ;

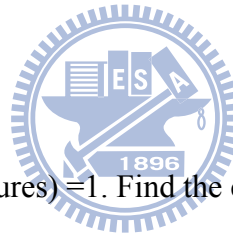
and  $N_i(\bar{x}|\bar{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{M/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x} - \bar{\mu}_i)^T \Sigma_i^{-1} (\bar{x} - \bar{\mu}_i)\right\}$ ,  $i = 1 \dots M$ , is the i-th

Gaussian density function with  $D \times 1$  mean vector  $\bar{\mu}_i$  and  $D \times D$  covariance matrix  $\Sigma_i$ . For numerical and computational reasons, we use only diagonal covariance matrices in this thesis. In general, estimating parameters of a full-covariance GMM

requires much more training data and is computationally expensive. And empirical evidence shows that diagonal matrix GMMs can perform equally well or out-perform the full matrix GMMs [27].

Training a GMM is to estimate the parameters  $\lambda = \{\bar{\mu}_i, \Sigma_i, w_i\}_{i=1}^M$  from a given collection of training vectors. The basic approach uses Vector Quantization (VQ) to get the initial parameters, and then the maximum likelihood (ML) model parameters are estimated via the expectation-maximization (EM) algorithm.

Vector Quantization (VQ) is one of the most efficient and useful methods in source-coding techniques. The disorderly speech feature vector distribution can be classified into codewords by VQ. The following states the LBG algorithm [28] which is commonly used in VQ.



### LBG algorithm

- 1. Set M (number of mixtures) = 1. Find the centroid of all feature vectors.
- 2. Split M into 2M partitions. As shown in equation (2-2), where  $\delta = 0.01$ .

$$\begin{aligned}\bar{\mu}_i^+ &= \bar{\mu}_i (1 + \delta) \\ \bar{\mu}_i^- &= \bar{\mu}_i (1 - \delta)\end{aligned}\tag{2-2}$$

- 3. Use K-means iterative algorithm to re-classify feature vectors and find the new centroid of each partition until old centroids equal to new centroids.
- 4. Repeat step 2 and 3 until meet the desired total number of mixtures.
- 5. Calculate variance  $\bar{\sigma}_i^2$  and weight  $w_i$ . As shown in equation (2-3), where n is the total feature number;  $n_i$  is the feature number in the i-th partition;  $\sigma_{ik}^2$  is the variance in the i-th partition and j-th dimension.

$$\sigma_{ik}^2 = \frac{\sum_j (x_{jk} - \mu_{ik})^2}{n_i}, \text{ calculated } x_{jk} \text{ which belongs to the centroid } \mu_{ik}$$

$$w_i = \frac{n_i}{n}$$

(2-3)

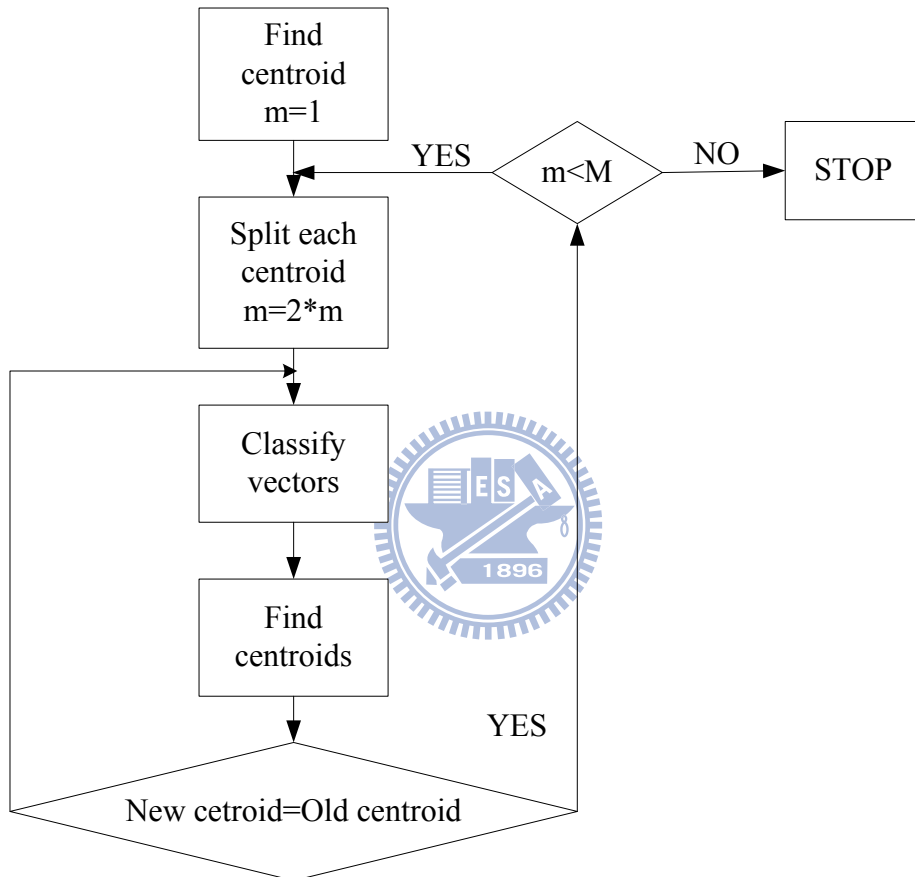


FIGURE 2-2 LBG algorithms.

Although VQ can classify the speech feature vectors by centroids, it can't describe the size and spatial shape of the speech feature vector distribution of each partition. As a result, the ML estimation via the EM algorithm for model parameters is used after VQ. For a set of i.i.d. feature vectors  $\mathbf{x} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T\}$ , the ML estimate of parameters of a GMM, is:

$$\begin{aligned}\lambda_{ML} &= \arg \max p(\lambda|\bar{x}) = \arg \max \frac{p(\bar{x}|\lambda)p(\lambda)}{p(\bar{x})} \\ &= \arg \max p(\bar{x}|\lambda)\end{aligned}\quad (2-4)$$

Therefore, the ML estimation is to find the best  $\lambda$  to achieve the highest probability. The ML parameter estimation can be accomplished iteratively via the EM algorithm. The basic idea of the EM algorithm is, beginning with initial parameters  $\lambda = \{\bar{\mu}_i, \Sigma_i, w_i\}_{i=1}^M$  obtained from VQ, to estimate a new model  $\hat{\lambda}$  provided  $p(\bar{x}|\hat{\lambda}) \geq p(\bar{x}|\lambda)$ . The new model  $\hat{\lambda}$  then becomes the initial model for the next iteration and the process is repeated until some convergence criterion is met. Generally speaking, five to ten iterations are sufficient for parameters convergence.

1. E-Step:  $Q(\lambda|\lambda^k) = E[\log p(x|\lambda)|y, \lambda^k]$
2. M-Step:  $\arg \max_{\lambda} Q(\lambda|\lambda^k)$

Given a training feature vector set  $\mathbf{x} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T\}$ , the following re-estimation formulae, which guarantee a monotonic converge to a local maximum, are used in each EM iteration:

**Posteriori probability:**

$$p(i|\bar{x}_t, \lambda) = \frac{w_i N(\bar{x}_t|\lambda_i)}{\sum_{j=1}^M w_j N(\bar{x}_t|\lambda_j)}, \quad i=1 \dots M \quad (2-5)$$

**Weight:**

$$\hat{w}_i = \frac{1}{T} \sum_{t=1}^T p(i|\bar{x}_t, \lambda) \quad (2-6)$$

**Mean:**

$$\hat{\mu}_i = \frac{\sum_{t=1}^T p(i|\bar{x}_t, \lambda) \cdot \bar{x}_t}{\sum_{t=1}^T p(i|\bar{x}_t, \lambda)} \quad (2-7)$$

**Variance:**

$$\hat{\sigma}_i^2 = \frac{\sum_{t=1}^T p(i|\bar{x}_t, \lambda) \cdot \bar{x}_t^2}{\sum_{t=1}^T p(i|\bar{x}_t, \lambda)} - \hat{\mu}_i^2 \quad (2-8)$$

To avoid the variance converging to zero, a variance floor of 0.000001 is adopted.

After estimating each speaker's model parameters  $\{\bar{\lambda}_1, \bar{\lambda}_2, \dots, \bar{\lambda}_S\}$ , the speaker identification decision is made according to the probability measures as follows:

$$\text{speaker ID} = \arg \max_{1 \leq k \leq S} p(\bar{x}|\lambda_k) \quad (2-9)$$

## 2.3 Maximum A Posteriori Adapted Gaussian Mixture Model

Conventional GMMs trained by the EM algorithm as depicted in Section 2.2 perform well when a large amount of training data is available to characterize speakers. In other words, training a GMM model for a particular speaker needs data from all his possible pronunciation. However, in the real world, this condition is somehow not feasible. Adaptations of the acoustic models have been studied for solving this problem. One successful adaptation approach, namely the Universal background model-maximum a posteriori (UBM-MAP) approach, has been widely used in text-independent speaker verification tasks in recent years.

The UBM is a large GMM which is usually set with 256~2048 mixtures depending on the size of the training data. Lower order mixtures are often used in applications with constrained speech (such as digits or a fixed vocabulary), while 2048 mixtures are used with unconstrained speech (such as the conversational speech). This approach firstly pools a huge amount of speech data gathered from a large number of background speakers to train a universal background model (UBM) by the LBG and EM algorithm. Unlike the standard approach of maximum likelihood training of a model for a particular speaker, independently with the UBM, this adaptation approach is to derive the speaker's model by adapting the well-trained UBM parameters to a speaker model  $\lambda$  using this speaker's training speech via the MAP estimation technique. For a UBM and training vectors from the hypothesized speaker,  $\mathbf{x} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T\}$ , the MAP estimate of parameters of the speaker GMM is:

$$\begin{aligned}\lambda_{MAP} &= \arg \max p(\lambda | \bar{x}) = \arg \max \frac{p(\bar{x} | \lambda) p(\lambda)}{p(\bar{x})} \\ &= \arg \max p(\bar{x} | \lambda) p(\lambda)\end{aligned}\tag{2-10}$$

Like the EM algorithm, the adaptation is a two step of the EM algorithm. The specifics of the adaptation are as follows. Given training vectors from target speaker  $\mathbf{x} = \{\bar{x}_1, \bar{x}_2, \dots, \bar{x}_T\}$  and a UBM  $\lambda_{UBM} = \{\bar{\mu}_i, \Sigma_i, w_i\}_{i=1}^M$ , we first compute the posteriori probability  $p(i|\bar{x}_t, \lambda_{UBM})$  for the mixture  $i$  in the UBM:

**Posteriori probability:**

$$p(i|\bar{x}_t, \lambda) = \frac{w_i N(\bar{x}_t | \lambda_i)}{\sum_{j=1}^M w_j N(\bar{x}_t | \lambda_j)}, \quad i=1 \dots M \quad (2-11)$$

We then use  $p(i|\bar{x}_t, \lambda_{UBM})$  and  $\bar{x}_t$  to compute the sufficient statistics for the mixture weight, mean, and second moment parameters:

**Weight:**

$$n_i = \sum_{t=1}^T p(i|\bar{x}_t, \lambda) \quad (2-12)$$

**Mean:**

$$E_i(\bar{x}_t) = \frac{1}{w_i} \sum_{t=1}^T p(i|\bar{x}_t, \lambda) \cdot \bar{x}_t \quad (2-13)$$

**Second moment:**

$$E_i(\bar{x}_t^2) = \frac{1}{w_i} \sum_{t=1}^T p(i|\bar{x}_t, \lambda) \cdot \bar{x}_t^2 \quad (2-14)$$



This is the same as the expectation step in the EM algorithm. Finally, these new statistics from the training data are used to update the old UBM statistics for mixture  $i$  to create the adapted parameters for mixture  $i$  with the equations:

$$\begin{aligned}\hat{w}_i &= [\alpha_i n_i / T + (1 - \alpha_i) w_i] \gamma \\ \hat{\mu}_i &= \alpha_i E_i(\bar{x}_t) + (1 - \alpha_i) \mu_i \\ \hat{\sigma}_i^2 &= \alpha_i E_i(\bar{x}_t^2) + (1 - \alpha_i) (\sigma_i^2 + \mu_i^2) - \mu_i^2\end{aligned}\quad (2-15)$$

where the scale factor  $\gamma$  is computed over all adapted mixture weights to ensure their summation equals unity and the adaptation coefficients  $\alpha_i$  controlling the balance between the old and new estimates. It is defined as:

$$\alpha_i = \frac{n_i}{n_i + r} \quad (2-16)$$

where  $r$  is a fixed relevance factor which can be viewed as an adaptation coefficient. If  $r$  is large, adaptation is slow and if  $r$  is small, adaptation is fast. We set  $r = 16$  in this thesis. In [27], only adapting the means shows the best performance in simulations. The performance of adapting the means and variances is similar to the performance of adapting the means only. After adaptation, the mixture components of the adapted GMM for each speaker retain a correspondence with the mixtures of the original UBM.

## Chapter 3 Auditory Model and Features

The speaker recognition system is described in Chapter 2. This chapter provides a brief review of the auditory model, which contains an early cochlear (ear) and a central auditory cortex (A1) module, proposed by Shamma et al. [26]. The auditory spectral features are extracted from the early cochlear module and used in the speaker recognition simulations. The Spectro-Temporal Modulation Filtering (STMF) is performed by the cortical module and produces cleaned spectral features. This approach was first investigated by Hung [29] in digit recognition tasks.

### 3.1 The Motivated Use of Auditory Model

In recent years, there is an increasing interest in adopting properties of human hearing perception for speech-related applications to overcome various types of distortion such as additive noises, convolutional noises and degradations from channel mismatch. It has been shown that human hearing is very robust against any noises in any recognition tests [22].

For instance, perceptual linear predictive (PLP) coefficients [30], which are one of the most used coefficients, embed two hearing perception properties: the equal loudness pre-emphasis and the intensity-loudness conversion. To compensate the fact that humans have non-equal hearing thresholds at different frequencies, the speech power spectra are multiplied by the magnitude response of an equal loudness pre-emphasis filter. The intensity-loudness conversion addresses the non-linear relation between the intensity of the sound level and the perceived loudness. And it has been shown that PLP is more robust to noise than the LPCC (linear predictive cepstral coefficients).

Here, we evaluate the performance of a noise suppression algorithm, which works on the internal perceptual representation of an auditory model, in text-independent speaker identification tasks and compare its robustness to other features or algorithms. The auditory model is inspired by psycho-acoustical and neuro-physiological findings along the mammal's hearing pathway: the cochlea and the cortex. Figure 3-1 shows a schematic plot of the auditory pathway. In the following sections, we introduce these two stages (cochlear and cortical stages) and related functions in the auditory model.

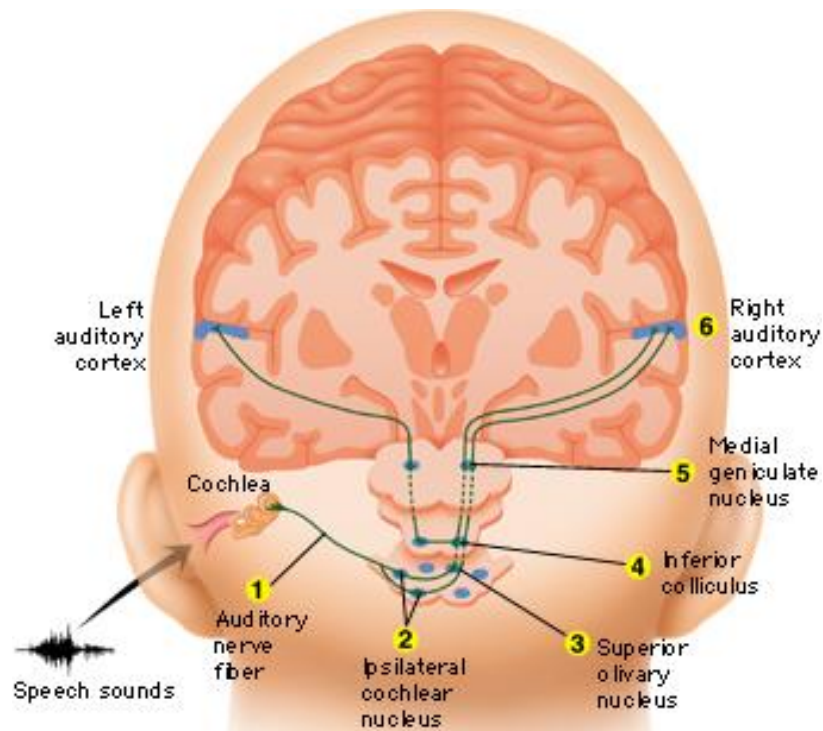


FIGURE 3-1 Hearing pathway.

(<http://brainconnection.positscience.com/topics/?main=anat/auditory-anat2>)

## 3.2 Cochlear Module and Auditory cepstral coefficients

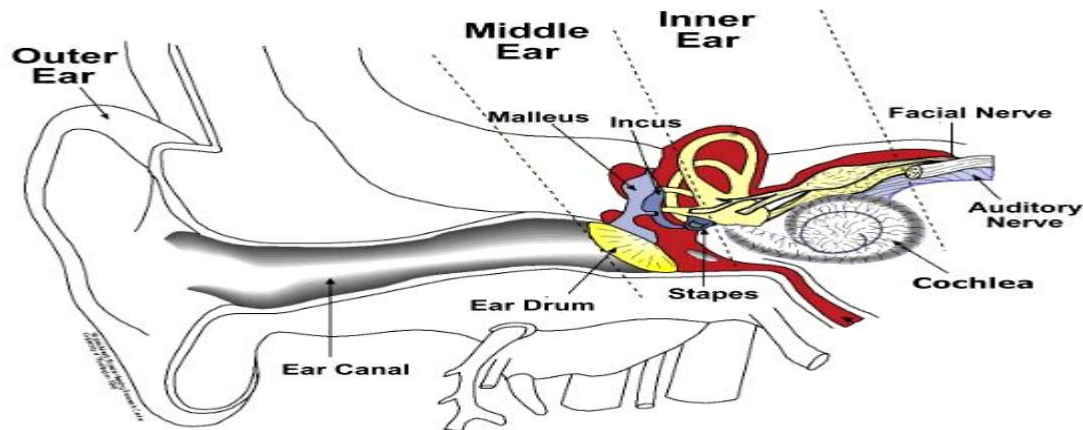


FIGURE 3-2 The anatomy of the ear.  
([http://www.advcoch.com/I2\\_Hearing\\_Physiology.htm](http://www.advcoch.com/I2_Hearing_Physiology.htm))

The ear can be divided into three parts – the outer ear, middle ear and inner ear, which are shown in Figure 3-2. The inner ear consists of the cochlea, which is composed of three chambers with full lymph, as shown in the top left panel of Figure 3-3. The basilar membrane (BM) dividing the scala media and the scala tympani plays a significant role in hearing. After the mechanical vibration reaches the oval window, a traveling wave is generated and propagates along the basilar membrane. Different locations of the BM achieve the maximum responses with respect to traveling waves with different frequencies. The right panel of Figure 3-3 shows the responsive frequencies along the basilar membrane. The inhibitions between neighboring frequencies produced by the traveling wave might be the main cause of the well-known “frequency masking” phenomenon in audition.

The traveling wave generates displacement along the BM, and then the hair cells residing along the basilar membrane transform the displacement to sensory nerve action potentials. There are two different hair cells: inner hair cells and outer hair cells. Most of the transformation from mechanical vibrations to electrical potentials is done

by inner hair cells, which connects with the auditory nerve. Due to the fact that a relaxation time is needed between consecutive firings of neurons, firing rates can not keep up with high frequency inputs, as demonstrated in Figure 3-4. The firing rate of the auditory nerve is bounded by 4-5k Hz and the rate of the midbrain is bounded by about 1k Hz.

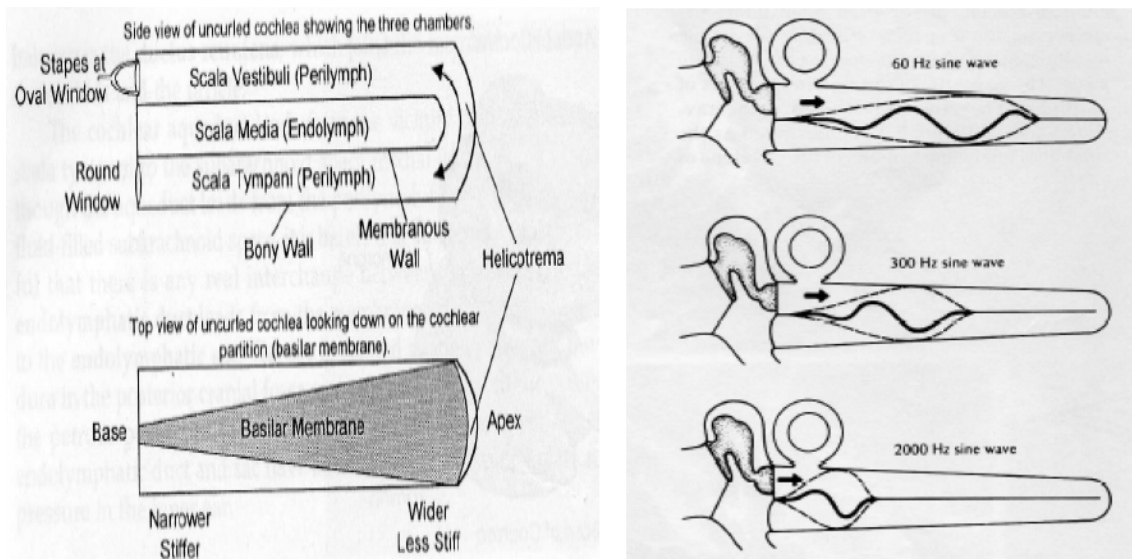


FIGURE 3-3 The basilar membrane diagram (left) and the characteristic frequency at the basilar membrane (right). (Hearing Physiology Handout, AAIP)

### Auditory Nerve Fiber Discharge: Firing Rate

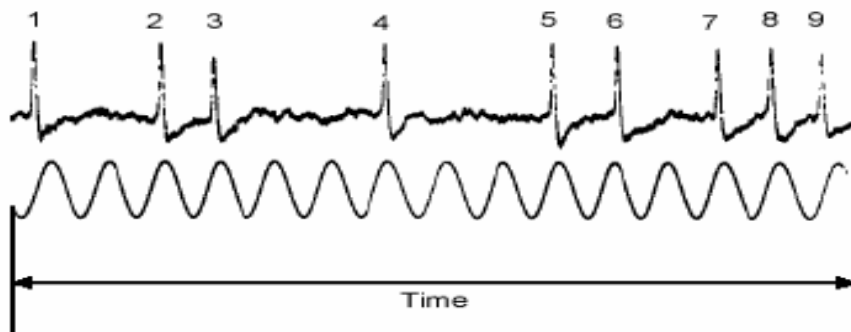


FIGURE 3-4 The firing rate of auditory nerve correspond to the monotone audio input. (Hearing Physiology Handout, AAIP)

Figure 3-5 depicts the early cochlear module and the derivation of auditory cepstral coefficients (ACCs). The speech signal is first filtered by a set of 128 overlapping asymmetric constant-Q filters whose magnitude responses can be expressed as:

$$|H(f)| = \begin{cases} (f_h - f)^\alpha e^{-\beta(f_h - f)}, & 0 \leq f \leq f_h \\ 0, & f > f_h \end{cases} \quad (3-1)$$

where  $f_h$  is the cut-off frequency (in log-frequency axis) and  $\alpha=0.3$  and  $\beta=8$ . These cochlear filters evenly distribute over 5.3 octaves with 24 filter/octave frequency resolution. The cochlea filters represent the selectivity of the basilar membrane to different frequencies.

The output of each filter is then passed through a lateral inhibitory network (LIN), a half-wave rectifier and a lowpass filter. The LIN is implemented by a first order differentiator along the log-frequency axis to roughly account for the frequency masking effect between neighboring neurons. It equivalently sharpens the frequency response of each cochlear filter. The half-wave rectifier combined with the following lowpass filter extract the envelope of filtered speech in each cochlear band.

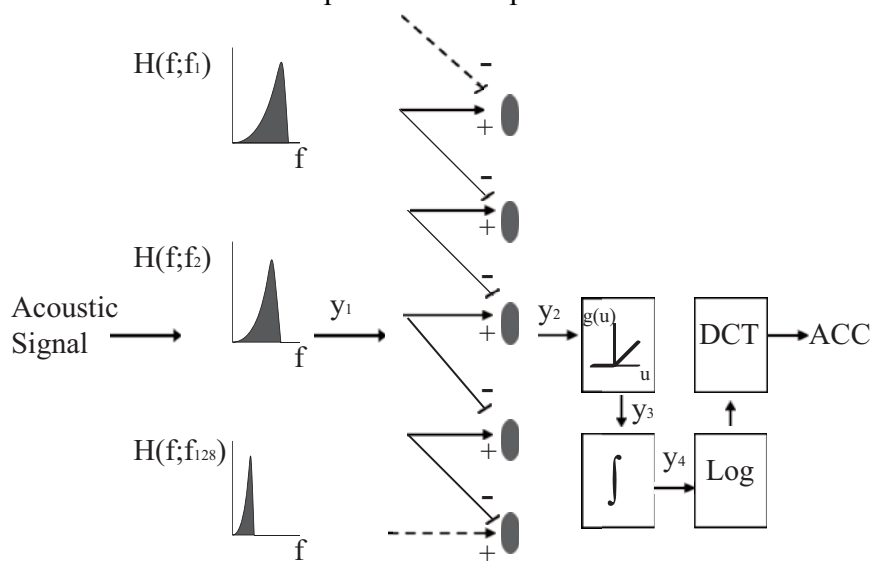


FIGURE 3-5 Stages of the early cochlear module.

Outputs at various stages can be stated as follow:

$$y_1(t, f_i) = s(t) *_t h(t; f_i) \quad (3-2)$$

$$y_2(t, f_i) = \partial_f y_1(t, f_i) = y_1(t, f_i) - y_1(t, f_{i-1}) \quad (3-3)$$

$$y_3(t, f_i) = \max(y_2(t, f_i), 0) \quad (3-4)$$

$$y_4(t, f_i) = y_3(t, f_i) *_t \mu(t; \tau) \quad (3-5)$$

where  $h(t; f_i)$  is the impulse response of the  $i$ -th constant-Q filter with center frequency  $f_i$ ,  $i=1\dots 128$ ;  $*_t$  is the convolution in the time domain; and the integration window  $\mu(t; \tau) = e^{-t/\tau} \cdot u(t)$  models the current leakage along the neural pathway to the auditory cortex.

The output  $y_4(t, f_i)$  is referred to as an auditory spectrogram which captures spectro-temporal envelopes of input speech along the log-frequency and the time axes. Similar to the derivation of conventional MFCCs, the Auditory Cepstral Coefficients (ACCs) are obtained by the discrete cosine transform (DCT) on the logarithm amplitude of the auditory spectrum at any time instant. The ACCs  $A(k, t)$   $k=0\dots N-1$ , can be written as:

$$A(k, t) = \sqrt{\frac{2}{N}} \sum_{i=1}^N \ln |y_4(t, f_i)| \cos\left(\frac{\pi k}{N} (i - 0.5)\right) \quad (3-6)$$

Intuitively, ACCs represent smoothed auditory spectra, which reflect vocal tract information of the speaker, along human's hearing pathway.

## 3.3 Cortical Module and Spectro-temporal Modulation

### Filtering

The second cortical module is inspired from neural activities of the auditory cortex (A1) to different spectro-temporal variations. Such spectro-temporal variations are encoded in two parameters: rate and scale. The rate (or velocity) parameter  $\omega$  in Hz depicts how fast the signal's energy varies along the temporal axis. The scale (or density) parameter  $\Omega$  in cycle/octave characterizes how broad the signal's energy distributed along the log-frequency axis. In addition, cortical neurons also show different selectivity of FM sweeping directions (upward and downward), which is represented in this module by the sign of the rate parameter (positive/negative for downward/upward sweeping direction).

To derive the spectro-temporal impulse responses of neurons in A1, moving ripple stimuli, the basis functions in the two-dimensional spectro-temporal domain, are used to drive the cortex. Figure 3-6 shows one example of the moving ripple stimulus of rate=+4 Hz and scale=0.5 cycle/octave. Therefore, each neuron in A1 has its own impulse response, which represents its preference on the spectro-temporal pattern shown in the input spectrogram, and is modeled by a 2D filter. To sum up, the first cochlear module of the auditory model produces a two-dimensional auditory spectrogram full of spectro-temporal amplitude modulations. The second cortical module then analyzes the auditory spectrogram by a bank of two-dimensional filters which are tuned to different spectro-temporal modulation parameters. Figure 3-7 demonstrates eight 2D cortical filtering of A1 on a sample spectrogram. The small top panels in each subplot are the impulse responses of different typical neurons tuned to slow/fast rates and coarse/fine scales. The bottom panels are envelopes (local energies) of outcomes of these 2D spectro-temporal filters.



Therefore, a four-dimensional output  $r(t, f, \omega, \Omega)$  of this module can be formulated as:

$$r(t, f, \omega, \Omega) = y_4(t, f) *_{tf} STIR(t, f; \omega, \Omega) \quad (3-7)$$

where  $STIR(t, f; \omega, \Omega)$  is the spectro-temporal impulse response of the two-dimensional filter tuned to  $\omega$  and  $\Omega$ ; and  $*_{tf}$  is the two-dimensional convolution in the time and log-frequency axes.

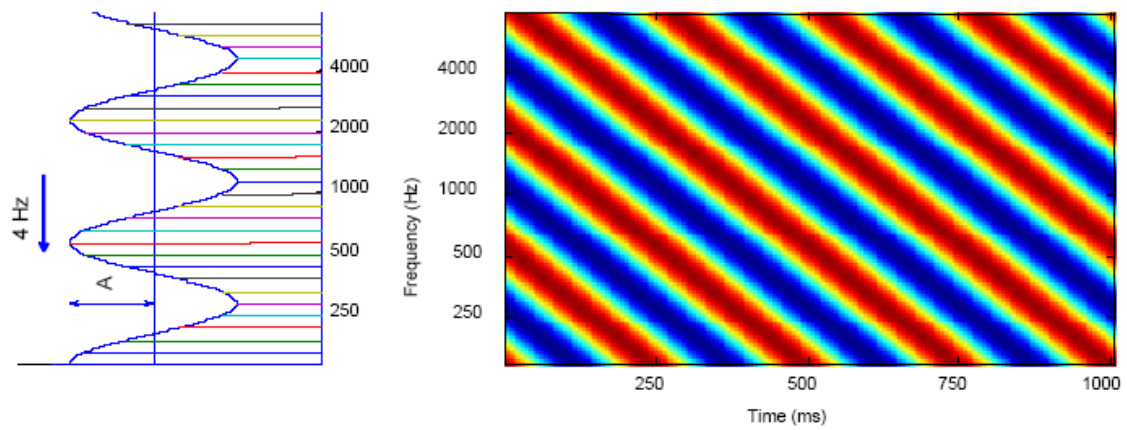


FIGURE 3-6 An example of moving ripple stimulus.

( Auditory Model Handout, AAIP)

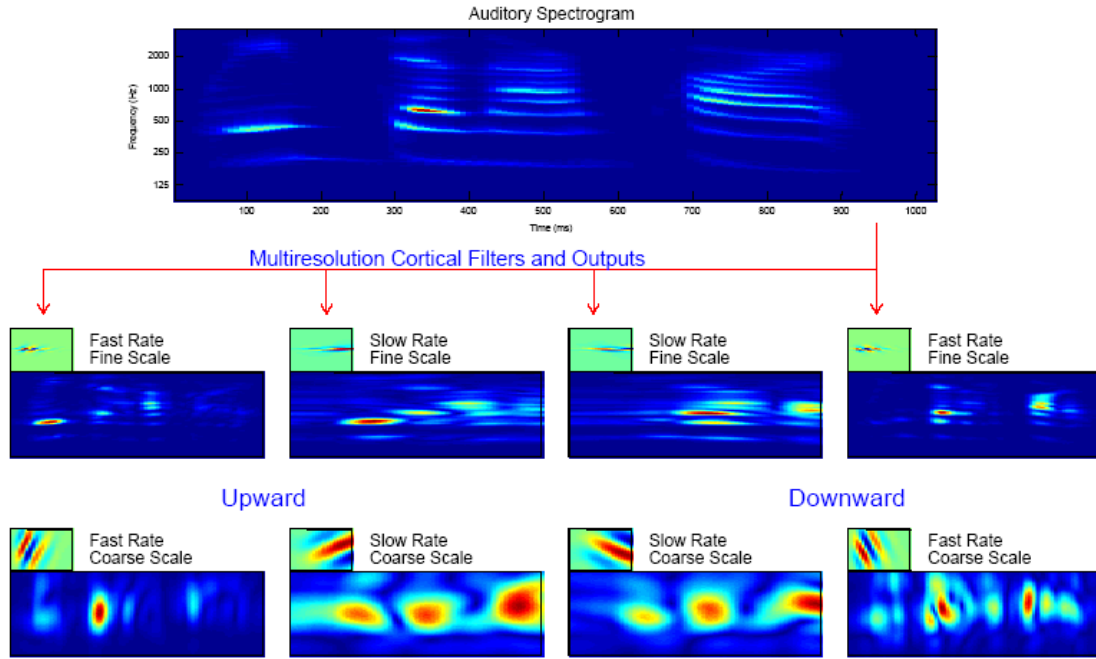


FIGURE 3-7 The response for 8 modeled neurons in the cortex. (Auditory Model Handout, AAIP)

The local energy of the four-dimensional output is then computed as:

$$E(t, f, \omega, \Omega) = \left| r(t, f, \omega, \Omega) + jH[r(t, f, \omega, \Omega)] \right| \quad (3-8)$$

where  $H[\cdot]$  is the Hilbert transform along the log-frequency axis. Therefore, for any fixed t-f point in the auditory spectrogram,  $E(\omega, \Omega; t, f)$ , which is referred to as the rate-scale representation, records energies of local modulations at different combinations of rate, scale and directionality. As shown in Figure 3-8, the left panel demonstrates an auditory spectrogram and right panels are corresponding rate-scale representations of those two points indicated by 'x' in the spectrogram. As seen in the figure, those two 'x' points have local modulations dominated at (8 Hz, 4 cycle/octave, upward) and (8~16 Hz, 2~4 cycle/octave, downward) respectively.

In summary, the early cochlear module estimates a two-dimensional auditory

spectrogram from a one-dimensional acoustic signal. The second cortical module analyzes amplitude modulations of the 2D auditory spectrogram in the rate-scale-directionality parameter space. Much more extensive details of the description, mathematic formulation and output examples of these two modules can be found in [26].

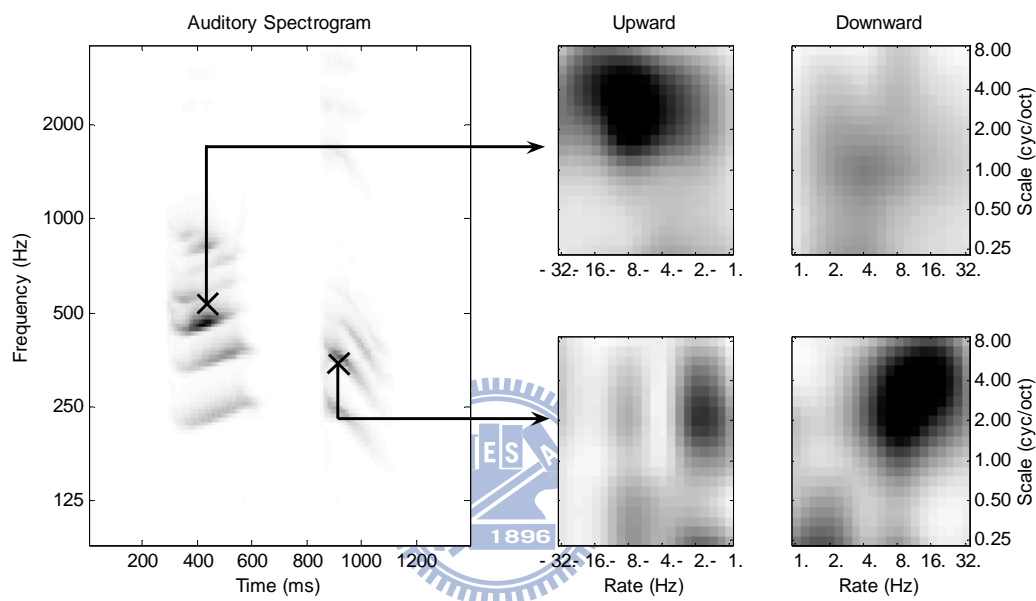


FIGURE 3-8 Rate-scale representation from the A1 module.

It is known that human hearing analyzes not only spectral contents but also temporal behaviors of the sound. In our auditory model, such ability is well characterized by the joint spectro-temporal modulation analysis performed by the second cortical module. In addition to spectral contents estimated in the first cochlear module, certain high-level features, such as speaking rate and FM sweeping directions, are well caught by the second cortical module. It has been shown that joint spectro-temporal modulations below 16 Hz and 8 cycle/octave well preserve the intelligibility of speech [31]. Not surprisingly, as shown in [32], the long-term averaged rate-scale energy pattern of speech falls roughly within these ranges. On the

other hand, rate-scale patterns of noises would differ from those of speech, indicating different high-level information between speech and noises. For example, Figure 3-9 shows auditory spectrograms ((a), (b)) and rate-scale energy representations ((c), (d)) of clean speech and white noise. This figure demonstrates that most of the spectro-temporal modulations of speech are within the range of rate=2-16 Hz and scale=0.5-8 cycle/octave, while the white noise has spectro-temporal modulations dominated at high rates and high scales.

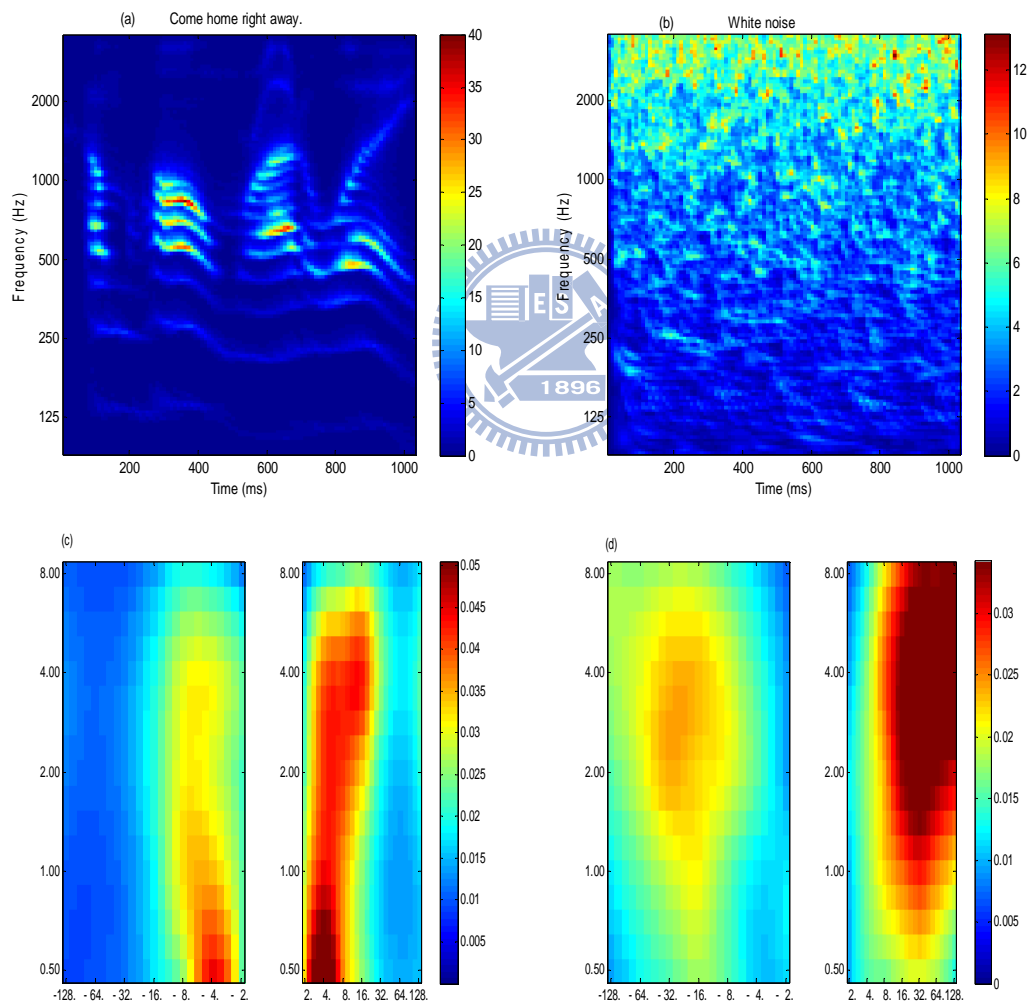


FIGURE 3-9 Auditory spectrograms of (a) clean speech, and (b) white noise. Rate-scale representations (with rate and scale in x- and y- axis) of (c) clean speech, and (d) white noise.

Accordingly, a noise suppression algorithm by the joint spectro-temporal modulation filtering (STMF) is proposed in [29]. For an input noisy speech, spectro-temporal modulations only within 2~32 Hz and 0.5~8 cycle/octave are kept in the STMF process and a cleaner spectrogram is generated:

$$y_5(t, f) = \sum_{\pm 2 \leq |\omega| \leq \pm 32, 0.5 \leq \Omega \leq 8} r(t, f, \omega, \Omega) *_{tf} STIR_1^*(-t, -f; \omega, \Omega) \quad (3-9)$$

where  $STIR_1(t, f; \omega, \Omega)$  is the normalization of  $STIR(t, f; \omega, \Omega)$ .

Figure 3-10 demonstrates procedures of our STMF noise suppression algorithm. The noisy auditory spectrogram is passed through the STMF process. Then, a simple threshold  $\delta$  (a certain percentile of the maximum value of the cleaned spectrogram) is used to determine the speech versus non-speech regions in the cleaned spectrogram. The threshold  $\delta$  bears the trade-off between effects of speech distortion and noise suppression. Finally, a  $\alpha$ — 1 template ( $\alpha$  for non-speech regions and 1 for speech regions) is generated and multiplied with the original noisy spectrogram to produce a noise-suppressed spectrogram. ACCs are then derived from the noise-suppressed spectrogram for our speaker recognition simulations.

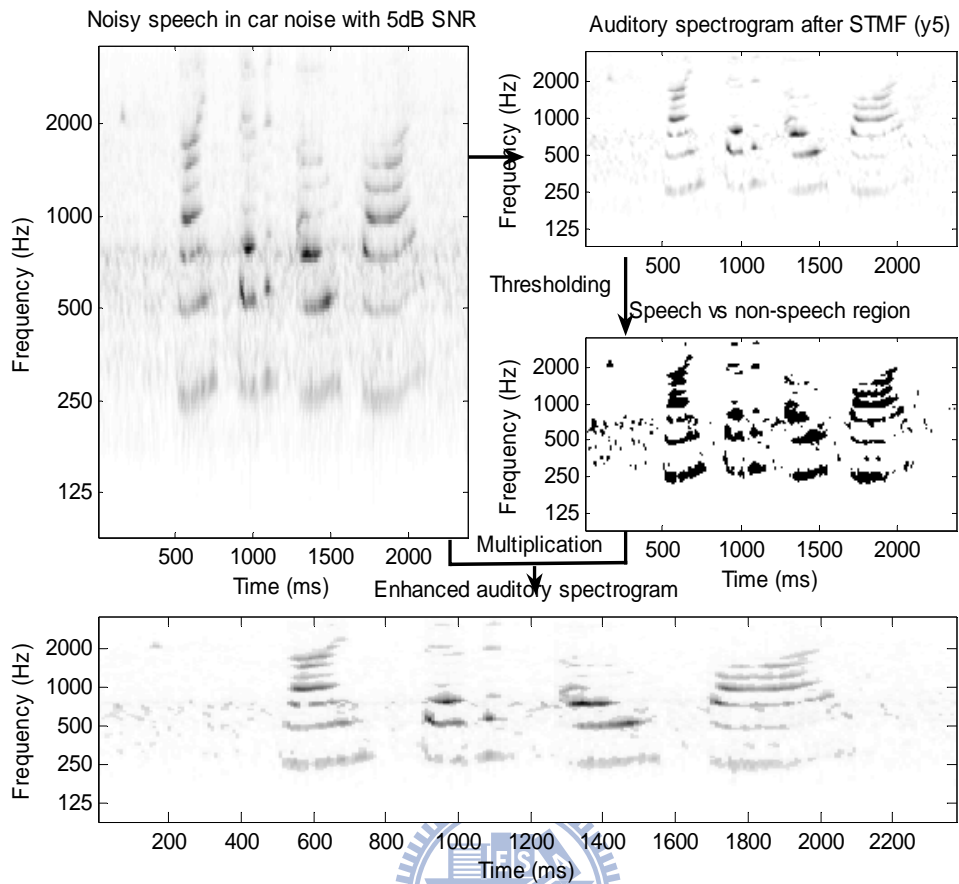


FIGURE 3-10 Noise suppression by STMF.

# Chapter 4 Evaluation

The robustness of ACCs before and after the STMF process is evaluated in text-independent closed-set speaker identification simulations. Experimental settings follow the ones in [25] and results are compared to results from Auditory-based Nonnegative Tensor Cepstral Coefficients (ANTCCs) proposed in [25] as well. Speech samples from TIMIT and GRID [33] corpora are tested. Note that these corpora do not consider session variability. In this chapter, we first introduce the TIMIT and GRID database and the evaluation measurements used in this thesis. Then, simulation results will be shown in Section 4.2. Finally, discussions for these evaluations will be given in Section 4.3.

## 4.1 Database and Evaluation Measurements

Speech samples in TIMIT corpus were recorded at Texas Instruments (TI) and transcribed at Massachusetts Institute of Technology (MIT). Thus, it is called “TIMIT”. TIMIT was designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech and speaker recognition systems. The TIMIT corpus is with high quality speech, which is ideal for testing a technique without the interference of noise and channel variations. It contains a total of 6300 clear sentences, sampled at 16 kHz, 10 sentences uttered by each of the 630 speakers (438 males and 192 females) from 8 major dialect regions in the United States. In this thesis, the first 8 utterances (two sa sentences, three si sentences and two sx sentences) and the remaining 2 utterances per speaker are used as the training and testing sets, respectively. The training data for each speaker is approximately 24 seconds.

The GRID is an audio-visual corpus for the speech separation and recognition tasks. It is also with high quality speech from 34 speakers (18 males and 16 females), each saying 1000 three-second phrases. Each phrase consists of a sequence of 6 characters shown in figure 4-1 and is sampled at 24 kHz. As in [25], for the GRID testing, speech samples are downsampled to 8 kHz and 50/60 utterances per speaker are randomly chosen for training/testing purposes.

<b>command</b>	<b>color</b>	<b>preposition</b>	<b>letter</b>	<b>number</b>	<b>adverb</b>
bin (b) lay (l) place (p) set (s)	blue (b) green (g) red (r) white (w)	at (a) by (b) in (i) with (w)	A – Z excluding W	1-9 and zero (z)	again (a) now (n) please (p) soon (s)

**FIGURE 4-1** The 6 characters in GRID corpus.

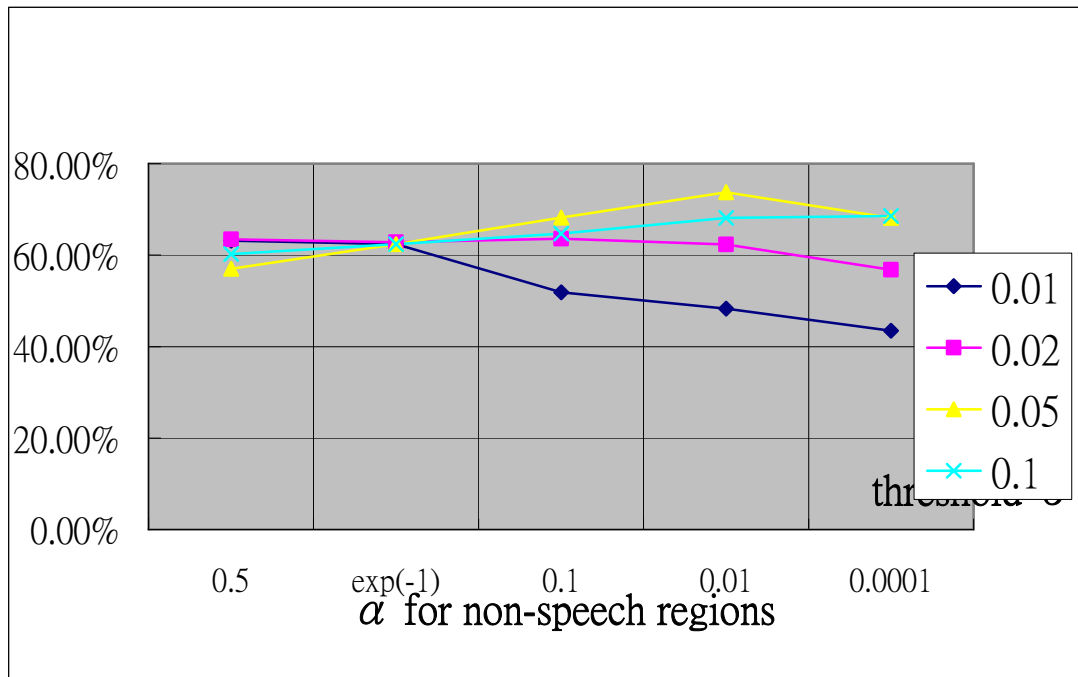
Four different types of noises (factory, pink, white and F-16) are extracted from Noisex-92 [34] and mixed with the clean speech in a wide range of SNRs (0, 5, 10, 15dB and Clean). 30-coefficient feature vectors (excluding the 0th coefficient) of conventional MFCCs and ACCs are calculated from a 25 ms window with 10 ms frame increment. As in [35], each speaker is modeled by a 32-mixture GMM derived from the EM algorithm in the training process. The K-means is used to initialize the EM algorithm. Additionally, a simple technique — cepstral mean subtraction (CMS) [6], which has been used in speaker recognition, is also considered in our works. Note, no VAD is used in this study.



## 4.2 Results

### 4.2.1 Results in GMM

To implement our STMF based speaker identification system, we first determine the parameters, the threshold  $\delta$  and the template  $\alpha$ , used in the STMF process. Figure 4-2 and Table 1 shows the identification performance of a GMM based recognizer on the GRID corpus with different STMF parameter sets. Clearly, the best average recognition rate can be obtained by  $\delta=0.05$  and  $\alpha=0.01$ . However, Table 1 further shows that using  $\delta=0.02$  &  $\alpha=e^{-1}$  outperforms the parameter set of  $\delta=0.05$  &  $\alpha=0.01$  in high SNR conditions. Not surprisingly, it demonstrates the threshold  $\delta$  bears the trade-off between effects of speech distortion and noise suppression. And these two effects are desired contradictorily in high and low SNR conditions. Therefore, we adopt two parameter sets (A:  $\delta=0.02$  &  $\alpha=e^{-1}$ , B:  $\delta=0.05$  &  $\alpha=0.01$ ) in following simulations.



**FIGURE 4-2** Average recognition rates (in %) over 0~15dB of the STMF process with different parameters.

**Table 1.** Correct recognition rates (in %) with different STMF ( $\delta$ ,  $\alpha$ ) parameters under various SNRs of the pink noise.

	$\delta = 0.01$				$\delta = 0.02$			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
$\alpha = 0.5$	90.83	82.21	54.80	24.95	90.88	81.23	58.97	22.75
$\alpha = \exp(-1)$	85.88	74.61	57.25	31.86	90.54	81.49	60.64	19.71
$\alpha = 0.1$	90.20	80.39	30.20	6.86	89.61	83.82	63.68	17.40
$\alpha = 0.01$	88.04	77.99	23.87	3.43	90.10	83.53	66.32	9.36
$\alpha = 0.001$	80.49	66.57	22.50	4.36	86.23	79.61	56.37	4.90
	$\delta = 0.05$				$\delta = 0.1$			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
$\alpha = 0.5$	85.74	74.17	54.61	13.43	85.88	74.07	52.01	28.97
$\alpha = \exp(-1)$	85.88	74.61	57.25	31.86	85.88	74.61	57.25	31.86
$\alpha = 0.1$	85.44	78.28	67.50	41.72	84.95	72.99	58.48	42.30
$\alpha = 0.01$	85.54	81.57	75.83	52.01	83.28	73.68	62.94	52.79
$\alpha = 0.001$	83.58	79.12	71.13	38.77	80.54	74.46	64.22	55.20

Correct recognition rates of the MFCC baseline, our proposed features ACCs before and after STMF, and ANTCCs from [25] are presented in Table 2 and Figure 4-3. The 70-speaker population is randomly chosen from the TIMIT corpus and all speech samples are with 16 kHz sampling frequency. Without CMS, our features and ANTCCs clearly achieve much higher recognition rates than the MFCCs under noisy conditions. Under all tested SNR conditions, ACCs outperform MFCCs by a wide margin. In addition, the ACCs after STMF also outperform ANTCCs in almost all conditions (except in the condition of 15dB F16 noise).

It can be observed that our ACCs perform poorer in the white and F16 noises than in the pink and factory noises, especially in high SNR conditions. One possible reason for that is the white and F16 noises both possess higher energies in high-frequency regions than the pink and factory noises. The 128 constant-Q cochlear filters possess constant frequency resolution and are normalized to have an almost flat

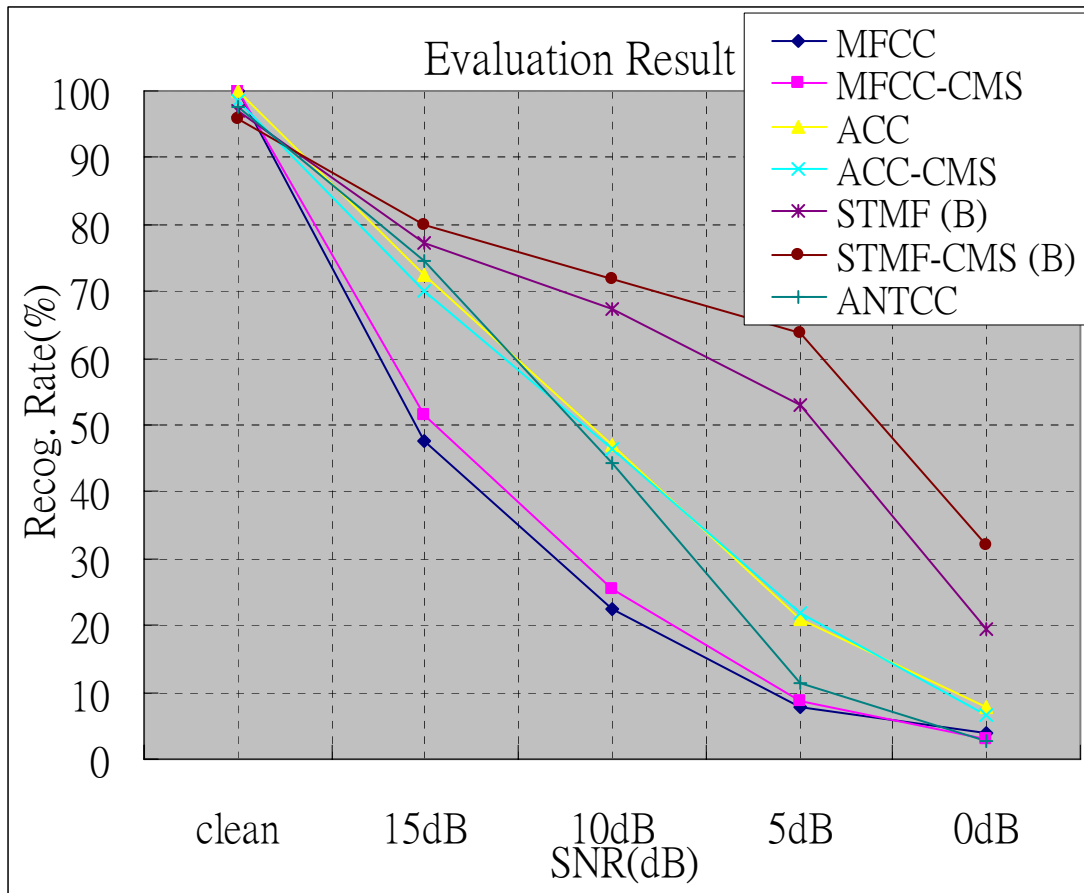
overall frequency response *along the log-frequency axis*. For the white and F16 noises, high energies within high-frequency regions would repetitively appear in several high-frequency cochlear channels due to their wide bandwidth. This phenomenon produces more severe mismatch between ACCs from noisy speech and from clean speech by high-frequency noises than by low-frequency noises.

**Table 2.** Correct recognition rates (in %) of 70 people in TIMIT corpus.

	<b>Factory1</b>				<b>Pink</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC</b>	66.43	34.29	10.71	5.71	47.86	22.86	8.57	2.86
<b>MFCC-CMS</b>	65.00	32.86	11.43	2.86	46.43	21.43	5.00	2.86
<b>ACC</b>	93.57	80.00	37.86	12.86	87.86	62.14	25.00	9.29
<b>ACC-CMS</b>	82.14	52.86	27.86	6.43	73.57	42.14	18.57	4.29
<b>STMF (A)</b>	94.29	86.43	51.43	10.00	90.71	70.00	41.43	7.86
<b>STMF-CMS (A)</b>	85.00	75.00	57.14	11.43	80.00	64.29	45.00	12.14
<b>STMF (B)</b>	85.00	78.57	62.86	21.43	84.29	76.43	64.29	29.29
<b>STMF-CMS (B)</b>	84.29	75.00	66.43	31.43	80.71	72.14	65.00	38.57
<b>ANTCC</b>	78.1	49.52	12.86	2.43	78.57	50.95	13.81	2.43

	<b>White</b>				<b>F16</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC</b>	30.00	15.71	7.86	4.29	45.71	16.43	3.57	2.14
<b>MFCC-CMS</b>	36.43	18.57	10.00	2.86	58.57	28.57	7.86	2.86
<b>ACC</b>	51.43	17.14	10.00	4.29	56.43	28.57	11.43	5.00
<b>ACC-CMS</b>	51.43	34.29	20.71	11.43	73.57	56.43	20.00	4.29
<b>STMF (A)</b>	71.43	45.00	16.43	5.00	69.29	38.57	18.57	4.29
<b>STMF-CMS (A)</b>	67.86	51.43	37.86	12.14	82.14	68.57	55.00	10.71
<b>STMF (B)</b>	72.14	60.71	48.57	16.43	67.86	53.57	36.43	10.71
<b>STMF-CMS (B)</b>	76.43	66.43	60.00	28.57	78.57	73.57	63.57	29.29
<b>ANTCC</b>	64.29	29.52	3.81	2.9	77.62	47.14	15.24	2.9

	<b>MFCC</b>	<b>MFCC-CMS</b>	<b>ACC</b>	<b>ACC-CMS</b>	<b>ANTCC</b>
<b>Clean</b>	100	100	100	98.57	97.62
	<b>STMF (A)</b>	<b>STMF-CMS(A)</b>	<b>STMF (B)</b>	<b>STMF-CMS(B)</b>	
<b>Clean</b>	98.57	98.57	97.14	95.71	



**FIGURE 4-3** Average recognition rates (in %) of 70 people in TIMIT corpus.

On the other hand, the CMS helps the recognition rates of ACCs before and after STMF in low SNR conditions (0~10 dB) as shown in Figure 4-3. However, the CMS could not only mitigate the noise effect, but also reduce the speaker variability. The recognition rates, therefore, might be diminished in high SNR conditions as the clean condition in Table 2 or Figure 4-3.

Evaluation results of the GRID corpus are presented in Table 3 and Figure 4-4. First, we consider results without CMS normalization. Similar to TIMIT results, ACCs outperform MFCCs in all noises (all SNR conditions) and the STMF process enhances recognition rates further. Compared with ANTCCs, STMF perform 25.3% better (average recognition rate) in low SNR conditions (5 and 0 dB), but 10.24% (in average) worse in high SNR conditions (15 and 10 dB). Note, not identical training

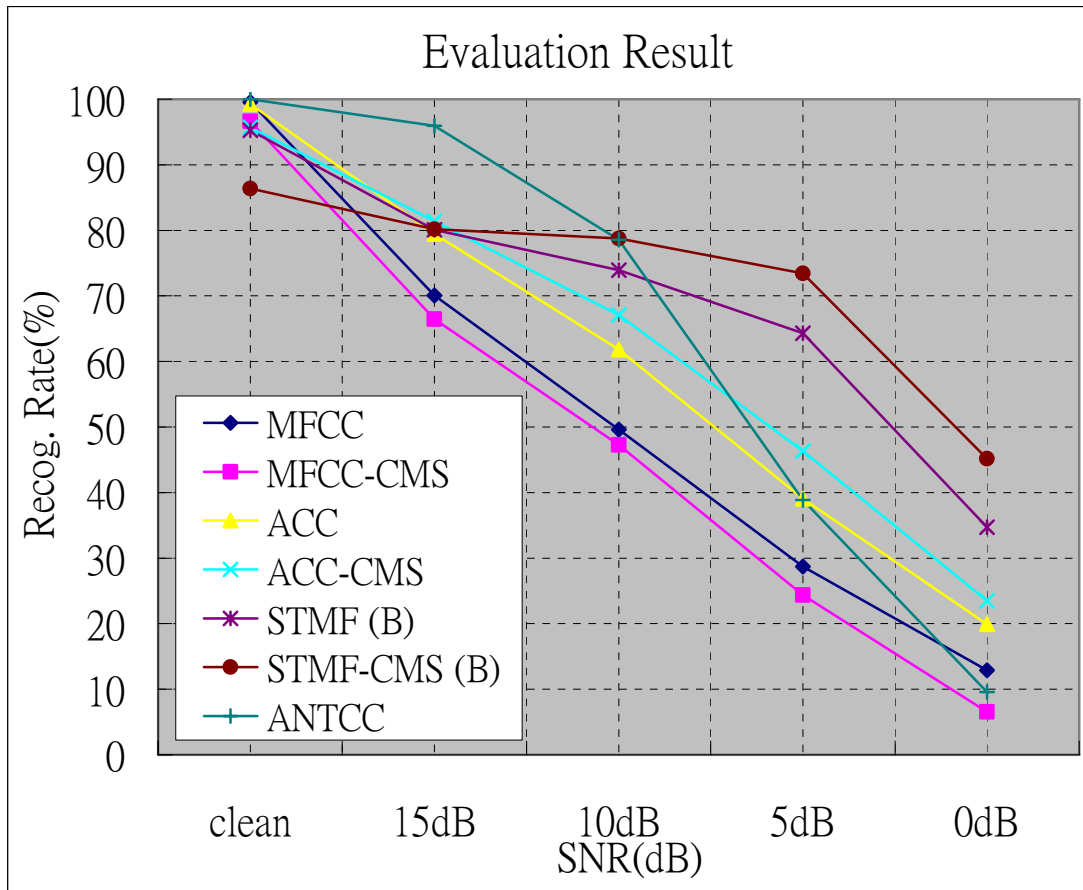
and testing sets are used in this study and in [25] for both TIMIT and GRID corpora evaluations shown in Table 2 and Table 3. Clearly, our features perform slightly worse against the white and F16 noises than against the pink and factory noises in both TIMIT and GRID corpora evaluations.

**Table 3.** Correct recognition rates (in %) of GRID corpus.

	<b>Factory1</b>				<b>Pink</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC</b>	79.46	60.78	34.26	13.97	75.98	51.91	26.27	7.99
<b>MFCC-CMS</b>	73.28	51.72	24.46	5.69	69.12	47.99	22.55	5.83
<b>ACC</b>	87.84	75.29	55.49	31.81	87.79	72.40	49.17	23.14
<b>ACC-CMS</b>	86.32	73.63	52.35	20.34	83.92	70.83	47.94	22.45
<b>STMF (A)</b>	90.39	81.91	58.58	20.39	90.54	80.49	60.64	19.71
<b>STMF-CMS (A)</b>	89.71	84.46	55.78	12.30	88.73	83.24	57.79	11.52
<b>STMF (B)</b>	85.78	81.32	75.00	41.47	85.54	81.57	75.83	52.01
<b>STMF-CMS (B)</b>	81.62	80.00	74.71	46.86	80.15	78.73	75.15	51.32
<b>ANTCC</b>	97.55	87.75	44.61	8.82	95.59	87.75	45.1	9.31

	<b>White</b>				<b>F16</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC</b>	55.74	38.97	27.16	14.61	69.07	46.76	27.06	15.05
<b>MFCC-CMS</b>	52.16	38.58	27.50	8.92	71.18	50.83	22.99	5.78
<b>ACC</b>	76.47	55.29	30.69	14.66	65.69	44.36	20.93	10.10
<b>ACC-CMS</b>	72.25	55.54	40.20	32.06	83.09	68.43	44.90	19.17
<b>STMF (A)</b>	83.28	72.65	45.83	19.71	88.14	82.65	52.55	8.04
<b>STMF-CMS(A)</b>	88.87	84.61	74.61	46.18	90.88	87.65	78.68	36.42
<b>STMF (B)</b>	82.30	77.01	67.06	30.49	66.76	55.83	39.26	14.90
<b>STMF-CMS (B)</b>	78.09	76.81	71.91	45.54	80.83	79.41	72.01	36.86
<b>ANTCC</b>	95.59	69.61	38.24	10.29	95.1	69.12	27.49	9.8

	<b>MFCC</b>	<b>MFCC-CMS</b>	<b>ACC</b>	<b>ACC-CMS</b>	<b>ANTCC</b>
<b>Clean</b>	99.56	96.62	99.26	95.69	100
	<b>STMF (A)</b>	<b>STMF-CMS(A)</b>	<b>STMF (B)</b>	<b>STMF-CMS(B)</b>	
<b>Clean</b>	98.28	94.07	95.25	86.37	

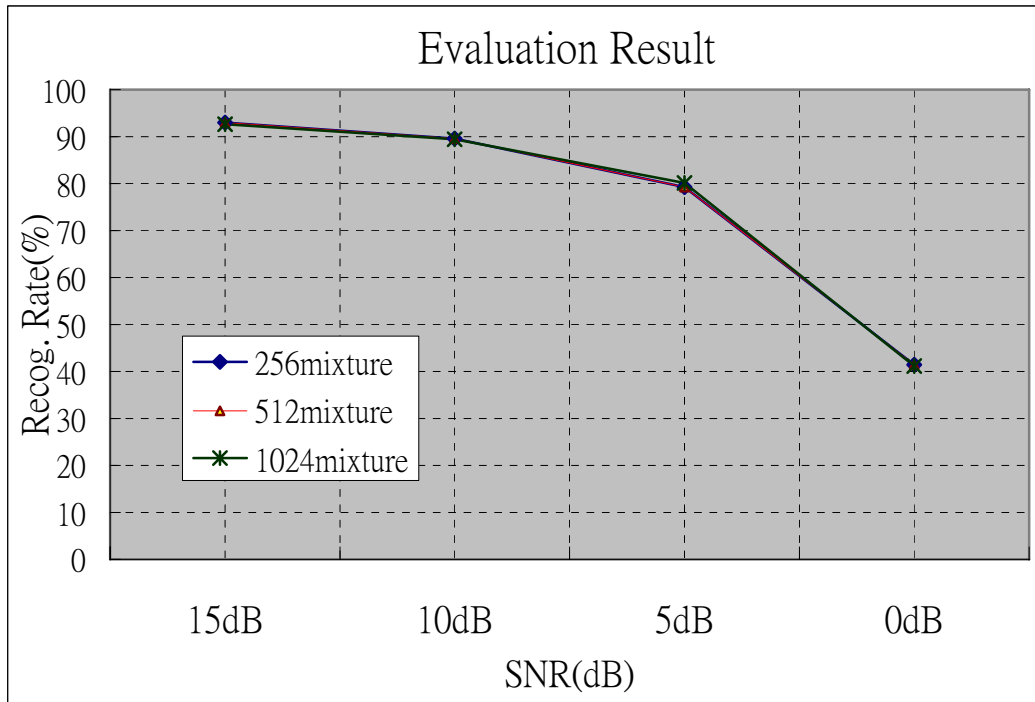


**FIGURE 4-4** Average recognition rates (in %) of GRID corpus.

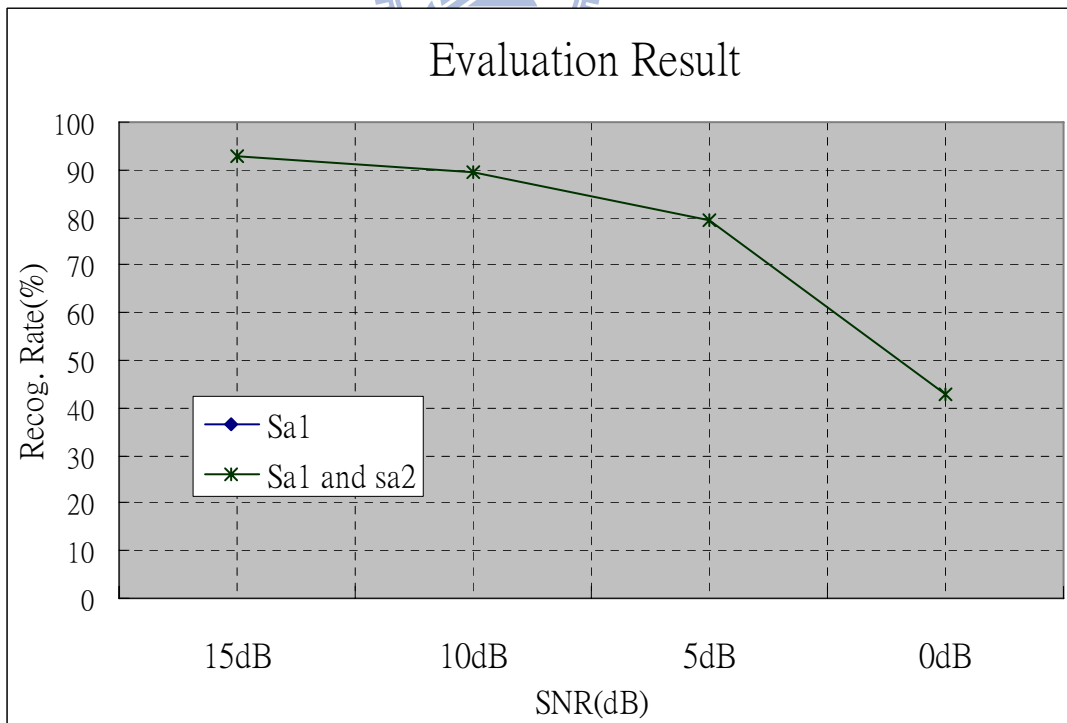
The CMS normalization yields similar trends for ACCs and STMF features in both corpora evaluations as shown in Figure 4-3 and 4-4. That is, the CMS enhances average recognition rates in low SNR conditions but degrades the performance in the 15dB and clean condition. However, the CMS produces worse average performance for MFCCs. It is interesting to note that all features (MFCCs, ACCs and STMF) for both corpora (TIMIT and GRID) benefit from the CMS normalization in the F16 noise as shown in Table 2 and Table 3. We could conclude that the CMS normalization is particularly effective against the F16 noise.

## 4.2.2 Results in MAP-GMM

In this section, we further adopt the MAP-GMM method to boost the performance of the STMF-CMS feature. In implementing the MAP-GMM in the speaker identification system, we first determine the number of mixtures used in the UBM (Universal Background Model). Effects from different numbers of mixtures are investigated by building different UBMs from the sa1 sentence of all 630 speakers in the TIMIT corpus. In order to match testing samples from the GRID corpus with 8 kHz sampling frequency, the training sa1 sentences are downsampled to 8 kHz. Then, each speaker's model is built from his training speech and the UBM by adapting the mean and variance via the MAP-GMM approach as shown in Section 2.3. Recognition rates in testing the GRID corpus using a 256-, 512- and 1024-mixture UBM are listed in Figure 4-5. Since higher-mixture UBMs only slightly improve the performance but with much heavier computational loads, the 256-mixture UBM is used in this thesis. Figure 4-6 shows the effects of building the 256-mixture UBM by different training data sets. The UBM is built either on 630 sa1 sentences only or combined with 630 sa2 sentences. As shown in Figure 4-6, the recognition rates of two different training sets are very close. Therefore, the 256-mixture UBM trained from sa1 sentences is used in succeeding tests.

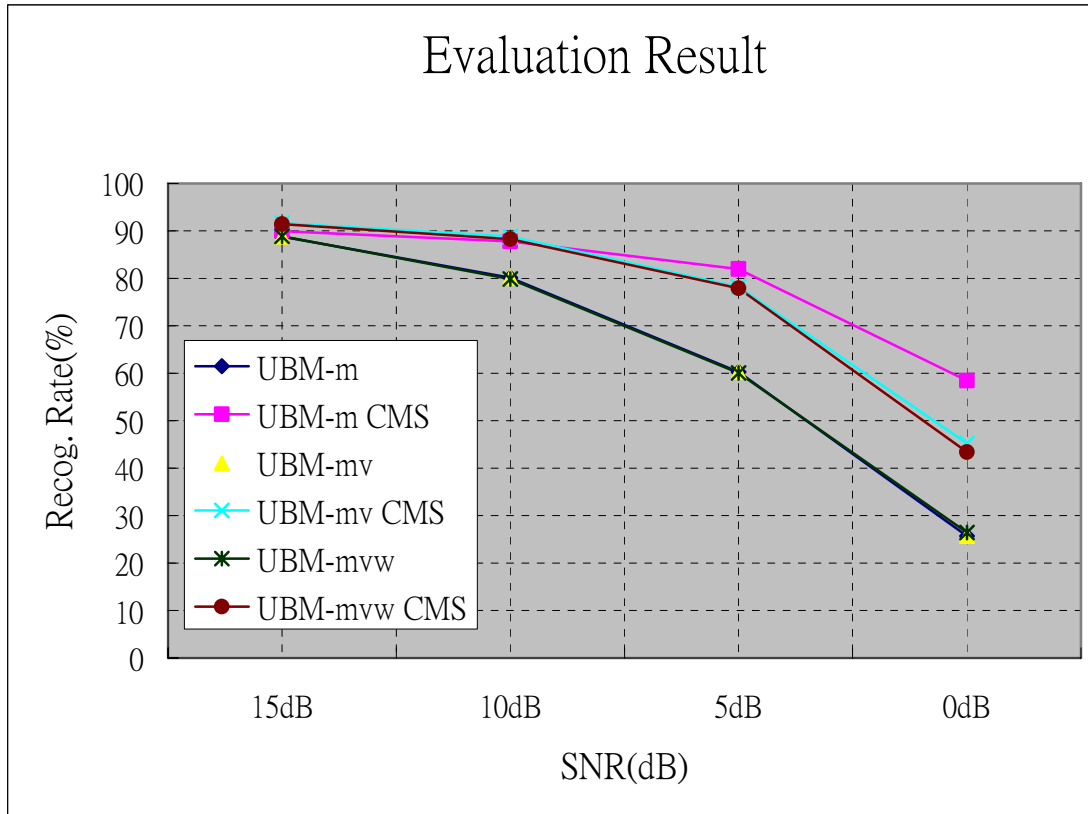


**FIGURE 4-5** Correct recognition rates (in %) for GRID corpus: MAP-GMM (adapted mean and variance) for the STMF-CMS(A) feature with different mixture numbers



**FIGURE 4-6** Correct recognition rates (in %) for GRID corpus: a 256-mixture MAP-GMM for the STMF-CMS(A) feature with different training sets for the UBM





**FIGURE 4-7** Average recognition rates (in %) of GRID corpus using the STMF(A) feature with various adaptations and CMS normalization

The 256-mixture UBM is utilized to evaluate performance by adaptations of parameters (mean, mean and variance, mean variance and weight) with and without the CMS normalization. Figure 4-7 indicates that the best overall performance is from adapting only the mean vectors with the CMS normalization. Thus, simulations in following sections are done in the scenario of a 256-mixture UBM with mean adaptation and the CMS normalization. Two UBMs built from 8 kHz and 16 kHz sampling frequency training sentences are used to match different sampling frequencies of GRID and TIMIT corpora.

**Table 4.** Correct recognition rates (in %) of 70 people in TIMIT corpus.

	<b>Factory1</b>				<b>Pink</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC-GMM-CMS</b>	65.00	32.86	11.43	2.86	46.43	21.43	5.00	2.86
<b>MFCC-UBM-CMS</b>	53.57	33.57	10.00	3.57	42.14	22.86	7.86	3.57
<b>ACC-GMM-CMS</b>	82.14	52.86	27.86	6.43	73.57	42.14	18.57	4.29
<b>ACC-UBM-CMS</b>	91.43	72.86	45.71	16.43	86.43	60.71	29.29	6.43
<b>STMF-GMM-CMS (A)</b>	85.00	75.00	57.14	11.43	80.00	64.29	45.00	12.14
<b>STMF-UBM-CMS (A)</b>	92.14	85.00	62.14	16.43	88.57	78.57	56.43	10.00
<b>STMF-UBM-CMS (B)</b>	84.29	75.00	66.43	31.43	80.71	72.14	65.00	38.57
<b>STMF-GMM-CMS (B)</b>	77.14	71.43	62.86	33.57	75.71	65.71	51.43	41.43

	<b>White</b>				<b>F16</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC-GMM-CMS</b>	36.43	18.57	10.00	2.86	58.57	28.57	7.86	2.86
<b>MFCC-UBM-CMS</b>	39.29	27.86	19.29	7.14	54.29	30.71	10.71	2.86
<b>ACC-GMM-CMS</b>	51.43	34.29	20.71	11.43	73.57	56.43	20.00	4.29
<b>ACC-UBM-CMS</b>	68.57	42.14	22.14	12.14	76.43	60.00	33.57	10.00
<b>STMF-GMM-CMS (A)</b>	67.86	51.43	37.86	12.14	82.14	68.57	55.00	10.71
<b>STMF-UBM-CMS (A)</b>	83.57	68.57	52.14	15.00	82.14	72.86	54.29	15.00
<b>STMF-UBM-CMS (B)</b>	76.43	66.43	60.00	28.57	78.57	73.57	63.57	29.29
<b>STMF-GMM-CMS (B)</b>	69.29	55.00	48.57	35.71	70.71	63.57	57.14	25.71

	<b>MFCC-GMM-CMS</b>	<b>MFCC-UBM-CMS</b>	<b>ACC-GMM-CMS</b>	<b>ACC-UBM-CMS</b>
<b>Clean</b>	100	100	98.57	98.57
	<b>STMF-GMM-CMS</b>	<b>STMF-UBM-CMS</b>	<b>STMF-GMM-CMS</b>	<b>STMF-UBM-CMS</b>
	(A)	(A)	(B)	(B)
<b>Clean</b>	98.57	98.57	95.71	92.86

Table 4 and Figure 4-8 show recognition rates of six approaches (three features with CMS; GMM or MAP-GMM) for the TIMIT corpus. Table 5 and Figure 4-9 show the comparison for the GRID corpus. Based on these results, the MAP-GMM clearly has the superior performance to the GMM for all features (MFCCs, ACCs and STMF). Compared with GMM, MFCCs, ACCs and STMF via the MAP-GMM approach

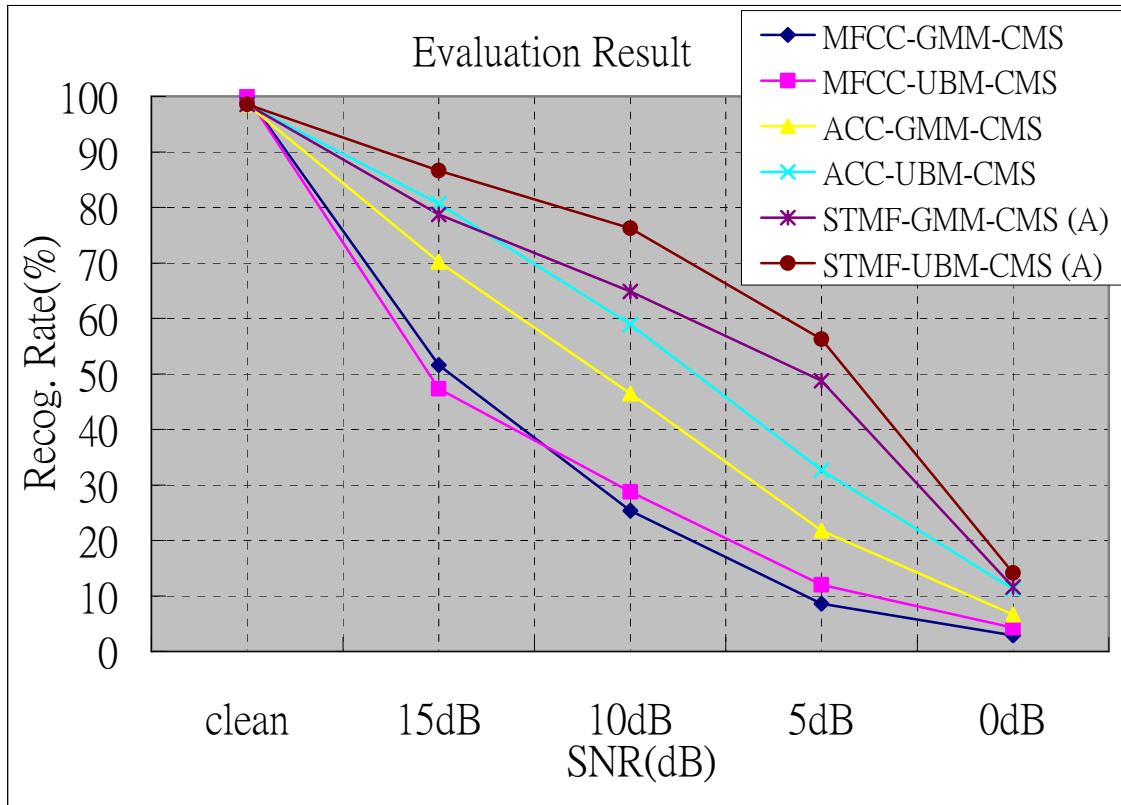
perform 9.9%, 12.46% and 13.92% better (average recognition rate in 5dB and 10dB SNR condition) in TIMIT corpus, and perform better 14.97%, 15.15% and 16.48% in GRID corpus. Therefore, the computational complexity, which is not addressed here, will be the only issue in choosing GMM versus MAP-GMM for a practical system.

**Table 5.** Correct recognition rates (in %) of GRID corpus.

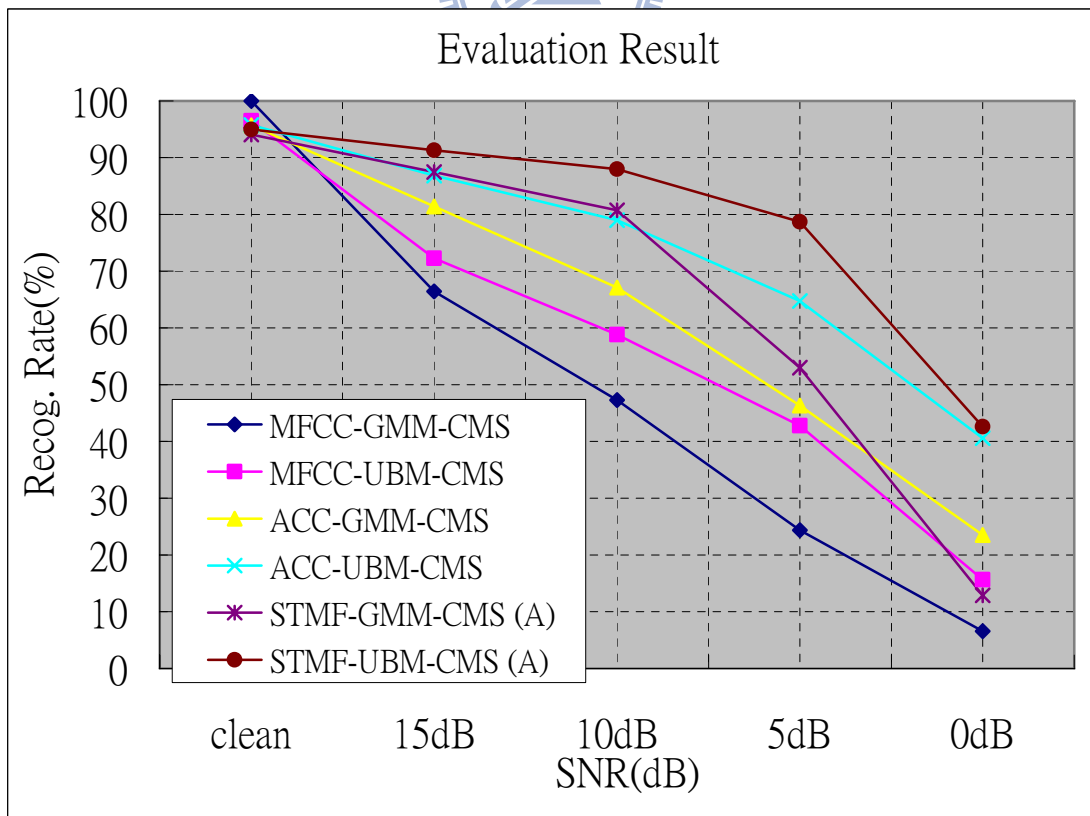
	<b>Factory1</b>				<b>Pink</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC-GMM-CMS</b>	73.28	51.72	24.46	5.69	69.12	47.99	22.55	5.83
<b>MFCC-UBM-CMS</b>	75.44	62.25	44.51	15.25	72.94	59.36	42.11	12.06
<b>ACC-GMM-CMS</b>	86.32	73.63	52.35	20.34	83.92	70.83	47.94	22.45
<b>ACC-UBM-CMS</b>	89.71	83.92	69.22	39.80	88.53	81.52	66.96	39.85
<b>STMF-GMM-CMS (A)</b>	89.71	84.46	55.78	12.30	88.73	83.24	57.79	11.52
<b>STMF-UBM-CMS (A)</b>	92.75	90.25	80.64	45.69	92.75	89.41	80.93	42.01
<b>STMF-UBM-CMS (B)</b>	81.62	80.00	74.71	46.86	80.15	78.73	75.15	51.32
<b>STMF-GMM-CMS (B)</b>	80.93	78.09	71.81	36.52	79.41	78.14	73.19	49.80

	<b>White</b>				<b>F16</b>			
	15dB	10dB	5dB	0dB	15dB	10dB	5dB	0dB
<b>MFCC-GMM-CMS</b>	52.16	38.58	27.50	8.92	71.18	50.83	22.99	5.78
<b>MFCC-UBM-CMS</b>	66.08	54.95	44.61	25.64	74.61	58.68	39.90	9.75
<b>ACC-GMM-CMS</b>	72.25	55.54	40.20	32.06	83.09	68.43	44.90	19.17
<b>ACC-UBM-CMS</b>	83.53	73.19	59.31	47.11	85.64	77.30	63.58	35.39
<b>STMF-GMM-CMS (A)</b>	83.28	72.65	45.83	19.71	88.14	82.65	52.55	8.04
<b>STMF-UBM-CMS (A)</b>	88.87	84.61	74.61	46.18	90.88	87.65	78.68	36.42
<b>STMF-UBM-CMS (B)</b>	78.09	76.81	71.91	45.54	80.83	79.41	72.01	36.86
<b>STMF-GMM-CMS (B)</b>	77.06	74.90	70.49	47.50	80.44	79.75	68.28	26.91

	<b>MFCC-GMM-CMS</b>	<b>MFCC-UBM-CMS</b>	<b>ACC-GMM-CMS</b>	<b>ACC-UBM-CMS</b>
<b>Clean</b>	100	96.52	95.69	95.78
	<b>STMF-GMM-CMS</b>	<b>STMF-UBM-CMS</b>	<b>STMF-GMM-CMS</b>	<b>STMF-UBM-CMS</b>
	(A)	(A)	(B)	(B)
<b>Clean</b>	94.07	94.95	86.37	84.41



**FIGURE 4-8** Average recognition rates (in %) of 70 people in TIMIT corpus.



**FIGURE 4-9** Average recognition rates (in %) of GRID corpus.

## 4.3 Discussions

From the experiment results in the previous section, it is clearly that ACCs are more robust than MFCCs. In addition, the STMF process could further enhance the performance of ACCs. Here are some reasons for these phenomenons.

The ACCs are derived from the auditory spectrum which represents speech energy along the log-frequency axis with the 24 cochlear filters per octave frequency resolution. This constant frequency resolution is high enough to characterize the 1/3~1/6 octave critical bandwidth measured in human hearing. In addition, the lateral inhibitory network in the first cochlear module sharpens the cochlear filters to have narrower bandwidth. On the other hand, the MFCCs use FFT to transform a time domain signal into the frequency domain. Such conventional approach has the trade-off between the time and frequency resolution. This constant-frequency-resolution versus time-frequency-resolution-trade-off might be the main reason for why hearing based features usually perform better than conventional FFT-based features.

The STMF process comprises several significant concepts. First, it works on joint spectro-temporal modulations not spectral or temporal modulations separately. Thus, we can extract more high-level features, such as the speaking rate (by temporal modulations) and FM sweeping directions (by joint spectro-temporal modulations). Secondly, the  $10^{-2}$  - 1 mask ( $10^{-2}$  for non-speech regions and 1 for speech regions) is intuitively similar to the conventional VAD approach. The major difference is that the VAD masks the non-speech regions on a frame-by-frame basis in the time domain while the STMF process masks the non-speech t-f units in the joint spectro-temporal domain.

As shown in Table 2 and Table 3, the parameter set STMF(B) performs better

within 0~10dB than the parameter set STMF(A), but worse in clean and 15dB conditions with the GMM based recognizer. It is not surprising since adopting the higher threshold would inevitably degrade the intelligibility of the clean speech such that the recognition rates decrease in high SNR conditions. On the other hand, the parameter set of STMF(A) produces better results than the parameter set of STMF(B) in the MAP-GMM based recognizer as shown in Table 4 and Table 5. The reason for that is the UBM is trained by the clean speech through the STMF process. The MAP-GMM is then to derive the speaker's model by adapting the well trained UBM. As mentioned above, the parameter set of STMF(B) performs worse in clean conditions, therefore, to construct a worse UBM. Thus, MAP-GMM speaker models adapted from this UBM would have worse performance due to the worse of speaker variability of this UBM. Consequently, the parameter set of the STMF(B) is suitable for GMM based recognizers while the parameter set of the STMF(A) is more favorable for MAP-GMM based recognizers.

# Chapter 5

## Conclusion and Future Works

In this thesis, we propose auditory spectral features (ACCs) further enhanced by the spectro-temporal modulation filtering (STMF) process for speaker recognition tasks in additive noises and demonstrate their superior performance of robustness to conventional MFCCs. Performance comparisons are also done between our proposed features and newly developed ANTCCs reported in [7]. For a randomly drawn 70-people testing set from TIMIT corpus, our STMF features are more robust than ANTCCs. For GRID corpus, ANTCCs and our STMF features achieve higher recognition rates in high SNR (15 and 10 dB) and low SNR (5 and 0 dB) conditions, respectively. We also demonstrate the MAP-GMM can further improve performance of proposed features provided the amount of training data is considered insufficient. However, the down side of the MAP-GMM is its computational complexity which makes the real-time implementation infeasible for now.

The STMF process produces joint spectro-temporal smoothed spectrogram, which highlights certain high-level information of the speaker. Such high-level information appears to be less variable than low-level information (for example, the static spectrum or MFCCs) in presences of noises. In the future, we will inspect the benefit of STMF process in presences of convolutional noises (such as channels or handsets mismatch).

## REFERENCE

- [1] D.A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech Comm.*, vol. 17, pp. 91–108, 1995.
- [2] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Comm.*, vol. 52, pp. 12–40, 2010.
- [3] D.A. Reynolds, et al., "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," in *Proc. ICASSP*, pp. 784-787, 2003.
- [4] R. Saeidi, J. Pohjalainen, T. Kinnunen, P. Alku, "Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification", in *IEEE Signal Processing Letters*, vol. 17, pp. 599-602, 2010.
- [5] J. Ming, et al., "Robust speaker recognition in noisy conditions," *IEEE trans. on Audio, Speech, and Language processing*, vol. 15, no. 5, pp. 1711–1723, July, 2007.
- [6] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE trans. on Audio, Speech, and Language processing*, vol. 29, pp. 254–272, 1981.
- [7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE trans. on Audio, Speech, and Language processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.
- [8] M. Rahim, Y. Bengio, and Y. Lecun, "Discriminative feature and model design for automatic speech recognition," in *Proc. Eurospeech'97*, Rhodes, Greece, 1997, pp. 75–78.
- [9] D. A. Reynolds, "Channel robust speaker verification via feature mapping," in *Proc. ICASSP, 2003*, vol. 2, pp. II-53–II-6.
- [10] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. A Speaker Odyssey—The Speaker Recognition Workshop*, Crete, Greece, 2001, pp. 213–218.
- [11] B. Xiang, U. Chaudhari, J. Navratil, G. Ramaswamy, and R. Gopinath, "Short-time Gaussianization for robust speaker verification," in *Proc. ICASSP'02*, Orlando, FL, 2002, pp. 681–684.
- [12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.
- [13] C. Barras and J. L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. ICASSP'03*, Hong Kong, China, 2003,



- pp. 49–52.
- [14] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, “Score normalization for text-independent speaker verification systems,” *Digital Signal Process.*, vol. 10, pp. 42–54, 2000.
- [15] R. Teunen, B. Shahshahani, and L. P. Heck, “A model based transformational approach to robust speaker identification,” in *Proc. ICSLP, 2000*, vol. 2, pp. 495–498.
- [16] C. H. Lee, C. H. Lin, and B. H. Juang, “A study on speaker adaptation of the parameters of continuous density hidden markov models,” *ASSP-39*, vol. 39, no. 4, pp. 806–814, April 1991.
- [17] Yiu, K.K., Mak, M.W., Kung, S.Y., “Environment adaptation for robust speaker verification by cascading maximum likelihood linear regression and reinforced learning,” *Computer Speech and Language*, vol. 21, pp. 231-246, 2007.
- [18] J. Ortega-Garcia and L. Gonzalez-Rodriguez, “Overview of speaker enhancement techniques for automatic speaker recognition,” in *Proc. ICSLP’96*, Philadelphia, PA, 1996, pp. 929–932.
- [19] Suhadi, S. Stan, T. Fingscheidt, and C. Beaugeant, “An evaluation of VTS and IMM for speaker verification in noise,” in *Proc. Eurospeech’03*, Geneva, Switzerland, 2003, pp. 1669–1672.
- [20] J. A. Nolasco-Flores and L. P. Garcia-Perera, “Enhancing acoustic models for robust speaker verification,” in *Proc. ICASSP*, pp. 4837–4840, Las Vegas, U.S.A., April 2008.
- [21] S. G. Pillay, A. Ariyaeinia, M. Pawlewski, and P. Sivakumaran, “Speaker verification under mismatched data conditions,” *IET Signal Processing*, vol. 4, no. 3:236–246, July 2009.
- [22] R.P. Lippmann, "Speech recognition by machines and humans," *Speech Comm.*, vol. 22, pp. 1–15, 1997.
- [23] W.H. Abdulla, "Robust speaker modeling using perceptually motivated feature", *Pattern Recognition letters*, pp.1333-1342, 2007.
- [24] Y. Shao, et al., "Incorporating auditory feature uncertainties in robust speaker identification," in *Proc. ICASSP*, vol. IV, pp. 277-280, 2007.
- [25] Q. Wu and L. Zhang, "Auditory Sparse Representation for Robust Speaker Recognition Based on Tensor Structure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2008, Article ID 578612, 2008.
- [26] T. Chi, P. Ru, and S.A. Shamma, "Multi-resolution spectro-temporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887-906, 2005.
- [27] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using

- adapted Gaussian mixture models,” *Digital Signal Process.*, vol.10, pp. 19–41, Jan. 2000.
- [28] Y. Linde, A. Buzo, and R. M. Gray, “An algorithm for vector quantizer design,” *IEEE trans. on communication*, com-28:84 95, 1980.
- [29] Y.-N. Hung, "Speech Enhancement Method based on Auditory Perceptual Model" in *Communication Engineering*. vol. master Hsin-Chu, Taiwan: National Chiao Tung University, 2008.
- [30] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Am.*, vol.87, no. 4, pp. 1738–1752, 1990.
- [31] M. Elhilali, T. Chi and S. A. Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech Comm.*, 41(2-3), pp.331–348, 2003.
- [32] T. Chi, et al., "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2719–2732, 1999.
- [33] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoust. Soc. Am.*, vol. 120, no. 5, pp. 2421–2424, 2006.
- [34] A. Varga and H.J.M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Comm.*, vol.12(3), pp. 247-251, 1993.
- [35] D.A. Reynolds, and R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions Speech Audio Processing*, vol. 3, pp. 72-83, 1995.