

# 國立交通大學

電信工程學系碩士班  
碩士論文

在感知訊號上使用子空間分析  
之語音增強技術

Subspace Decomposition of Perceptual Representations  
for Speech Enhancement

研究生：蕭任伯

Student: Hsiao, Jen-Po

指導教授：冀泰石 博士

Advisor: Dr. Chi, Tai-Shih

中華民國九十八年八月

在感知訊號上使用子空間分析  
之語音增強技術

Subspace Decomposition of Perceptual Representations  
for Speech Enhancement

研究生：蕭任伯

Student: Hsiao, Jen-Po

指導教授：冀泰石 博士

Advisor: Dr. Chi, Tai-Shih



A Thesis

Submitted to Department 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  
August 2009  
Hsin-Chu, Taiwan, Republic of China

中華民國九十八年八月

# 在感知訊號上使用子空間分析 之語音增強技術

學生：蕭任伯

指導教授：冀泰石 博士

國立交通大學電信工程學系碩士班  
感知訊號處理實驗室

## 中文摘要



在早期的語音訊號處理，是從時域或頻域兩種不同維度分開處理。近年來隨著聽覺模型的建立，我們確認了人類在聽覺上是同時在時、頻兩的維度上處理，基於這樣高維度的分析，人類比之現存的任何演算法擁有更高的健全性。

本論文中，使用了馬里蘭大學 NSL (Neural Systems Laboratory) 實驗室所開發出來的聽覺感知模型，模擬訊號透過耳朵往上傳遞到中腦聽神經的傳遞路徑，在其時-頻域分析階段先濾出語音最顯著的區域，接著使用子空間分析進一步壓抑殘存之雜訊。最後利用聽覺模型抽取出的語音特徵參數 (Auditory Spectrogram Coefficients) 在隱藏式馬可夫模型套件 (HTK) 上做連續數字的語音辨識，由辨識率的提升來印證此演算法的強健性。

# Subspace Decomposition of Perceptual Representations for Speech Enhancement

Student: Hsiao, Jen-Po

Advisor: Dr. Chi, Tai-Shih

Department of Communication Engineering

National Chiao-Tung University

Perception Signal Processing Laboratory



## English Abstract

In early years, conventional speech enhancement techniques have been developed separately in time domain and in frequency domain. Recent years, with the auditory model being introduced, enhancement techniques are developed in joint spectro-temporal domains to incorporate hearing perception perspectives to enhance their robustness.

In this thesis, we use the auditory model, which simulates the hearing physiology from cochlea to cortex, introduced by NSL (Neural Systems Laboratory), Maryland university. At first, the spectrograms are selected within speech regions in cortical domain. Second, we adopt the subspace algorithm to filter the noise that exists in speech regions. Finally, the Auditory Cepstrum Coefficients (ACC) is extracted for HTK recognition task. From HTK evaluations, the robustness of the proposed algorithm is proven.

# 誌 謝

能完成這篇論文，要歸功於很多人的幫助。

要感謝實驗室的學長，勇樣、尚儒、廷宇，在我還是菜鳥時給我很大的指導；實驗室的同學及學弟妹，柏宏、大師、阿郎、大樹、勝哥、叮咚、蘭雲及基奴禮偉，平時嬉嬉鬧鬧又能認真做研究，這兩年真是相當棒的回憶；IT跟語音實驗室的同學小玄子、谷嶸、no 哥、pulu、小宋、小帥哥以及大學同學 renve、仔仔、whyme，在課餘能一起討論功課還有出去玩，大家都是很棒的一群人。

再來要感謝我的家人，讓我能後顧之憂的就讀研究所。想對爸媽說：感謝你們的體諒與包容，讓我這不成材的兒子順利完成學業，這也是第一次寫給你們的感謝詞，致上我十二萬分的感謝。另外要謝謝女友小雪的陪伴，讓我阿宅的生活能走出戶外，給予我更充足的陽光、水果與點心。

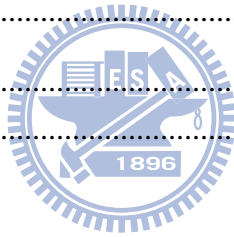
最後是我的指導老師，冀泰石教授。兩年來帶領我完成碩士班的研究，從你的身上學到很多人生的哲理，說是亦師亦友一點也不為過。這篇論文，應該也只有這頁不會被你訂正吧。(笑)

蕭任伯謹誌 交通大學

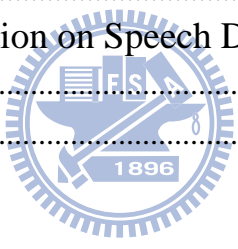
2009 年夏天

# 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.....	2
<b>Chapter 2 Literature Review</b> .....	3
2.1 Hearing Physiology .....	3
2.1.1 Hearing Physiology.....	4
2.1.2 Spectrum Estimation of Auditory Perceptual Model .....	7
2.1.3 Cortical Analysis.....	10
2.2 Basic Subspace Algorithms in Speech Enhancement .....	13
2.2.1 Time-Domain Constrains .....	14
2.2.2 Pre-whitening for Colored Noise .....	16
2.3 Supervector : 2D image processing.....	17
<b>Chapter 3 Subspace Decomposition of Perceptual Representations for Speech Enhancement</b> .....	19



3.1 Introduction .....	19
3.2 The 2D Neural Patterns in the Cortex.....	23
3.2.1 Dimension Redundancy Problem .....	24
3.2.2 Frequency Band Division .....	25
3.2.3 Window Length for Noise Estimation .....	27
3.3 The weighted mask for HTK evaluation .....	27
3.4 Summary .....	30
<b>Chapter 4 Evaluation .....</b>	<b>33</b>
4.1 Database and Evaluation Measurements Introduction.....	33
4.1.1 AURORA 2.0 .....	34
4.1.2 Advance Front-end feature Extraction.....	34
4.1.3 HTK Setting.....	35
4.1.4 Speech Distortion and Residual Noise.....	37
4.2 HTK Results .....	39
4.3 Performance Evaluation on Speech Distortion and Residual Noise .....	42
4.4 Summary .....	45
<b>Chapter 5 Conclusion and Future Works.....</b>	<b>46</b>
<b>Appendix I Pre-whitening Verification .....</b>	<b>48</b>
<b>Appendix II The AFE and Yung's Result .....</b>	<b>49</b>
<b>Appendix III The Proposed algorithm HTK Recognition</b>	
<b>Result.....</b>	<b>51</b>
<b>REFERENCE.....</b>	<b>52</b>



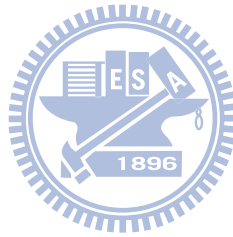
## List of Figures

FIGURE 2-1 The anatomy of the ear. ....	4
FIGURE 2-2 The basilar membrane diagram and the characteristic frequency at the basilar membrane. ....	5
FIGURE 2-3 Firing rate of auditory neuron. ....	6
FIGURE 2-4 The diagram of first stage of auditory model.....	7
FIGURE 2-5 Filterbank in NSLtool box .....	8
FIGURE 2-6 An example of wav2aud “come home right away”.....	10
FIGURE 2-7 An example of moving ripple stimulus. ....	11
FIGURE 2-8 Responses for 8 basic neurons in the cortex. ....	12
FIGURE 2-9 Rate-scale plot squeezed spectra and temporal.....	13
FIGURE 2-10 The realignment diagram. ....	18
FIGURE 3-1 The clean speech and the noisy speech spectrogram .....	20
FIGURE 3-2 Flowchart of the proposed algorithm. ....	22
FIGURE 3-3 The STCRs of clean speech. (rv=1,2,4, sv=1,2,4) .....	23
FIGURE 3-4 The STCR downsampling diagram.....	25
FIGURE 3-5 The statistic of speech in auditory spectrogram. ....	26
FIGURE 3-6 The silence (noise) frame of clean, enhanced and noisy speech.....	28
FIGURE 3-7 The weighting curve.....	29
FIGURE 3-8 Binary mask and smoothed mask.....	30
FIGURE 3-9 Weighted mask through the enhanced auditory spectrogram.....	32
FIGURE 4-1 AFE block scheme.....	35
FIGURE 4-2 Recognition rate of ACC Baseline and the Yung’s result.....	37
FIGURE 4-3 Speech frames and noise frames. ....	38
FIGURE 4-4 Simulation for different $\mu$ and threshold. ....	39
FIGURE 4-5 Recognition result of AFE, Yung’s and the proposed algorithm.....	40
FIGURE 4-6 Average recognition rate of AFE, Yung’s and the proposed algorithm. .	42
FIGURE 4-7 Speech distortion for HTK evaluation. ....	43
FIGURE 4-8 Residual noise for HTK evaluation.....	43
FIGURE 4-9 Speech distortion for subspace algorithm. ....	44
FIGURE 4-10 Residual noise for subspace algorithm.....	44



## List of Tables

Table 3-1 The downsample multiply for each scale. ....	24
Table 3-2 The downsample multiply for each rate. ....	24
Table 3-3 The estimated noise region corresponding to each rate. ....	27
Table 4-1 Hit / insertion rate of AFE, Yung's and the proposed algorithm.....	41



# Chapter 1 Introduction

## 1.1 Introduction

In recent years, lots of speech applications, for instance, the cell phone, PDA, hearing aid device, etc., have been developed to provide convenience for our life activity. To conquer the noisy environment around us, the functions of devices are designed to be robust as human beings. Speech enhancement is one of the techniques that against noisy environment. Those techniques are often utilized to improve the speech recognition rate or the speech quality; depending on what applications on hand.

During the past decades, conventional speech enhancement techniques have been developed both in time domain and in frequency domain, such as spectral subtraction [1, 2], Wiener filter [3], statistical-model-based method [4] and subspace method [5]. Later on, with the auditory model being introduced [6-9], enhancement techniques are developed in separate or joint spectro-temporal domains to incorporate hearing perception perspectives to enhance their robustness. Here, we adopt the subspace method onto the joint spectro-temporal domain, an internal domain of our auditory model which is specifically called cortical domain [10].

In this thesis, we use (1) HTK to evaluate the speech recognition and (2) the speech distortion measure and (3) the noise residual error to examine the effectiveness and robustness of our proposed subspace algorithm. We review works done by other researchers in chapter 2. Our proposed method would be presented in Chapter 3. The evaluation results will be shown in Chapter 4. Brief the discussions and the future works will be given in chapter 5.

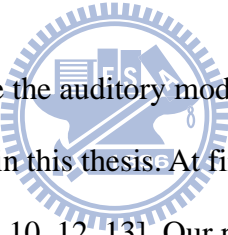
## 1.2 Motivation

Human-machine interface will be the killer application of next generation. Indeed, there are many people that not able to write but to speak. Also, many people would like to listen clearly instead of reading comprehensibly; apparently to the elder. Therefore, speech enhancement is more and more important to our society with the increasing elder population.

Auditory models have been evolved from one-dimensional into multi-dimensional models. Therefore, auditory model based speech enhancement techniques should be built on the multi-dimensional auditory representation. The preliminary work done by Yung showed some significant achievements in speech recognition rate [11], hence we propose a subspace decomposition coupled with Yung's method to further explore the robustness of the multi-dimensional speech enhancement technique.

# Chapter 2

## Literature Review



In this chapter, we briefly describe the auditory model and the subspace decomposition algorithm utilized in this thesis. At first, the auditory model developed by Shamma et al. is introduced [9, 10, 12, 13]. Our proposed approach works on the representations from this auditory model. In section 2.2, we shortly review basic subspace algorithms for speech enhancement [5, 14, 15]. Finally, the supervector technique, which is used to express higher dimensional representations in our subspace decomposition, will be described concisely [16, 17].

### 2.1 Hearing Physiology

During past decades, the idea of adopting properties of human hearing in speech-related applications becomes more and more popular within the group of speech researchers. Here, we adopt a similar idea to study the speech enhancement in

an internal perceptual representation of an auditory model. Basic hearing physiology and the auditory model, which is proposed by Shamma et al, are introduced step by step in this section

## 2.1.1 Hearing Physiology

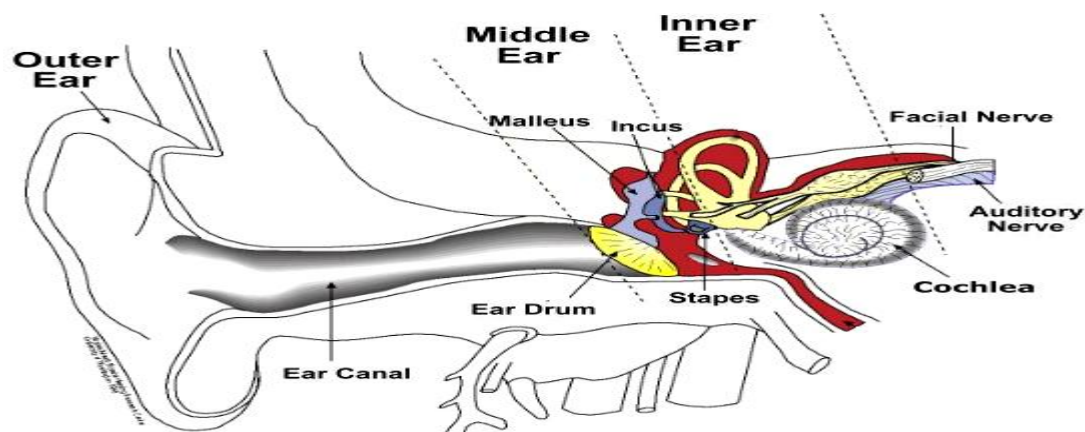


FIGURE 2-1 The anatomy of the ear.  
([http://www.advcoch.com/I2\\_Hearing\\_Physiology.htm](http://www.advcoch.com/I2_Hearing_Physiology.htm))

The ear could be divided into three parts – outer ear, middle ear and inner ear, and the anatomy of the ear is shown in figure 2-1.

The most important functions of the out ear are localization, amplification and protection. Because of the paired ears, we could use the phase delay and amplitude difference to judge the direction of sound source. Also, the ear canal is regarded as a filter that gives the largest gain at about 3,300 Hz.

The middle ear is the portion of the ear internal to the eardrum, external to the oval window of the cochlea. When the sound arrives at the eardrum, it is transferred from wave to vibration. By passing through the three ossicles, known as malleus, incus and stapes, the sound signal is conveyed to the oval window, the start of inner ear.

The cochlea in the inner ear plays a significant role in the auditory system. It consists of three chambers with full lymph, as shown in figure 2-2. By the time the mechanical vibration arrives the oval window, a traveling wave is generated and propagates along the basilar membrane (BM) of the cochlea. Different locations of the BM reach maximum responses in pertain to traveling waves with different frequencies. The basilar membrane is about 35mm in length with its width increasing and elasticity decreasing progressively from base to apex. The left panel of figure 2-2 shows the diagram of basilar membrane and the right panel shows the maximum responsive frequencies along the basilar membrane. The range of resonance frequency is about 20-20,000 Hz, which is the audible frequency range of human being.

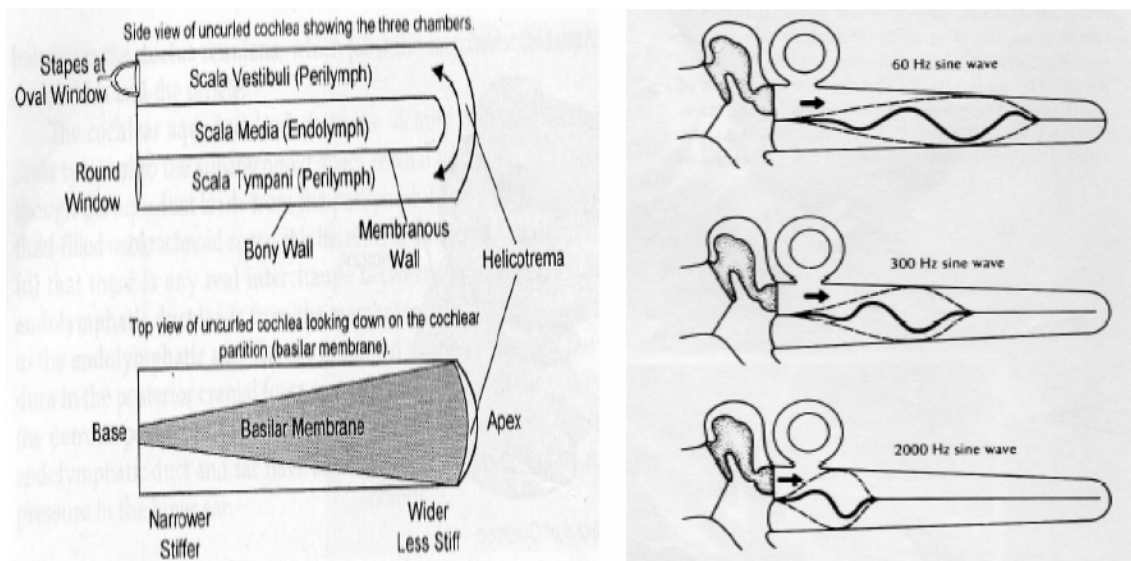


FIGURE 2-2 The basilar membrane diagram (left) and the characteristic frequency at the basilar membrane (right). (Hearing Physiology Handout, AAIP)

For a complex sound consisting of several frequencies, the overall pattern of the BM would be determined by resonances of all input frequency components. The mechanical inhibitions between neighboring frequencies on the BM might be the

main reason of the well-known “frequency masking” phenomenon of human audition.

The traveling wave generates displacement of the BM, then the hair cells distributed along the basilar membrane transform the displacement pattern to corresponding pattern of sensory nerve action potentials. There are two different hair cells: inner hair cells (IHCs) and outer hair cells (OHCs). Most of the transformation from mechanical vibrations to electrical potentials is done by the help of IHCs, a kind of sensor connects with the auditory nerve. On the other hand, OHCs are often for the amplification/reduction of action potentials through the auditory nerve to protect the auditory sensory system. Due to the fact that a relaxation time is needed between consecutive fires of auditory neurons, firing rates can not keep up with high frequency components, as demonstrated in Figure 2-3. Firing rates of IHCs are bounded by 4-5k Hz and rates of the midbrain are bounded by about 1k Hz.

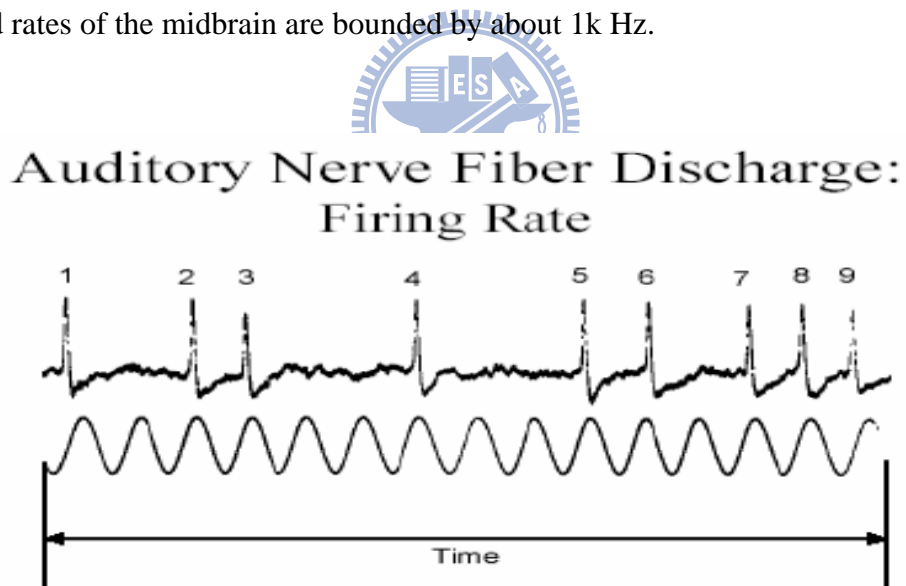


FIGURE 2-3 The firing rate of auditory nerve correspond to the monotone audio input. (Hearing Physiology Handout, AAIP)

## 2.1.2 Spectrum Estimation of Auditory Perceptual Model

The first stage of the auditory perceptual model is to simulate the sound pathway from the cochlea, hair cells and auditory nerves to the midbrain. It is divided into three substages – analysis stage, transduction stage and reduction stage, as shown in figure 2-4.

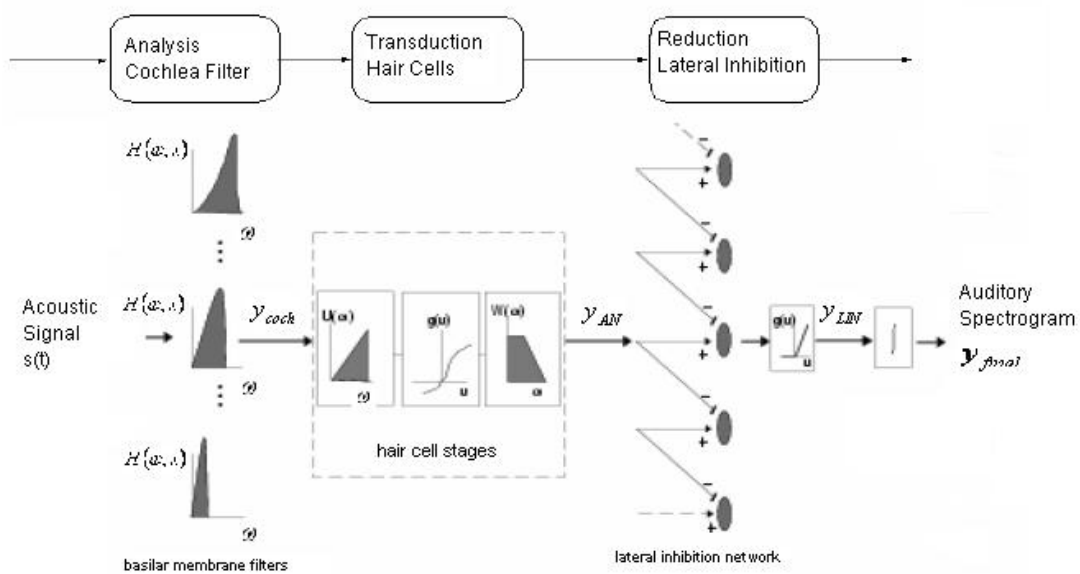


FIGURE 2-4 The diagram of first stage of auditory model. (Auditory Model Handout, AAIP)

The cochlea is often thought as a frequency analyzer, hence modeled by a bank of 128 constant-Q bandpass filters in the analysis stage. Figure 2-5 shows a filterbank consisting of 128 IIR filters uniformly distributed among 5.3 octaves with 24 filters/octave frequency resolution. The bandwidth and the center frequency of each filter satisfy the following equation:

$$f_{center}/bandwidth = Q \quad (2-1)$$

where  $Q$  is a constant ( $= 4$ ) in our implementation. It is obviously that with the



center frequency increasing, the corresponding bandwidth is increasing gradually. This property describes the general idea that the cochlea possesses higher frequency resolution (i.e., narrower bandwidth) at low frequency regions than at high frequency regions.

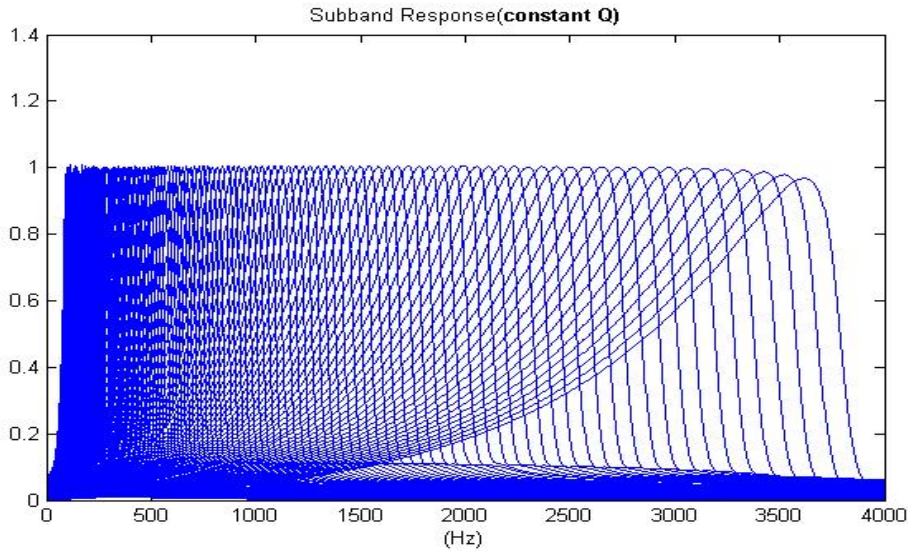


FIGURE 2-5 The filterbank consists of 129 filters which conforms to  $f_{center}/bandwidth = Q$ .

In the analysis stage, outputs of the cochlear filterbank can be represented by the following equation:

$$y_{coch}(t, x) = s(t) \otimes h(t, x) \quad (2-2)$$

where  $x$  encodes the location of a particular cochlear filter along the BM (i.e., the log-frequency axis from engineering point of view) and  $h(t, x)$  are impulse responses of the filterbank.

The transduction stage then models the behaviors of inner hair cells including (1) the transduction of the traveling pressure to the velocity in the lymph; (2) the neural saturation and (3) current leakages. This stage can be formulated as

$$y_{AN}(t, x) = g(\partial_t y_{coch}(t, x)) \otimes \omega(t) \quad (2-3)$$

where  $\partial_t$  models the transduction of the hydraulic pressure to velocity; the sigmoid function  $g$  is used to simulate the neural saturation as follows:

$$g(u) = 1/(1 + e^{-u}) \quad (2-4)$$

and the low-pass function  $\omega(t)$  is used to account for current leakages of auditory neurons.

The last reduction stage addresses two important observations in the auditory sensory system: (1) the lateral inhibition of auditory neurons, which might account for the frequency masking phenomenon shown in human hearing; and (2) the observed temporal dynamics reduction from the cochlea to the midbrain. The following two equations are formulated in the auditory model we used.

$$y_{LIN}(t, x) = \max(\partial_x y_{AN}(t, x), 0) \quad (2-5)$$

$$y_{final} = y_{LIN}(t, x) \otimes \mu(t; \tau) \quad (2-6)$$

where the first-order derivative  $\partial_x y_{AN}(t, x)$  simply approximates the lateral inhibition between neighboring neurons, the half-wave-rectifier puts the constraint on the negative potential, and the low-pass filter  $\mu(t; \tau) = e^{-t/\tau} \cdot u(t)$  with a time constant  $\tau$  models the temporal dynamics reduction of the midbrain.

The output of these three stages is a two-dimensional representation in the spectral (log-frequency) and temporal domain and is referred to as the auditory spectrogram [12]. Yung's study showed features extracted from auditory spectrograms are more robust in speech recognition tasks [11]. One example of the auditory spectrogram is shown in figure 2-6.

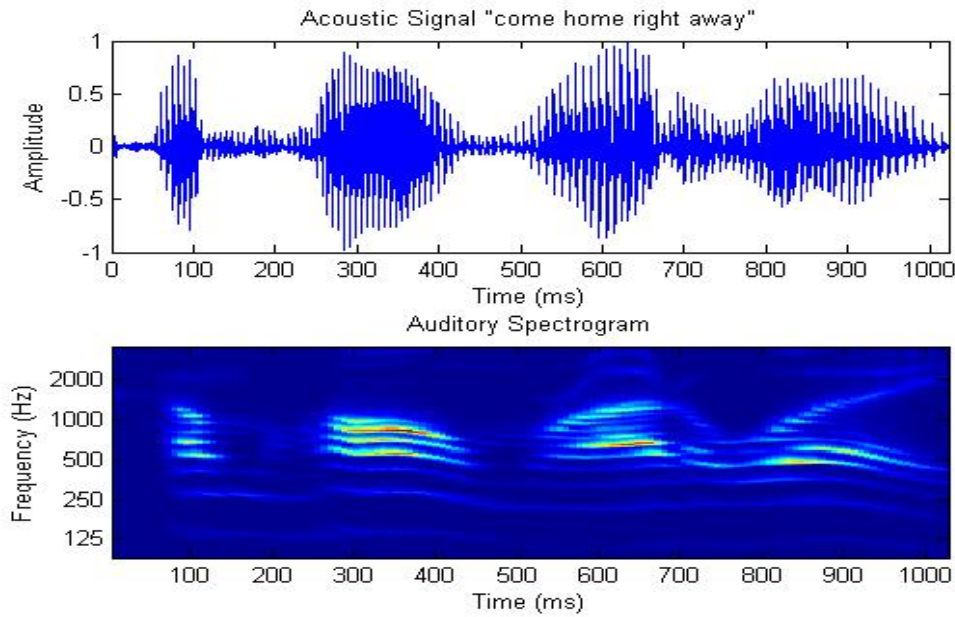
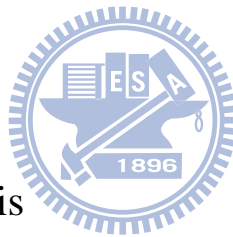


FIGURE 2-6 An example of wav2aud using sentence “come home right away”.



### 2.1.3 Cortical Analysis

The processing of generating the auditory spectrogram, an estimate of the spectrum by the inner ear, is introduced in the previous section. Furthermore, neurophysiological evidences reveal that neurons in the higher auditory cortex (AI) respond to different frequencies as well as to temporal structures of patterns generated by inner ears. In other words, AI’s neurons exhibit different spectro-temporal tunings and can be characterized by Spectro-Temporal Receptive Fields (STRFs), which can be considered as spectro-temporal two-dimensional impulse responses from engineering perspectives. To measure the 2D impulse responses of neurons in AI, one has to use orthogonal basis signals in the spectro-temporal domain to drive the cortex. Such spectro-temporal basis signals are so called moving ripple stimuli. Figure 2-7 shows one example of the moving ripple stimulus of rate=+4 (Hz, the temporal

velocity in time) and scale=0.5 (cycle/octave, the density in log-frequency). In addition to the rate and scale parameters, directional selectivity of the FM sweep is encoded by the sign of the rate parameter, in which positive sign of rate represents the downward direction, i.e., frequency decreasing with time, and negative sign represents the upward direction.

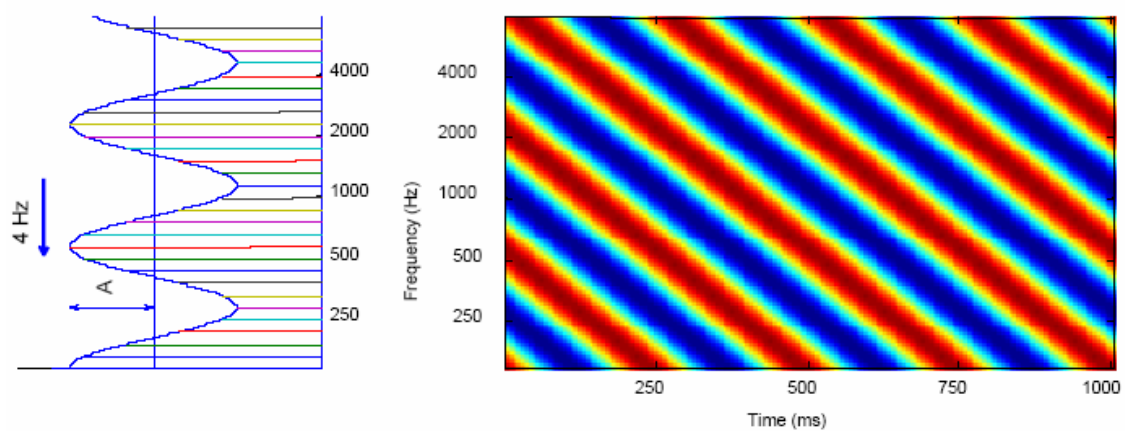


FIGURE 2-7 An example of moving ripple stimulus.

( Auditory Model Handout, AAIP)

By measuring impulse responses of many neurons, researchers conclude different AI's neurons roughly tune to combinations of different rate, scale and direction. Therefore, the auditory cortex can be modeled as a bank of 2D bandpass filters to analyze the input 2D auditory spectrogram. The schematic plot in figure 2-8 demonstrates the 2D cortical filtering of AI on a sample spectrogram. The small top panels on each subplot are the impulse responses of different typical neurons tuning to slow/fast rates and coarse/fine scales. The bottom panels are outcomes of these 2D spectro-temporal filters.

Overall outputs of the 2D filtering construct a four-dimensional representation (in rate, scale, log-frequency and time), which is hard for illustration. Therefore, we integrate the 4D output along both spectral and temporal axes to generate an energy pattern on the remaining rate-scale axes. Figure 2-9 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=0.5-4 cyc/oct, while the white noise has modulations distributed to high rates and all possible scales.

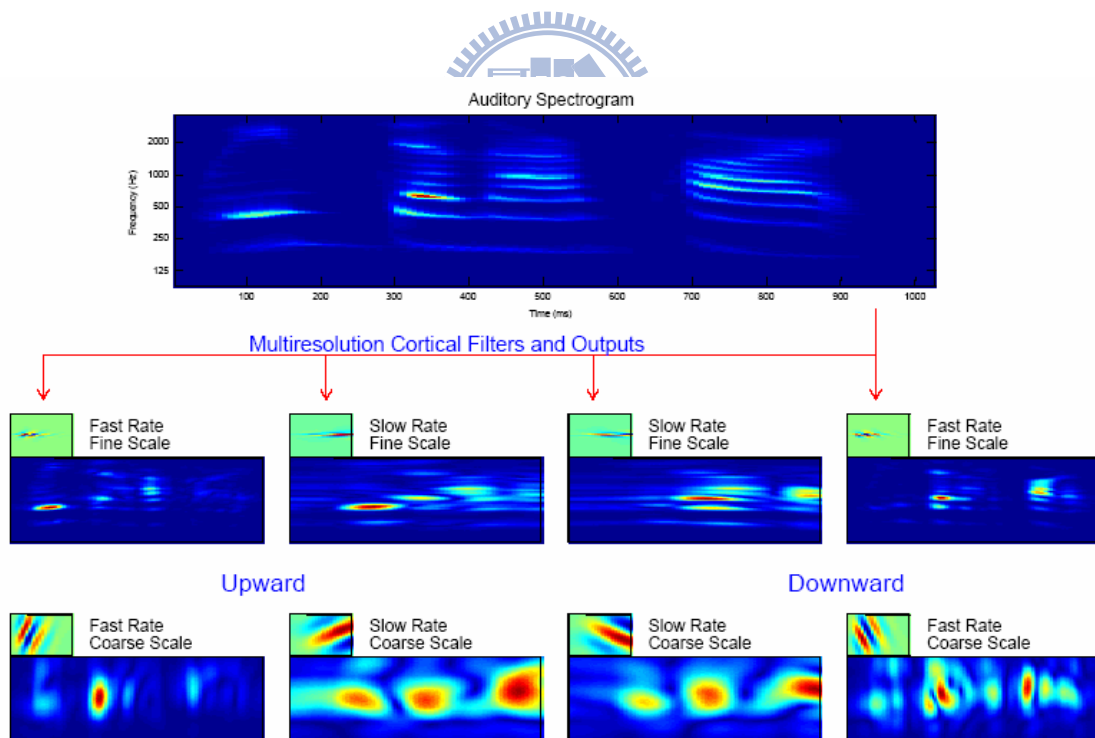


FIGURE 2-8 The response for 8 basic nerves in the cortex. (Auditory Model Handout, AAIP)

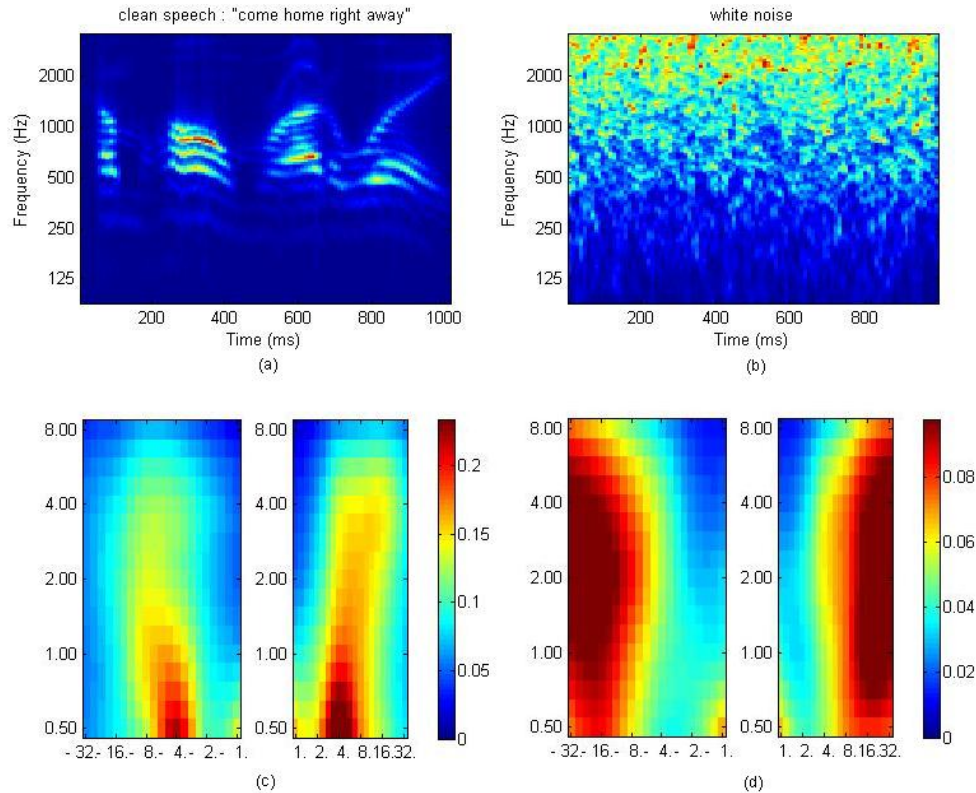


FIGURE 2-9 (a) clean speech. (b) white noise. (c) clean speech in rate-scale domain with rate and scale in x- and y- axis. (d) white noise in rate-scale domain.

## 2.2 Basic Subspace Algorithms in Speech Enhancement

There are many speech enhancement algorithms, such as spectral subtraction [1, 2], Wiener filtering [3] and statistical-model-based method [4]. In this study, a subspace decomposition algorithm based on linear algebra theory is utilized and introduced in this section. Subspace algorithms suppress noise by including signal components falling in “speech” space while excluding components in the “noise” space. In this section, we first introduce the time-domain linear optimal estimator

which minimizes the speech distortion from white noise under certain constraints.

Next, the colored noise, which is similar to the real noise around us, will be considered in our algorithm.

## 2.2.1 Time-Domain Constrains

Consider the noisy speech signal  $y = x + d$  containing samples of clean speech  $x$  and noise  $d$ . The cross-correlation matrix of  $y$  (of length  $K$ ) is defined as:

$$\begin{aligned} R_y &\equiv E[y \cdot y^T] \\ &= E[x \cdot x^T] + E[d \cdot d^T] + E[x \cdot d^T] + E[d \cdot x^T] \end{aligned} \quad (2-7)$$

The cross-correlation matrix  $R_y$  ( $K \times K$ ) is a symmetric and positive semi-definite, assuming  $x$  and  $d$  are wide-sense stationary signals. We postulate that the signal and the noise vectors are uncorrelated and zero mean, then the preceding equation can be reduced to:

$$R_y = R_x + R_d \quad (2-8)$$

where  $R_x \equiv E[x \cdot x^T]$  and  $R_d \equiv E[d \cdot d^T]$  are the auto-correlation matrices of the signal and noise, respectively. If we further assume that the noise is white, the noise correlation matrix will be diagonal and the equation (2-8) can be rewritten as:

$$R_y = R_x + \sigma_d^2 \cdot I \quad (2-9)$$

where  $\sigma_d^2$  is the noise variance.

Now let  $\hat{x} = H \cdot y$  be a linear estimator of the clean speech  $x$ , where  $H$  is a  $K \times K$  matrix. The residual error  $\varepsilon$  of this estimator is then given by:

$$\begin{aligned}
\varepsilon &= \hat{x} - x \\
&= (H - I) \cdot x + H \cdot d \\
&= \varepsilon_x + \varepsilon_d
\end{aligned} \tag{2-10}$$

where  $\varepsilon_x$  represents the speech distortion, and  $\varepsilon_d$  represents the residual noise.

Next we define the energy of  $\varepsilon_x$  and  $\varepsilon_d$  as:

$$\begin{aligned}
\bar{\varepsilon}_x^2 &= E[\varepsilon_x^T \cdot \varepsilon_x] \\
\bar{\varepsilon}_d^2 &= E[\varepsilon_d^T \cdot \varepsilon_d]
\end{aligned} \tag{2-11}$$

Thus we can obtain the optimum linear estimator by solving the following time-domain constrained problem:

$$\begin{aligned}
\min_H \bar{\varepsilon}_x^2 \\
\text{subject to: } \frac{1}{K} \bar{\varepsilon}_d^2 \leq \zeta^2
\end{aligned} \tag{2-12}$$

where  $\zeta$  is a positive constant. This constrained optimization problem can be solved as in [18]:

$$H_{opt} = R_x (R_x + \mu \cdot R_d)^{-1} \tag{2-13}$$

where  $\mu$  is the Lagrange multiplier. The formula of this optimal estimator

$H_{opt}$  is similar to the formula of the Wiener filter when  $\mu = 1$ . The major

difference is that  $H_{opt}$  works on the time domain, on the other hand, the Wiener

filter performs on the frequency domain. In addition, the constant  $\mu$  gives us

lots of degrees of freedom in designing our estimator.

Furthermore, equation (2-13) can be simplified by using eigen-decomposition of

$R_x = U \Lambda_x U^T$  yielding:

$$H_{opt} = U \Lambda_{opt} U^T \tag{2-14}$$



where  $\Lambda_{opt}$  is a  $K \times K$  diagonal matrix given by:

$$\Lambda_{opt} = \Lambda_x (\Lambda_x + \mu \cdot \sigma_d^2 \cdot I)^{-1} \quad (2-15)$$

## 2.2.2 Pre-whitening for Colored Noise

Only white noise with diagonal correlation matrix is considered in the previous section. However, in practical world, background noises are seldom white, but colored instead. A simple way to deal with colored noises is to transform them to white noises by a pre-whitening process which is introduced in this section.

The correlation matrix  $R_d$  of noise, which can be extracted from the speech absent segments, is factorized by the Cholesky factorization:

$$R_d = R^T R = L \cdot L^T \quad (2-16)$$

where  $L$  is a unique lower triangular  $K \times K$  matrix. Multiplying the pre-whitening matrix  $L^{-1}$  to the equation (2-8) yields:

$$\begin{aligned} L^{-1}y &= L^{-1}x + L^{-1}d \\ y' &= x' + d' \end{aligned} \quad (2-17)$$

where  $d'$  becomes white after the pre-whitening procedure. (See Appendix I for the proof.) Therefore, the correlation matrix  $R_{y'}$  of the noisy speech can be rewritten as:

$$\begin{aligned} R_{y'} &= L^{-1}R_x L^{-T} + I \\ &= R_{x'} + I \end{aligned} \quad (2-18)$$

After deriving the linear estimator of  $x'$  as mentioned in the previous section, we should multiply  $L$  to the estimator  $\hat{x}'$  to have the post-whitening estimator  $\hat{x}$ . These procedures can be formulated as:

$$\hat{x} = L \cdot H' \cdot L^{-1} y \quad (2-19)$$

where  $H'$ , the optimal estimator solution for pre-whitening elements as in equation (2-18), has the same form as the  $H_{opt}$  in equation (2-13).

The noise correlation matrix is not diagonal since  $U$ , the eigenvector matrix of  $R_x$ , diagonalizes  $R_x$  not  $R_d$ . It is shown [19] that there exists a matrix  $V$  which can diagonalize  $R_x$  and  $R_d$  simultaneously in the following way:

$$\begin{aligned} V^T R_x V &= \Delta_x \\ V^T R_d V &= I \end{aligned} \quad (2-20)$$

where  $\Delta_x$  and  $V$  are the eigenvalues matrix and eigenvector matrix respectively of  $\Sigma = R_d^{-1} R_x$ . Note that the eigenvector matrix  $V$  is not orthogonal. Hence, we can rewrite the optimal linear estimator from equation (2-15) as:

$$H_{opt} = V^{-1} \Delta_x (\Delta_x + \mu \cdot I)^{-1} V^T \quad (2-21)$$

## 2.3 Supervector : 2D image processing

Many perceptual properties in hearing and in vision share similar sensory mechanisms [20]. For example, the principles to group sounds from a spectrogram are the same principles to group objects from an image. Therefore, in this study, we treat the speech enhancement in spectrograms as a 2D image enhancement problem. The most common technique in 2D image enhancement is using the supervector technique to transform the 2D task into a 1D task, as shown in some eigenface studies [16, 17].

In image processing applications, the pattern of  $N$  by  $N$  elements is usually rearranged to a vector of  $1$  by  $N^2$ . This implies that characteristics of a  $N \times N$  matrix are equal to those of a  $1 \times N^2$  vector, as shown in figure 2-10.

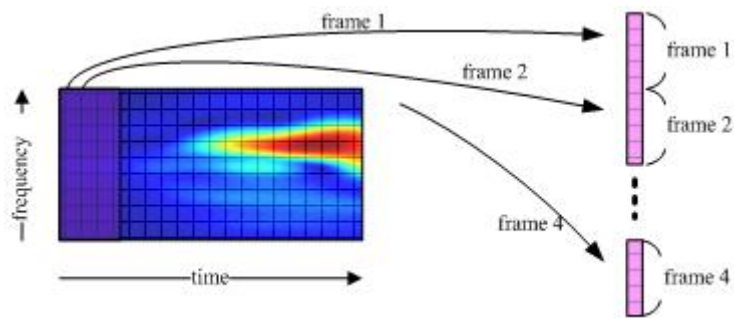
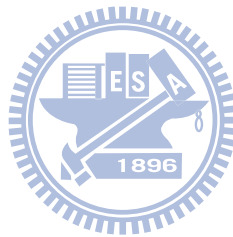
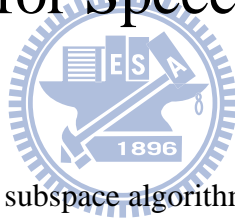


FIGURE 2-10 The realignment diagram showing the transition of 2D to 1D.



## Chapter 3

# Subspace Decomposition of Perceptual Representations for Speech Enhancement



The auditory model and the basic subspace algorithm were described in Chapter 2.

The subspace decomposition of perceptual representations will be fully expressed in this chapter.

### 3.1 Introduction

Most speech processing algorithms are developed in either temporal domain (channel by channel) or in spectral domain (frame by frame). However, from neuro-physiological evidence, human brain analyzes speech in a joint spectro-temporal fashion of considering temporal dynamics with spectral contents at the same time. Our approach of taking the joint spectro-temporal domain into consideration is inspired by such scientific findings. For example, one could easily

understand speech in noisy environments merely because of significant differences shown in spectro-temporal structures between speech and noise, as in figure 3-1. Following this concept, we propose the subspace decomposition algorithm in the joint spectro-temporal domain to extract speech-related features.

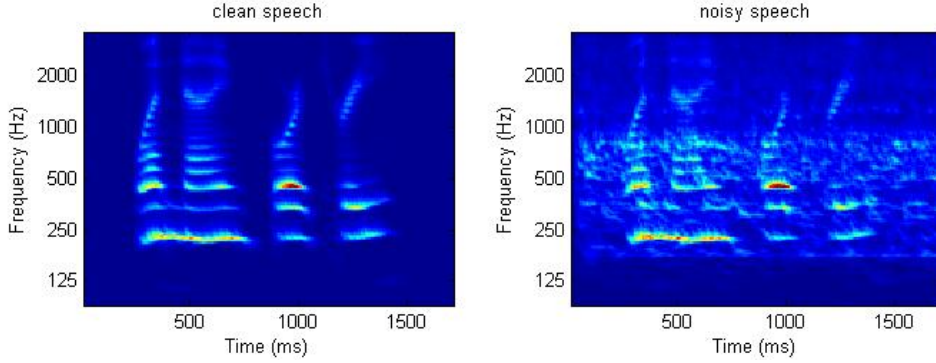
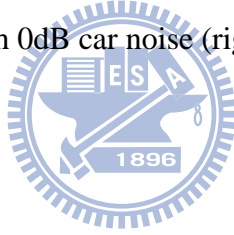


FIGURE 3-1 The auditory spectrogram of the clean speech (left) and the speech with 0dB car noise (right).



The spectro-temporal auditory representation used in this study was proposed in [9]. As pointed out in [9], the four-dimensional cortical impulse response is given by:

$$STRF(x, t; \Omega, \omega) = RF(x; \Omega, \phi) \cdot h_{IR}(t; \omega, \theta) \quad (3-1)$$

where  $RF(x)$  is the response field along the log-frequency (tonotopic) axis,  $h_{IR}(t)$  is the temporal impulse response. It has been shown that most of the modulations of speech signals fall in the range of rate = 2~16 Hz, scale = 0.5~8 cyc/oct [11]. Thus, we would use modulations within those ranges to extract spectro-temporal structures of speech in our enhancement application as:

$$STRF_{speech} = \begin{cases} STRF(x, t; \Omega, \omega), & 2 \leq \Omega \leq 16, 0.5 \leq \omega \leq 8 \\ 0, & otherwise \end{cases} \quad (3-2)$$

The Spectro-Temporal Cortical Response  $STCR_{\Omega, \omega}(x, t)$  within speech regions

can then be written as:

$$STCR_{\Omega,\omega}(x,t) = y(x,t) \otimes STRF(x,t;\Omega,\omega) \Big|_{\substack{\Omega=\pm 2,\pm 4,\pm 8,\pm 16 \\ \omega=0.5,1,2,4,8}} \quad (3-3)$$

where  $y(x,t)$  is an input spectrogram and  $\otimes$  is the 2D convolution. For every input spectrogram  $y(x,t)$ , we obtain 40 STs given the  $\omega(rate) = \pm 2, \pm 4, \pm 8, \pm 16$  Hz and  $\Omega(scale) = 0.5, 1, 2, 4, 8$  cycle/octave. Next, we adopt the subspace decomposition via the supervector technique to each STCR separately.

As shown in figure 2-10, we transfer each 2D STCR to a 1D vector, i.e.,

a matrix  $(M \times N) \Rightarrow$  a vector  $(M \cdot N \times 1)$ , by:

$$st_{\Omega,\omega}(i) = \Phi[ST_{\Omega,\omega}(x,t)] \quad (3-4)$$

$\Phi$  is the transition function of 2D to 1D.

Transferring a 2D matrix to a 1D vector is a conventional way to allow us applying the subspace decomposition to the perceptual representation STCR.

In the proposed subspace decomposition approach, better or worse noise estimate would definitely affect the enhancement result. In this study, we do not treak around this issue and roughly estimate the noise from a few ms at the beginning of the input signal, which will be described in the next section.

Figure 3-3 illustrates signal flows of our proposed algorithm. Panel (a), (b) and (c) shows the original time domain waveform, the original auditory spectrogram and the spectro-temporal modulation energies at different (rate, scale) combinations, respectively. Panel (d) shows filtered spectro-temporal responses ST within speech regions and the enhanced responses by our proposed subspace decomposition

algorithm is shown in panel (f). Panel (e) shows the enhanced spectrogram by reconstruction of responses from (d), modulations of speech only [11]. Furthermore, panel (g) shows the final enhanced spectrogram by reconstruction of all enhanced responses in (f) from (d).

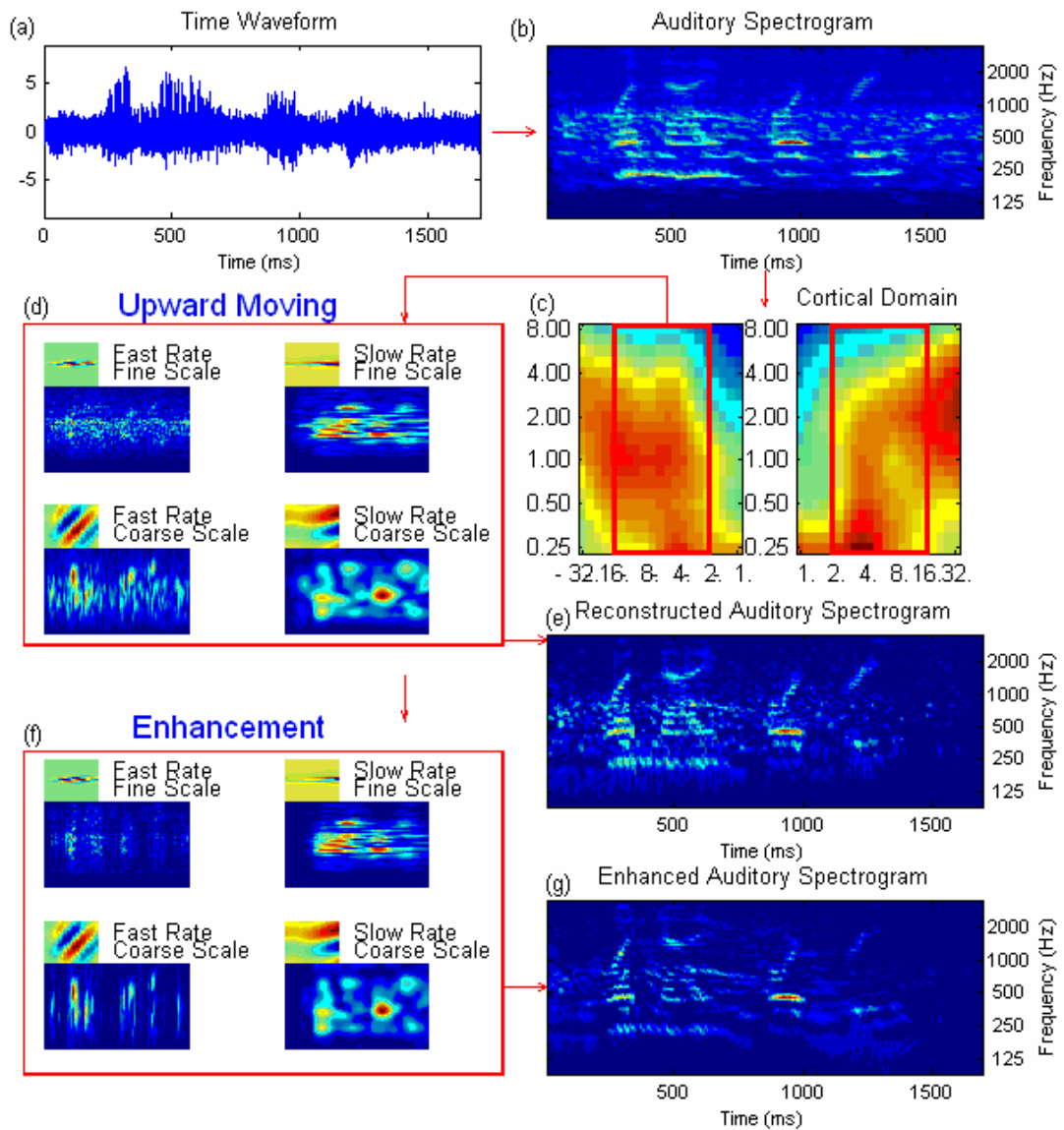


FIGURE 3-2 Flowchart of the proposed algorithm.

## 3.2 The 2D Neural Patterns in the Cortex

Equation (3-3) indicates the speech region in the cortical domain. Figure 3-3 shows STCRs in rate=1, 2, 4, scale=0.5, 1, 2, 4 combinations. It is noteworthy that (1) the lower the rate, the more time delay the STCR shows; (2) from the sampling theory, the upper bound of scale to avoid aliasing is 12 for the 24 samples per octave sampling in scale axis. In this section, we will discuss several issues related to the proposed algorithm, including (1) reduction of the computation and (2) a simple estimation of noise.

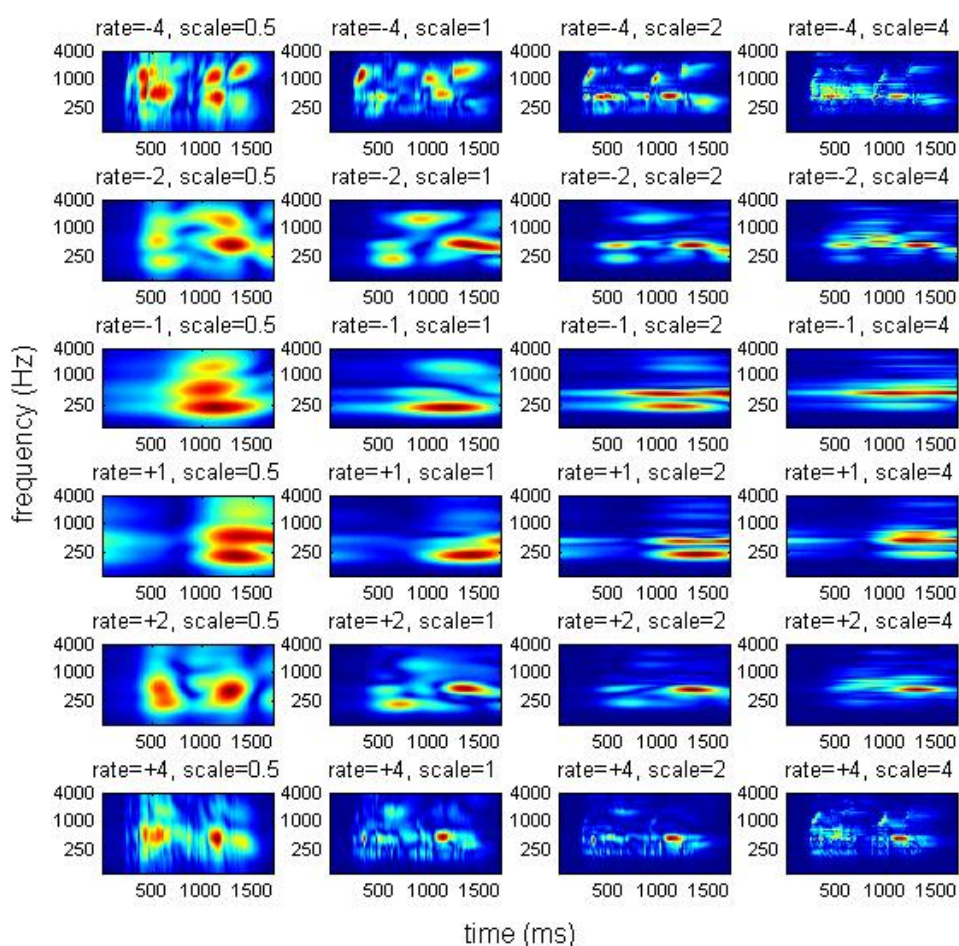


FIGURE 3-3 The STCRs of clean speech from fig 3-1 (left). Top to bottom are rate=-4, -2, -1, +1, +2, +4 and left to right are scale= 0.5, 1, 2, 4 respectively.



### 3.2.1 Dimension Redundancy Problem

Due to the high dimension of our spectrogram, our eigen-decomposition algorithm inherits much heavier computation than other speech enhancement algorithms, such as spectral-subtraction and Wiener filtering. To tackle such a problem, we can (1) reduce the dimension of the spectrogram or (2) partition the whole spectrogram into smaller segments for eigen-decomposition.

According to the sampling theory, bandwidth can be saved by down sampling the low-passed signals which has no high frequency components. Theoretically, in log-frequency dimension, we could downsample 3 times in the scale=4 cyc/oct channel since the upper bound of scale is 12 cyc/oct. However, in practice, we use less aggressive multiply numbers to avoid any possible aliasing. Table 3-1 shows the downsample multiply we use for channels at certain scales.

scale (cyc/oct)	0.5	1	2	4	8
downsample multiply	8	8	4	2	1

Table 3-1 The downsample multiply for scales.

For the same reason, in temporal dimension, we could downsample 25 times in the rate=2 Hz channel since the upper bound of rate is 50 Hz. Table 3-2 shows the downsample multiply we use corresponding to various rates.

rate (Hz)	2	4	8	16
downsample multiply	4	4	2	1

Table 3-2 The downsample multiply for rates.

Figure 3-4 shows original and downsampled versions of STCRs at various (rate, scale) combinations with downsample multiply as in Table 3-1 and 3-2. In the extreme case of rate=2 Hz and scale=0.5 cyc/oct, the size of the downsampled ST is reduced to 1/32 times of the original size. This downsampling dramatically decreases the overall computation.

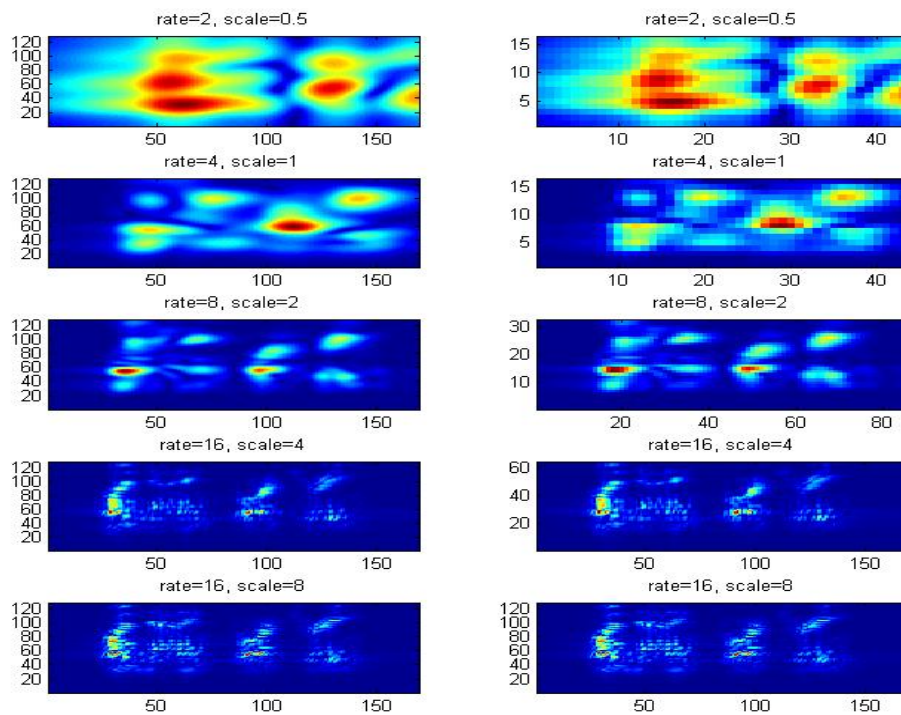


FIGURE 3-4 Examples of downsampled STCRs at various (rate, scale) combinations. Left column are the original STCRs and right column are the downsampled STCRs.

### 3.2.2 Frequency Band Division

In this work, we define four consecutive frames as a 40 ms “block” to be our 2D processing unit. In addition to downsampling the size of STCRs as mentioned in previous section, we further divide the processing unit along the frequency axis into several smaller units to reduce the computation. Another motivation of doing this is to

match hearing perceptions about frequency weighting. Dividing frequency bands in our auditory spectrogram might give us the flexibility of adjusting parameters in each band to fit certain noise sources, for instance, car noise in specific bands. However, more detailed study on frequency weighting is beyond the scope of this work. Here, we mainly consider computation reductions by this frequency band division.

Dorman et al. explored the influence of frequency bands on speech intelligibility [21]. From our viewpoints, the goal of speech enhancement is to sustain speech harmonics as much as possible while reducing the noise simultaneously. As shown in figure 3-5, we observe that most of the speech harmonics show up within the frequency range of around cochlear channel 28 (200 Hz) to channel 100 (1584 Hz). Our choices of channel 28 and 100 are for convenient implementation of down-sampling along the log-frequency axis. At the end, we divide each processing unit into three smaller units: below channel 28, from channel 28 to channel 100 and above channel 100.

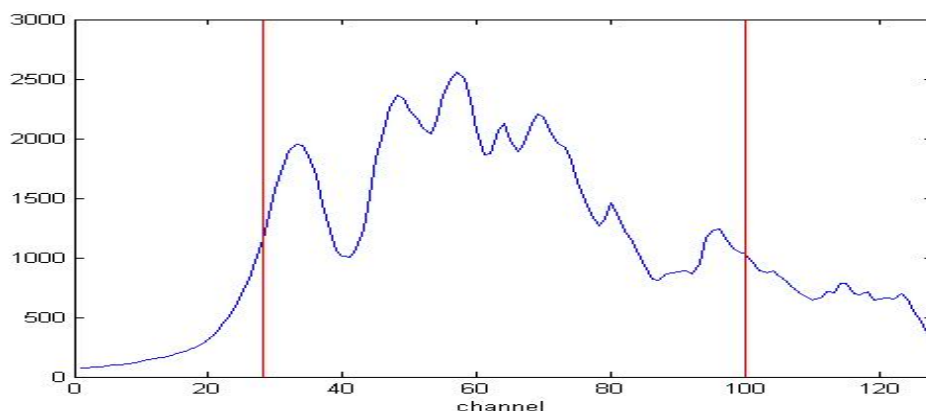


FIGURE 3-5 Example of an auditory spectrum along the cochlear channel (log-frequency) axis.

### 3.2.3 Window Length for Noise Estimation

The performance of each speech enhancement algorithm is largely affected by its accuracy in noise estimation. Considering different delays shown in STCRs in figure 3-4, estimating noise simply from a window with fixed duration at the beginning of signals is no longer valid. In this work, the 40 ms window selected has the strongest energy in that longer window. However, the duration of the longer window in lower rate STCRs is lengthened due to the severe temporal delays. Table 3-3 summarizes the longer window durations used here to find the 40 ms window to estimate noise for different rate STCRs.

rate (Hz)	2	4	8	16
estimated noise region (ms)	320	320	240	160

Table 3-3 The estimated noise region corresponding to each rate.

### 3.3 The weighted mask for HTK evaluation

Although the enhanced auditory spectrogram looks clean as shown in figure 3-2 (g), we still need to match testing features to training features as close as possible in order to achieve good recognition rates by HTK evaluations.

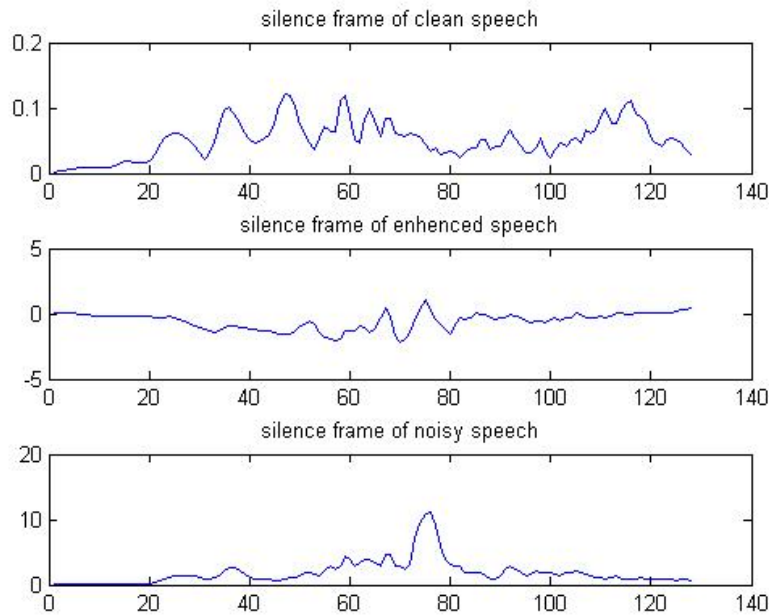


FIGURE 3-6 Silence/noise frame from clean speech (top), enhanced speech (middle) and noisy speech (bottom) at the 150<sup>th</sup> frame in the same speech as in figure 3-2.

As seen in Figure 3-6, the proposed algorithm does not make the noise spectrum identical to the silence spectrum after enhancement even though the noise energy is clearly suppressed. Such enhanced but distorted spectra won't give good performance while being used in HTK recognition evaluation. Therefore, a two-dimensional "mask" is generated by our enhancement algorithm and applied to the noisy (non-enhanced) spectrogram to reduce discrepancies between training and testing spectra in HTK evaluation.

First, we generate a binary mask by thresholding the enhanced spectrogram from our subspace decomposition algorithm. We set a small number instead of zero as the weight for non-speech portions and unity as the weight for speech portions in the spectrogram. Figure 3-7 shows the average recognition rates between 0 and 20 dB for various non-speech weights while  $\mu = 3$  (used in the subspace decomposition

algorithm) and  $threshold = (\max . \text{value of the spectrogram}) * 6\%$  . Different thresholds show similar performance curves as in Figure 3-7. Evidently, choosing lower weight for non-speech parts is not helpful to the speech recognition rates since highly suppressed non-speech bands in a speech frame make the spectrum easily mismatch to training spectra. Finally, the (1, 0.3) binary mask is smoothed by a 2D lowpass filter to avoid any sharp edges in the binary mask. Figure 3-8 shows the binary mask before and after smoothing.

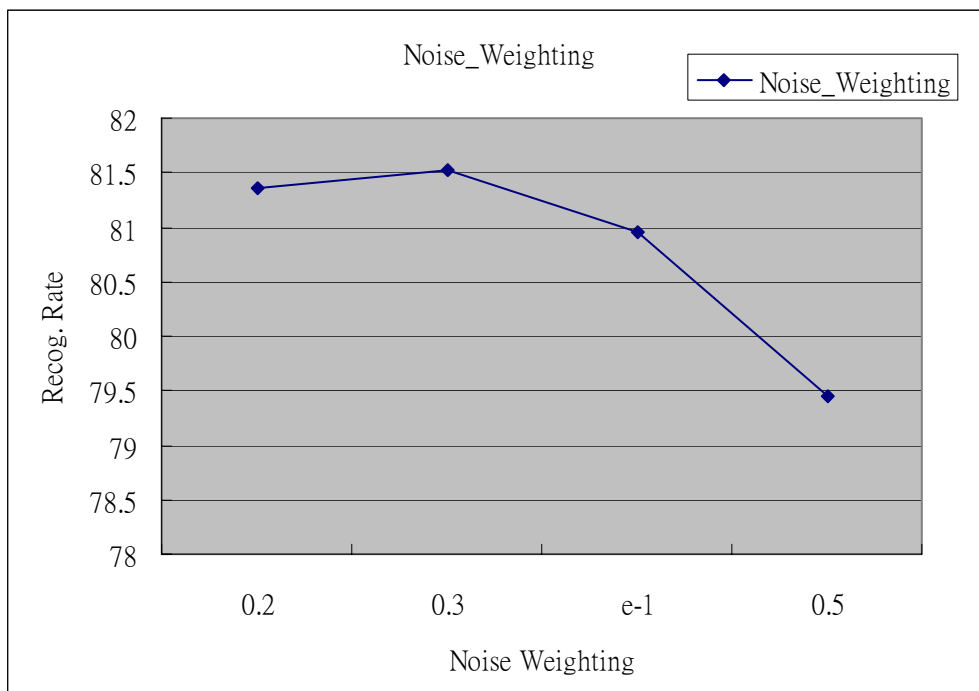


FIGURE 3-7 The noise weighting curve.

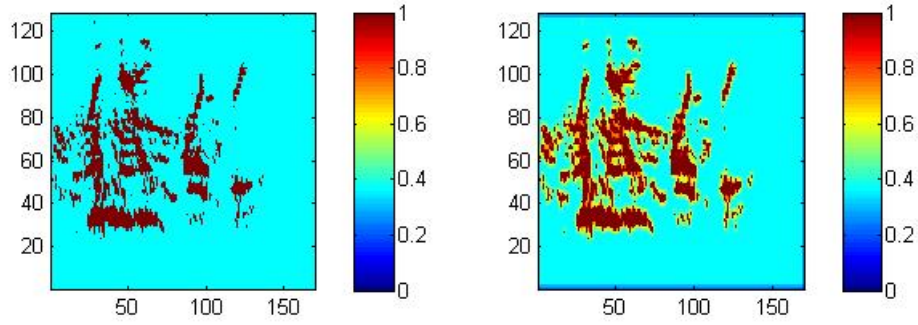


FIGURE 3-8 Binary mask derived from our enhancement algorithm. The left panel shows the original mask and the right panel shows the smoothed mask.

### 3.4 Summary

In this chapter, we present details of the proposed algorithm with following procedures:

1. Obtain the auditory spectrogram from auditory model analysis.
2. Generate smoothed spectrograms within speech regions in cortical domain (rate = 2~16 Hz, scale = 0.5~8 cyc/oct).
3. Downsample smoothed spectrograms by different multiply based on their rate and scale, divide each processing block into three broad subbands in frequency, and estimate noise in subbands as illustrated in section 3.2.2.
4. Align each subbanded segment of the spectrogram to an 1D representation (matrix => vector, equation 3-4) and apply the subspace decomposition algorithm in each segment as follows:
  - ◆ Apply eigen-decomposition of  $\Sigma = R_n^{-1} R_x$ . (equation 2-20)
  - ◆ Derive the optimal filter by:  $H_{opt} = V^{-1} \Delta_x (\Delta_x + \mu \cdot I)^{-1} V^T$ .
  - ◆ Obtain enhanced vector  $\hat{x} = H_{opt} \cdot y$ .

5. Reconstruct the 40 STCRs back to an auditory spectrogram and generate the weighting mask based on the enhanced spectrogram.
6. Multiply the weighting mask to the original spectrogram, as shown in figure 3-9, for HTK speech recognition evaluation.

In section 3.2, we depict the proposed enhancement algorithm in full details including (1) dimension reductions by downsampling, (2) frequency band division and (3) noise estimations in STCRs. Processes (1) and (2) above are purely for the sake of reducing computation complexity. As presented in section 3.3, we apply a weighting mask to reduce the discrepancies of silence between training and testing phases in the HTK evaluation.

Adjustable parameters in this proposed algorithm are the Lagrange multiplier  $\mu$  and the threshold which determines the noise region in the enhanced auditory spectrogram. If we set the  $\mu = 1$ , the equation of the subspace algorithm will become similar to the frequency-domain Wiener filter. However, unlike the Wiener filter, the subspace algorithm is in the eigen-space domain. Note, from equation 2-15, higher  $\mu$  has similar effects as with larger noise. Hence, with higher  $\mu$ , the optimal filter would not only eliminate more noise but also produce more speech distortions at the same time. Not surprisingly, the choice of the Lagrange multiplier is a trade-off decision between speech distortion and residual noise (quantitative evaluation will be given in next chapter). Similarly, the threshold that determines the noise region also has a trade-off effect between speech distortion and residual noise.



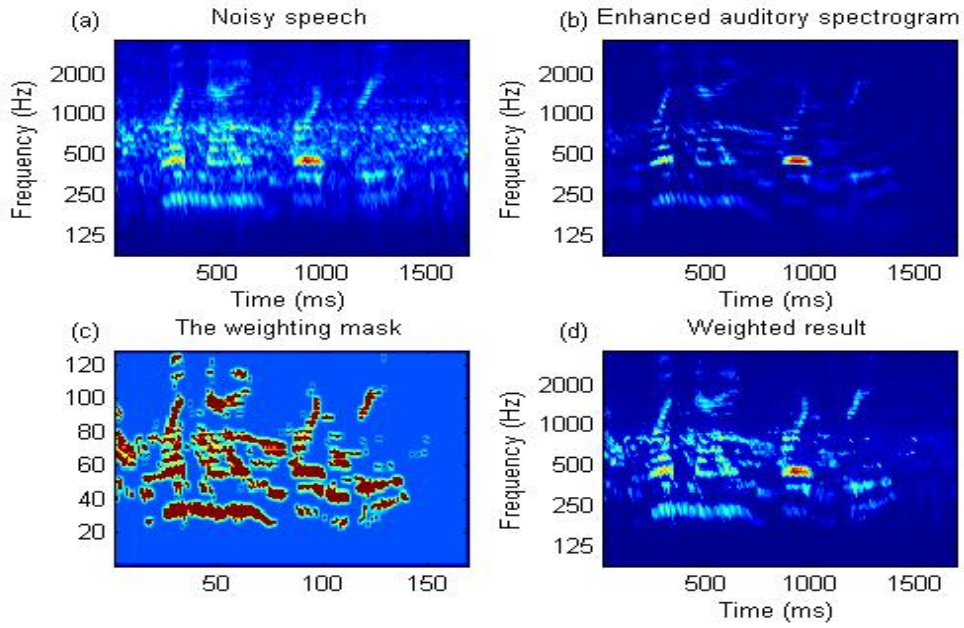


FIGURE 3-9 (a) The original noisy auditory spectrogram, (b) the enhanced auditory spectrogram, (c) the smoothed weighting mask and (d) the spectrogram, obtained by multiplying (c) to (a), used in HTK evaluation.



# Chapter 4

## Evaluation

In this chapter, we first introduce the (1) AURORA 2.0 database, (2) the compared algorithm, Advance Front-end feature Extraction (AFE), published by ETSI [22] and (3) the evaluation measurements used in this thesis. The HTK simulation results will be shown in section 4.2. Section 4.3 gives the speech distortion and residual noise error results from our proposed subspace decomposition algorithm. Summaries for these evaluations will be given at the end.

### 4.1 Database and Evaluation Measurements Introduction

AURORA 2.0 database is intended for the evaluation of front-end feature extraction algorithms in background noise and is used widely by speech researchers to evaluate and compare the performance of noise robust speech recognition algorithms.

The subspace algorithm is developed to minimize the speech distortion subject to certain levels of residual noise error. Therefore, we define measures of the speech

distortion and residual noise error to evaluate the proposed subspace algorithm.

### 4.1.1 AURORA 2.0

AURORA 2.0 is published by ETSI, European Telecommunication Standards Institute, for Distributed Speech Recognition (DSR) where the speech analysis is done at the telecommunication terminal and the recognition at central location in the telecom network.

The speech for this database is from TIDigits, consisting of connected digits spoken by American English speakers (downsampled to 8k Hz). A selection of 8 different real-world noises has been added to the speech over a range of signal to noise ratios. The 8 different noises are half grouped into class A (stationary noise), consisting of suburban train, babble, car and exhibition hall, and class B (non-stationary noise), consisting of restaurant, street, airport and train station.

The training data includes 8440 clean sentences spoken by 55 males and 55 females and the testing data is recorded by 52 males and 52 females who are different from those in clean dataset. The 8 different noises are added in 1001 sentences at 7 different SNR levels, including clean, 20dB, 15dB, 10dB, 5dB, 0dB and -5dB. Therefore, there are 56056 sentences for testing in total.

### 4.1.2 Advance Front-end feature Extraction

ETSI in 2003 specified algorithms for advanced front-end feature extraction and their transmission which form part of a system for distributed speech recognition.

Figure 4-1 shows the AFE terminal block scheme. VAD, in noise reduction, labels the non-speech frames. If VAD is enabled, non-speech frames could not be transmitted and therefore, it reduces the loading in the network transmission.

In this study, the VAD is disabled in order to emphasize and compare the noise reduction ability.

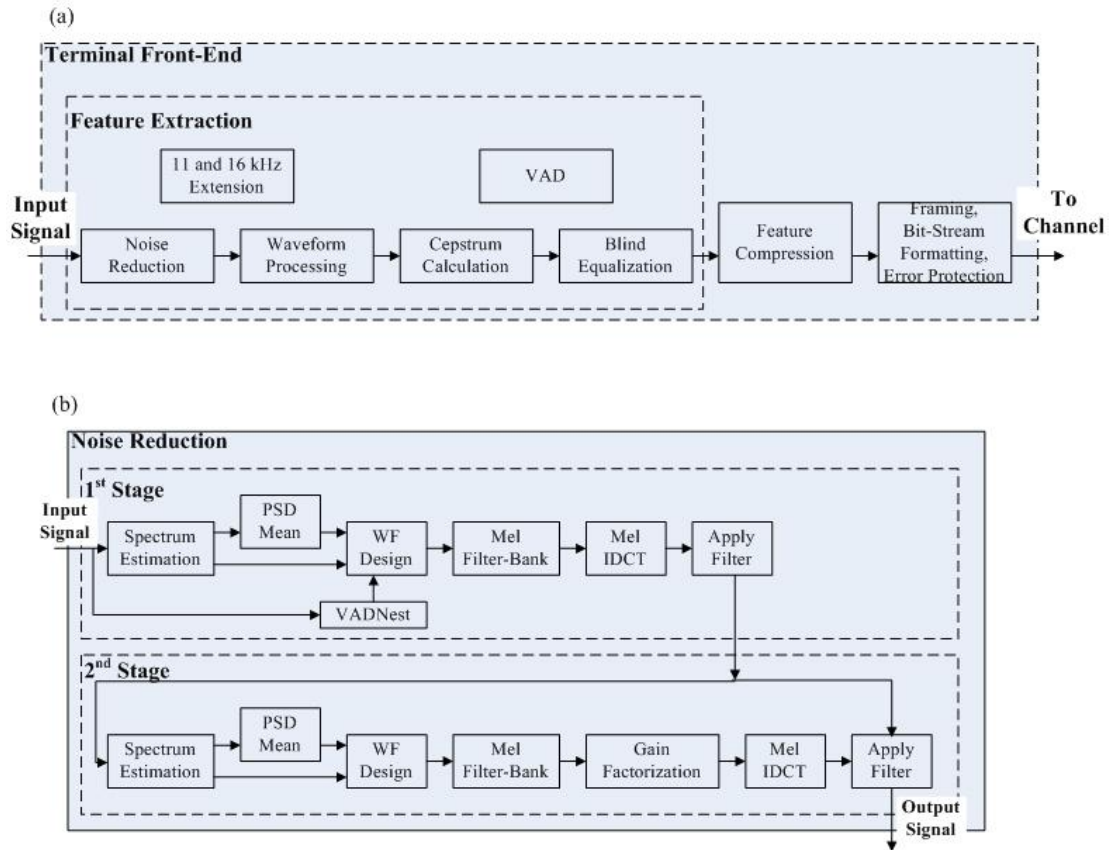


FIGURE 4-1 The AFE block scheme: (a) the terminal diagram and (b) the noise reduction block.

### 4.1.3 HTK Setting

We follow the training procedures presented in the AURORA 2.0 [23]. We use clean data training and match condition testing in this study. The match condition means the clean training data as well as testing data are both processed by the same enhancement algorithm.

Digits are modeled as whole-word HMMs with following parameters:

- 16 states per word (18 states in HTK notation with 2 dummy states at beginning and end).
- Simple left-to-right models without skips over states.
- 3 Gaussian mixtures per state.
- A feature vector size of 36 is used per frame for speech recognition. It is composed of 12 cepstral coefficients plus corresponding delta and acceleration coefficients.

Two pause models are defined. The first one called “sil” consists of 3 states with a mixture of 6 Gaussian models per state. The second pause model called “sp” is to model pauses between words. It consists of a single state which is tied with the middle state of the first pause model.

In this study, we use Auditory Cepstral Coefficients (ACCs) as the recognition feature and compare its performance to that of conventional Mel-Frequency Cepstral Coefficients (MFCCs). The robustness of ACCs over MFCCs has been demonstrated in [11].

HTK recognition results are expressed by three errors whose combinations determine the correct rate and accuracy rate. Related terminologies are defined as following:

- D : Deletion error, the number of non-recognized syllables.
- S : Substitution error, the number of wrongly recognized syllables.
- I : Insertion error, the number of syllables been recognized but not existed in answers.
- N : The total number of syllables.
- Correct rate =  $(N - D - S) / N \times 100\%$
- Accuracy rate =  $(N - D - S - I) / N \times 100\%$

In this study, the recognition rate stands for the accuracy rate. Figure 4-2 shows the ACC baseline and the performance of Yung's algorithm averaged over all kinds of noise in AURORA 2.0 database. Detailed results in each noise source are shown in Appendix II.

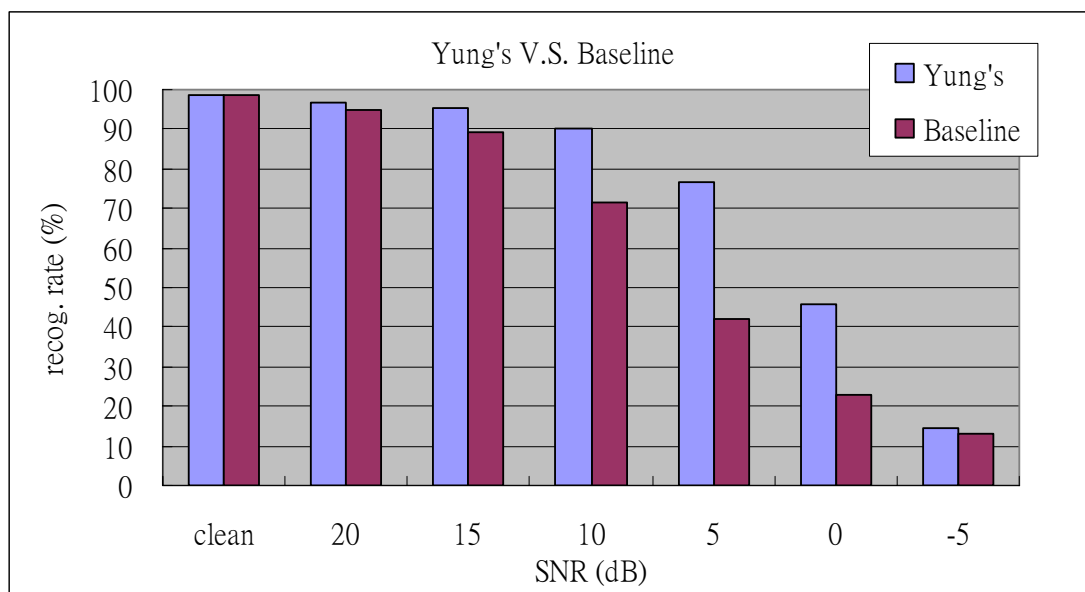


FIGURE 4-2 Recognition rate of ACC Baseline and Yung's result.

#### 4.1.4 Speech Distortion and Residual Noise

As mentioned in Chapter 3, the Lagrange multiplier  $\mu$  of the subspace algorithm would have opposite influences on speech distortion and residual noise. To calculate both measures, we first define the speech region and noise region in the spectrogram. The speech region is composed of those frames whose energies are greater than 2% of the maximum energy of the auditory spectrogram of clean speech. Other frames are considered as the noise region as shown in figure 4-3.

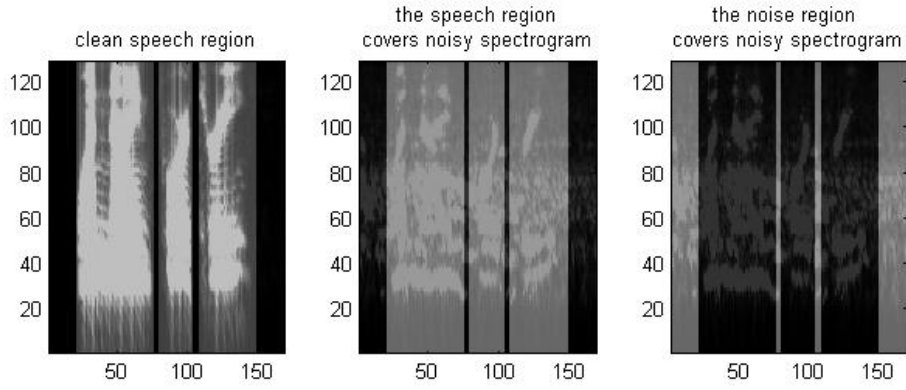


FIGURE 4-3 Speech region and the noise region from the clean auditory spectrogram (left). The middle and right subplot show the speech frames and noise frames cover the original noisy auditory spectrogram, respectively.

Measures of speech distortion and residual noise are defined by:

$$\text{Speech Distortion} = \frac{1}{\# \text{ of speech frame}} \sum_{\text{speech frame}} \frac{\|X - \hat{X}\|}{\|X\|} \quad (4-1)$$

$$\text{Residual Noise} = \frac{1}{\# \text{ of non-speech frame}} \sum_{\text{non-speech frame}} \|X - \hat{X}\| \quad (4-2)$$

where  $X$  and  $\hat{X}$  are auditory spectra of a certain frame of the clean speech and the enhanced speech. Note,  $\|X\|$  is close to zero in non-speech (silence) frames, hence, the residual noise measurement is not normalized by  $\|X\|$ . In addition,  $X$  and  $\hat{X}$  are first normalized by maximum values in auditory spectrograms of the whole clean sentence and the whole enhanced sentence, respectively.

## 4.2 HTK Results

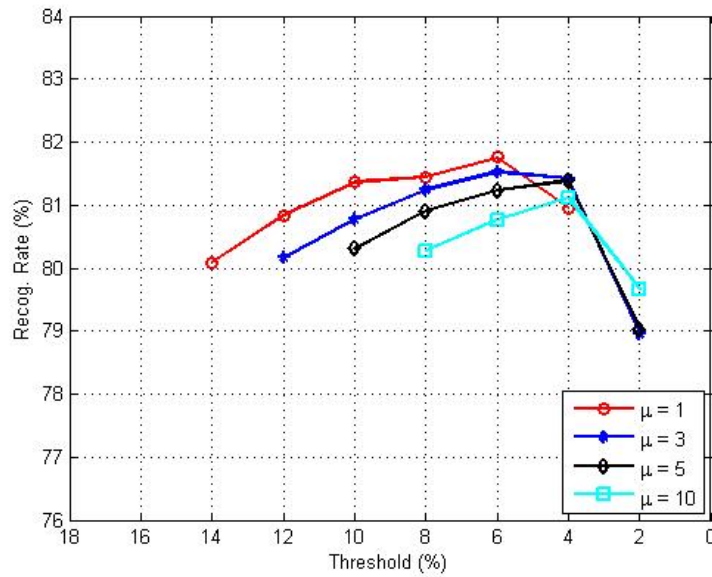


FIGURE 4-4 The simulation for different  $\mu$  and threshold.

Figure 4-4 shows the recognition rate of different  $\mu$  and threshold. We observe that the higher  $\mu$  has to be coupled with the lower threshold to achieve the same recognition rate. Not surprisingly, the higher the  $\mu$  is, the more severe the speech is degraded even the more suppressed the noise is. The highest speech recognition rate is achieved with  $\mu = 1$  and threshold= 6%. Details of recognition rates under such conditions are given in Appendix III.

Figure 4-5 shows the average recognition rates of AFE, Yung's method and the proposed algorithm. It shows our largest improvement over Yung's algorithm is in babble noise. This significant improvement is due to the decrease of the insertion errors as shown in table 4-1(a). On the other hand, the insertion errors in the car noise are low enough originally to not have further significant improvement, as shown in table 4-1(b).



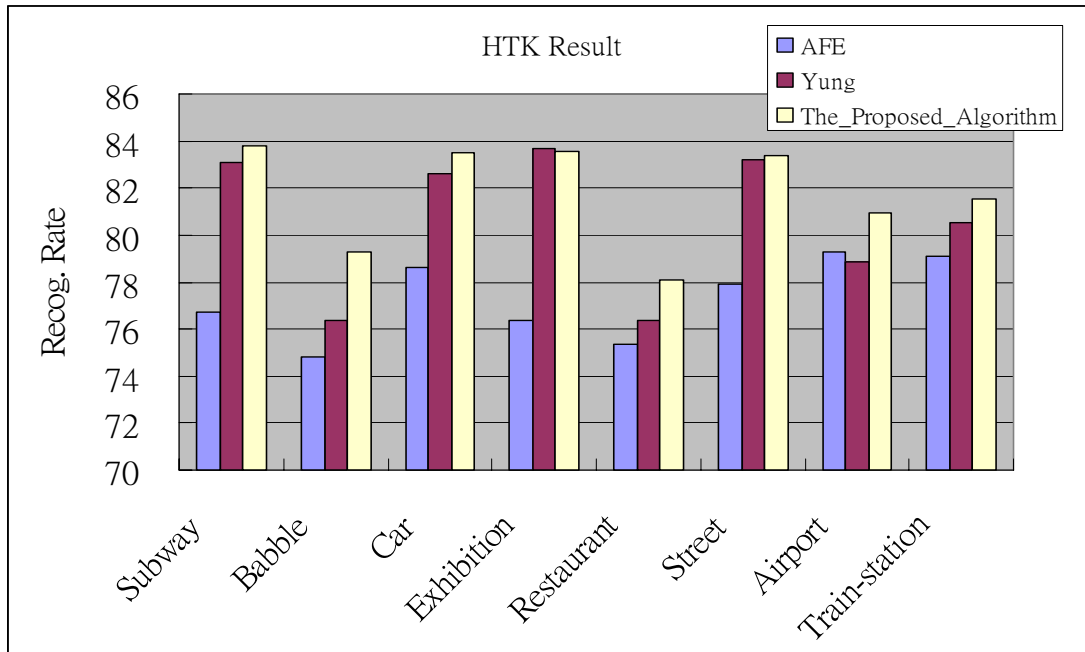


FIGURE 4-5 Average recognition rates (between 0~20 dB) of AFE, Yung's and the proposed algorithm in different noises.



The insertion error is the key factor to our improvement over Yung's method. In car noise environment, it is comparably easy to clean the noisy speech because of its stationarity. On the other hand, the babble noise is hard to compress by speech enhancement algorithms because it is relatively non-stationary and with characteristics (spectro-temporal modulations) comparatively close to speech. From our low insertion error in babble noise, we can say that our proposed algorithm not only enhances the speech but also suppresses the noise successfully. ( Appendix II shows the details about the hit and insertion rate. )

(a)	AFE		Yung		The_Proposed	
	Babble		Babble		Babble	
SNR/dB						
clean	98.76	1.00	98.58	0.60	98.37	0.60
20	96.95	4.66	97.10	0.63	97.04	0.67
15	94.47	4.84	95.62	0.79	95.41	0.67
10	87.94	4.90	91.93	2.63	91.29	1.12
5	72.04	4.47	80.26	10.40	81.65	4.14
0	44.35	3.23	50.76	19.47	52.84	15.39
-5	19.92	1.15	24.61	20.68	24.24	21.86
Average	79.15	4.42	83.13	6.78	83.65	4.40

(b)	AFE		Yung		The_Proposed	
	Car		Car		Car	
SNR/dB						
clean	98.78	0.89	98.84	0.69	98.51	0.75
20	97.58	0.78	97.38	0.42	96.96	0.48
15	95.91	0.78	96.06	0.42	95.47	0.60
10	90.40	0.54	92.48	0.48	92.01	0.63
5	74.11	0.21	81.69	0.45	81.66	0.36
0	39.01	0.03	49.63	2.51	54.28	0.69
-5	18.43	0.00	19.15	2.06	20.13	2.77
Average	79.40	0.47	83.45	0.85	84.07	0.55

Table 4-1 Hit / insertion rate of AFE, Yung's and the proposed algorithm in (a) babble and (b) car noise.

Figure 4-6 shows average recognition rates of AFE, Yung's method and our proposed algorithm. It shows the performance boost of around 4% in 0dB and 3% in 5dB over Yung's algorithm; and of around 6% in 0dB and 8% in 5dB over the AFE. In high SNR conditions, our performance is comparable to Yung's performance because less noise exists to work with. Overall speaking, our proposed algorithm performs better than Yung's and AFE algorithm, hence it's more robust.

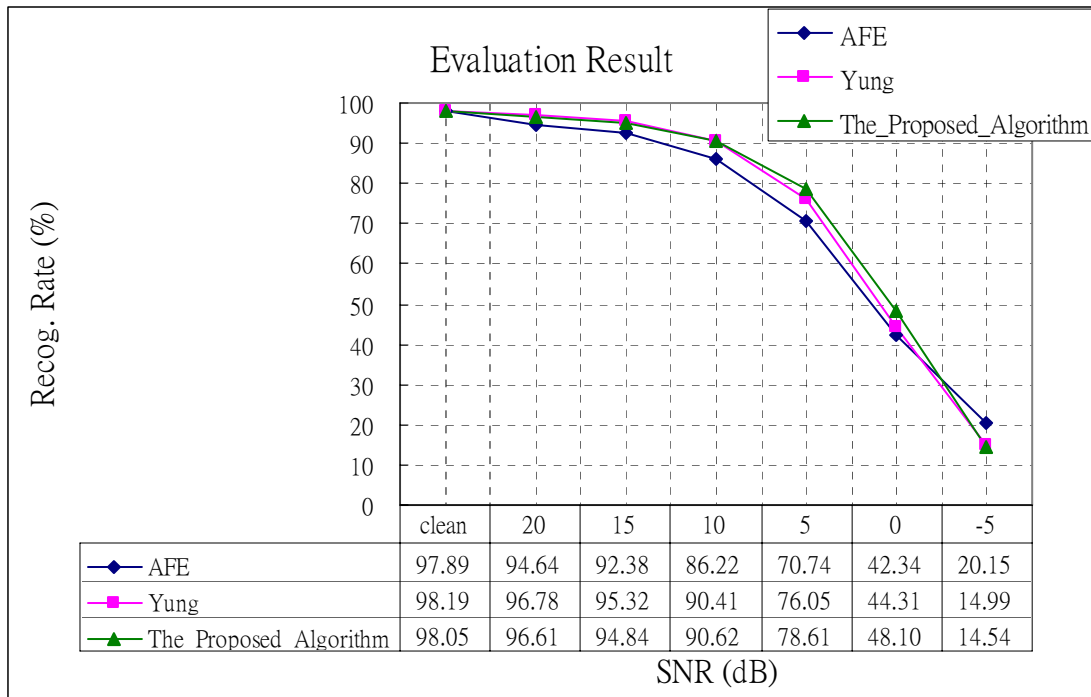
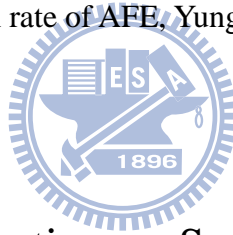


FIGURE 4-6 Average recognition rate of AFE, Yung’s and the proposed algorithm.



### 4.3 Performance Evaluation on Speech Distortion and Residual Noise

Average measures of speech distortion and residual noise between 20 ~ 0dB are shown in figure 4-7 and 4-8. Numbers are calculated by treating enhanced, masked clean speech as the clean pattern, and treating enhanced, masked noisy speech as the test pattern. Such procedures are designed to match the “match condition” scheme in HTK recognition. Obviously, the results reveal the effectiveness of the proposed algorithm. Distortions decrease gradually as  $\mu$  increases, especially visible in the measure of the residual noise.

The speech distortion and residual noise are proportional to the hit and insertion rate in HTK tasks in some way. From figure 4-7 and 4-8, our proposed algorithm has

superior performance in these two distortion measures than Yung's previous study, which is also consistent to the performance shown in the HTK recognition task. However, the mathematical correlation between these two measures and the speech recognition rate is beyond the scope of this thesis.

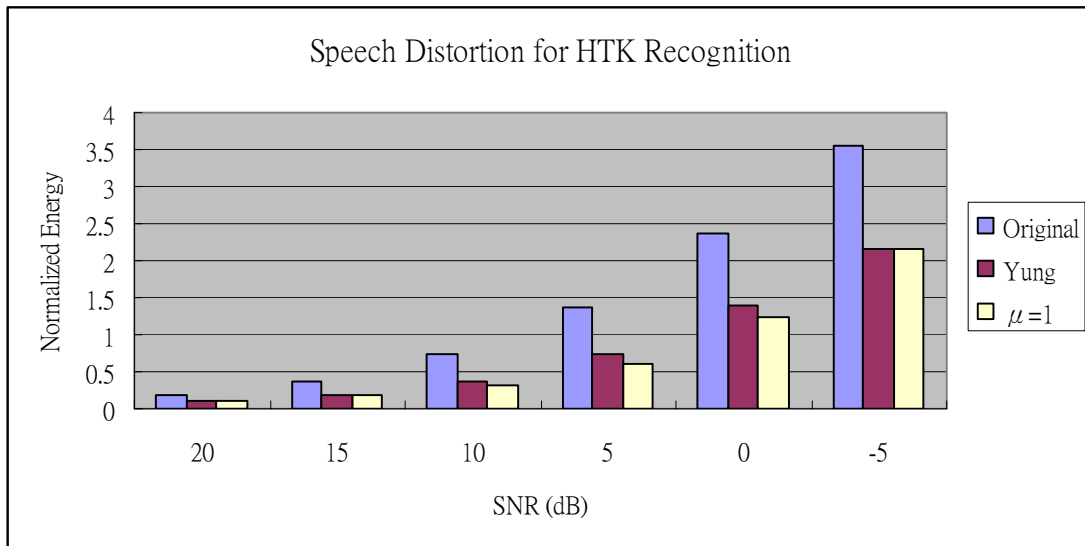


FIGURE 4-7 Average speech distortion shown in spectrograms.

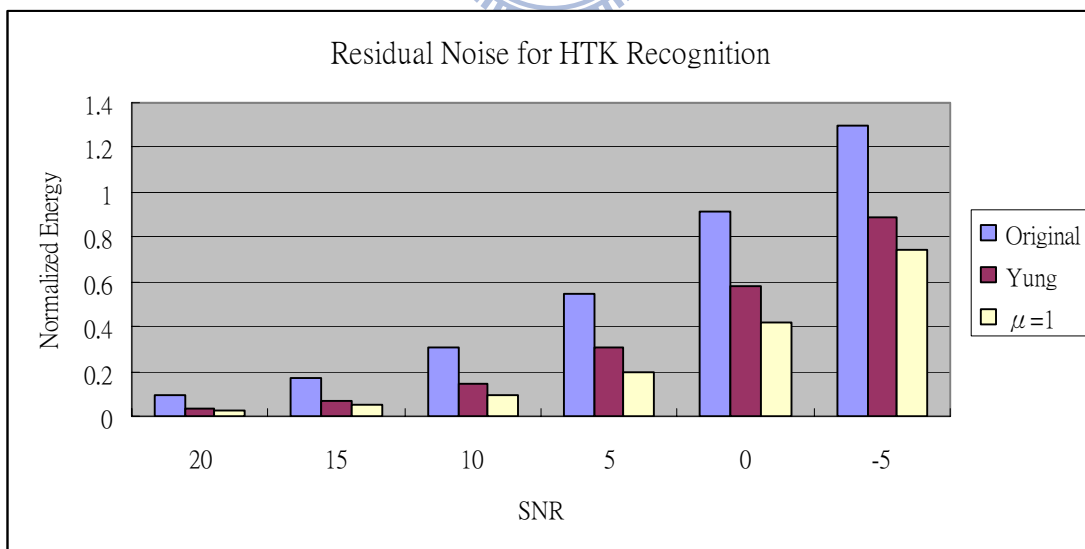


FIGURE 4-8 Average residual noise shown in spectrograms.

Figure 4-9 and 4-10 illustrate the trade-off phenomenon between speech distortion and residual noise. These two quantities are measured between the original clean speech and the enhanced noisy speech (which is not a “match condition” comparison as in the preceding comparison). When the parameter  $\mu$  increases, the speech distortion gradually increases but the residual noise gradually decreases. These results clear confirm that  $\mu$  controls degrees of speech distortion and residual noise in opposite directions.

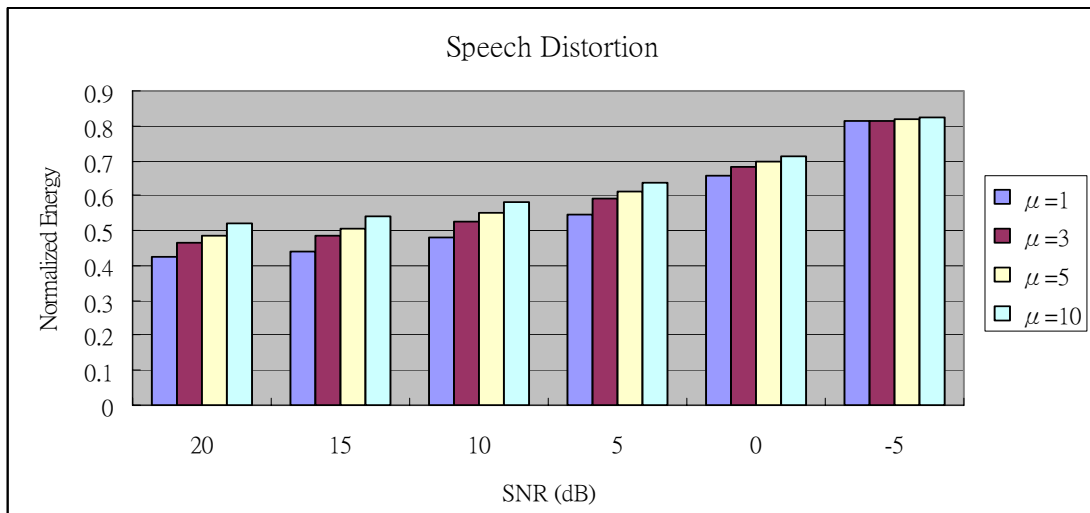


FIGURE 4-9 Speech distortion measures for different  $\mu$  and SNR.

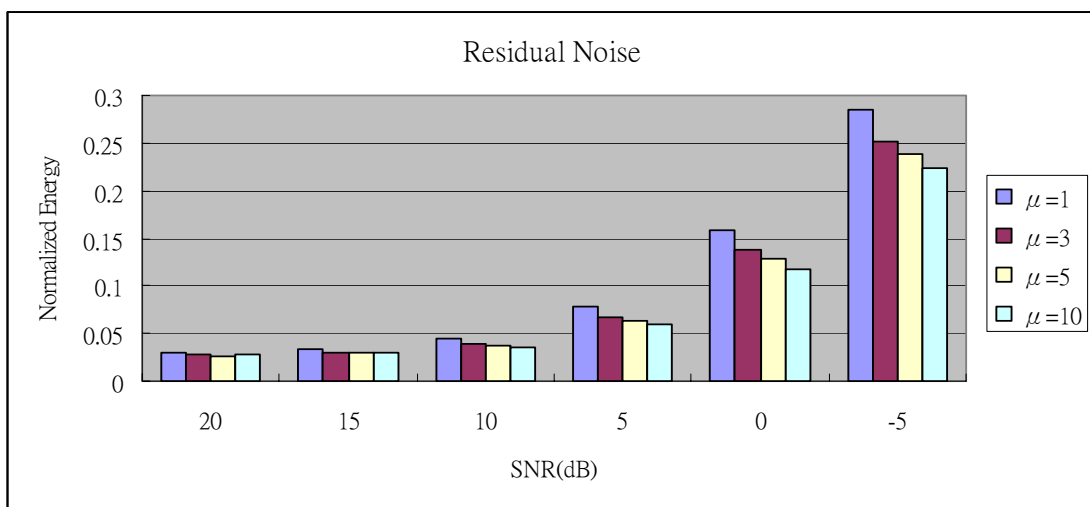
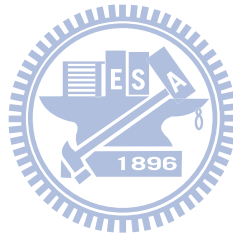


FIGURE 4-10 Residual noise measures for different  $\mu$  and SNR.

## 4.4 Summary

In this chapter, we introduced the AURORA 2.0 database, AFE algorithm and two evaluation measurements: (1) HTK speech recognition and (2) speech distortion and residual noise. Our goal of improving the recognition rate for DSR systems over AFE algorithm is reached by enhancing the speech and suppressing the noise simultaneously, especially in babble noise. In addition, the trade-off between speech distortion and residual noise was investigated and demonstrated. However, lack of methods of transferring auditory spectrogram back into waveform makes listening tests for enhanced speech quality infeasible. Although the auditory spectrogram looks clean, it does not guarantee the quality of the enhanced speech which is usually not considered in DSR systems.



# Chapter 5

## Conclusion and Future Works

The proposed subspace decomposition algorithm is performed on multi-dimensional cortical representations of the speech region (rate=2-16 Hz, scale=0.5-8 cycle/octave) . In each (rate, scale) combinational cortical representation, our algorithm suppresses the noise in the eigen-space domain through the eigen-decomposition analysis. We exhibit every aspect of the proposed algorithm in details and its performance in chapter 3 and 4. For HTK evaluations, the proposed algorithm gives the improvement of around 6% in 0dB and 8% in 5dB over the AFE. As for speech distortion and residual noise measurements, they clear confirm that  $\mu$  controls the trade-off phenomenon between both.

Here, we address major disadvantages of our proposed system:

- (1) The 2D eigen-decomposition analysis is with high computational complexity. It is not practical for real-time on-line systems.
- (2) Lack of phase information. The proposed algorithm works on various degrees of filtered modulations of the auditory spectrogram. It then

reconstructs the enhanced spectrogram, which is short of the phase information of the time-domain waveform. Therefore, it is not possible to invert our enhanced spectrograms back to acoustical sounds without further distortions for subjective sound quality listening tests.

- (3) The rough noise estimation. Noise estimation techniques play the important role in most of the speech enhancement algorithms. This work focuses on adopting the subspace decomposition to the perceptual representations; noise estimation should be fully studied in the future.

Finally, we point out several directions for future evolution of our speech enhancement algorithm. First, the noise estimation process needs further investigation since many speech enhancement techniques work well due to their accurate noise estimates. Second, build an inverse process to invert the auditory spectrogram back to time-domain waveform with acceptable distortions. A successful real-time one-shot inverse would be a huge contribution to our auditory model. Once it is done, any manipulations on the spectrogram can then be heard as acoustical sounds to make interactive listening tests feasible. Third, apply other feature normalization processes, such as Cepstral Mean Subtraction (CMS) and Cepstral Mean and Variance Normalization (CMVN), to our Auditory Cepstral Coefficients to further improve the performance of this perceptual feature.



# Appendix I

## Pre-whitening Verification

Here, we prove the pre-whitening approach that is used in the proposed algorithm. Recall equation (2-16),

$$R_d = R^T R = L \cdot L^T \quad (2-16)$$

where  $R_d = E[d \cdot d^T]$  is the auto-correlation matrix of noise vector  $d$ ,  $L$  is the transpose of  $R$  which is the factor of Cholesky factorizing to  $R_d$ .

Thus, the pre-whitening equation is given by (2-17):

$$\begin{aligned} L^{-1}y &= L^{-1}x + L^{-1}d \\ y' &= x' + d' \end{aligned} \quad (2-17)$$

Our goal is to demonstrate that  $R_{d'}$  is an identical matrix. The proof is as following:

$$\begin{aligned} R_{d'} &= E[d' \cdot d'^T] \\ &= E[L^{-1}d \cdot (L^{-1}d)^T] \\ &= E[L^{-1}d \cdot d^T \cdot L^{-T}] \\ &= L^{-1} \cdot R_d \cdot L^{-T} \\ &= L^{-1} \cdot (L \cdot L^T) \cdot L^{-T} \\ &= I \quad \# \end{aligned}$$

## Appendix II The AFE and Yung's Result

The AFE HTK result.

SNR/dB	Subway	Babble	Car	Exhibition	A-Average	
clean	98.00	97.76	97.88	97.90	97.89	
20	95.49	92.29	96.81	95.53	95.03	
15	92.88	89.63	95.14	92.97	92.66	
10	85.29	83.04	89.86	86.52	86.18	
5	69.67	67.56	72.41	68.59	69.56	
0	40.13	41.44	38.98	38.11	39.67	
-5	21.83	18.77	18.43	18.51	19.39	
Average	76.69	74.79	78.64	76.34	76.62	
SNR/dB	Restaurant	Street	Airport	Train-station	B-Average	Total Average
clean	98.00	97.76	97.88	97.90	97.89	97.885
20	91.80	95.95	94.21	95.00	94.24	94.635
15	89.07	93.23	92.63	93.52	92.11	92.38375
10	82.78	86.46	87.24	88.58	86.27	86.22125
5	68.59	71.52	73.64	73.93	71.92	70.73875
0	44.55	42.50	48.58	44.46	45.02	42.34375
-5	19.47	20.86	22.40	20.95	20.92	20.1525
Average	75.36	77.93	79.26	79.10	77.91	77.2645

The hit and insertion rate of AFE. (hit / insertion)

SNR/dB	Subway		Babble		Car		Exhibition	
clean	98.99	0.98	98.76	1.00	98.78	0.89	96.42	1.30
20	96.25	0.77	96.95	4.66	97.58	0.78	96.82	1.30
15	93.52	0.64	94.47	4.84	95.91	0.78	94.17	1.20
10	85.97	0.68	87.94	4.90	90.40	0.54	56.96	1.30
5	70.10	0.43	72.04	4.47	74.11	0.21	69.36	0.77
0	40.22	0.09	44.35	3.23	39.01	0.03	38.35	0.25
-5	21.83	0.00	19.92	1.15	18.43	0.00	18.73	0.22
SNR/dB	Restaurant		Street		Airport		Train-station	
clean	98.99	0.98	98.76	1.00	98.78	0.89	99.20	1.30
20	97.39	5.59	96.77	0.82	97.44	3.22	97.47	2.47
15	94.93	5.89	93.98	0.76	95.65	3.01	95.56	2.04
10	88.70	5.93	87.03	0.57	90.40	3.16	90.74	2.16
5	74.24	5.65	71.89	0.36	76.71	3.07	75.81	1.88
0	49.65	5.10	42.62	0.12	50.28	1.70	45.45	0.99
-5	21.86	2.39	20.86	0.00	23.74	1.34	5.92	0.40

### Yung's Result

SNR/dB	Subway	Babble	Car	Exhibition	A-Average	
clean	98.31	97.97	98.15	98.33	98.19	
20	97.02	96.46	96.96	95.80	96.56	
15	95.76	94.83	95.65	94.97	95.30	
10	91.62	89.30	92.01	90.77	90.93	
5	79.61	69.86	81.24	81.61	78.08	
0	51.27	31.29	47.12	55.23	46.23	
-5	20.69	3.93	17.09	22.83	16.14	
Average	83.06	76.35	82.60	83.68	81.42	
SNR/dB	Restaurant	Street	Airport	Train-station	B-Average	Total Average
clean	98.31	97.97	98.15	98.33	98.19	98.19
20	96.96	96.92	96.90	97.25	97.01	96.78375
15	94.60	95.68	95.26	95.80	95.34	95.31875
10	87.53	91.44	89.98	90.65	89.90	90.4125
5	67.79	79.84	72.89	75.55	74.02	76.049125
0	34.94	52.15	39.19	43.32	42.40	44.31375
-5	6.36	21.01	12.65	15.33	13.84	14.98625
Average	76.36	83.21	78.84	80.51	79.73	80.575575

### The hit rate and insertion rate of Yung's result. (hit / insertion)

SNR/dB	Subway		Babble		Car		Exhibition	
clean	98.96	0.64	98.58	0.60	98.84	0.69	98.98	0.65
20	97.61	0.58	97.10	0.63	97.38	0.42	97.44	1.64
15	96.56	0.80	95.62	0.79	96.06	0.42	96.36	1.39
10	92.94	1.32	91.93	2.63	92.48	0.48	92.81	2.04
5	83.88	4.27	80.26	10.40	81.69	0.45	83.71	2.10
0	58.95	7.68	50.76	19.47	49.63	2.51	57.33	2.10
-5	25.82	5.13	24.61	20.68	19.15	2.06	24.38	1.54
SNR/dB	Restaurant		Street		Airport		Train-station	
clean	98.96	0.64	98.58	0.60	98.84	0.69	98.98	0.65
20	97.88	0.92	97.34	0.42	97.32	0.42	97.78	0.52
15	96.59	2.00	96.13	0.45	95.85	0.60	96.58	0.77
10	92.69	5.16	92.23	0.79	91.89	2.06	92.59	1.94
5	80.84	13.05	81.35	1.51	79.78	6.89	79.91	4.38
0	55.51	20.57	54.38	2.15	52.22	13.03	50.85	7.53
-5	25.67	19.31	23.46	2.45	25.56	12.91	22.34	7.00

## Appendix III The Proposed algorithm

### HTK Recognition Result

SNR/dB	Subway	Babble	Car	Exhibition	A-Average	
clean	98.10	97.76	97.76	98.58	98.05	
20	96.75	96.37	96.48	96.39	96.50	
15	95.12	94.74	94.87	94.11	94.71	
10	91.74	90.18	91.38	89.97	90.82	
5	81.52	77.51	81.30	81.46	80.45	
0	53.82	37.45	53.59	55.82	50.17	
-5	19.04	2.39	17.36	21.47	15.07	
Average	83.79	79.25	83.52	83.55	82.53	
SNR/dB	Restaurant	Street	Airport	Train-station	B-Average	Total Average
clean	98.10	97.76	97.76	98.58	98.05	98.05
20	96.81	96.55	96.48	97.04	96.72	96.60875
15	95.09	94.71	95.05	95.06	94.98	94.84375
10	89.38	91.41	90.55	90.31	90.41	90.615
5	72.55	80.08	77.30	77.17	76.78	78.61125
0	36.66	54.05	45.21	48.16	46.02	48.095
-5	7.28	21.37	11.60	15.77	14.01	14.535
Average	78.10	83.36	80.92	81.55	80.98	81.75475

The proposed algorithm HTK result. (hit / insertion)

SNR/dB	Subway		Babble		Car		Exhibition	
clean	98.65	0.55	98.37	0.60	98.51	0.75	99.11	0.52
20	97.51	0.77	97.04	0.67	96.96	0.48	97.53	1.14
15	95.95	0.83	95.41	0.67	95.47	0.60	95.83	1.73
10	92.60	0.86	91.29	1.12	92.01	0.63	92.10	2.13
5	83.88	2.36	81.65	4.14	81.66	0.36	83.83	2.38
0	60.42	6.60	52.84	15.39	54.28	0.69	57.88	2.07
-5	26.90	7.86	24.24	21.86	20.13	2.77	23.26	1.79
SNR/dB	Restaurant		Street		Airport		Train-station	
clean	98.65	0.55	98.37	0.60	98.51	0.75	99.11	0.52
20	97.82	1.01	97.04	0.48	97.14	0.66	97.56	0.52
15	96.44	1.35	95.28	0.57	95.65	0.60	95.68	0.62
10	92.82	3.44	92.08	0.67	91.86	1.31	91.92	1.60
5	81.09	8.54	81.23	1.15	80.58	3.28	80.50	3.33
0	55.66	19.01	55.65	1.60	54.91	9.69	54.12	5.95
-5	24.04	16.76	23.85	2.48	25.26	13.66	22.28	6.51

## REFERENCE

- [1] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586-1604, 1979.
- [2] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, pp. 126-137, 1999.
- [3] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series: with engineering applications*: MIT press Cambridge, MA, 1964.
- [4] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*, 1993.
- [5] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," in *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, 1993, pp. 355-358 vol.2.
- [6] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and language*, vol. 1, pp. 109-130, 1986.
- [7] J. R. Cohen, "Application of an auditory model to speech recognition," *The Journal of the Acoustical Society of America*, vol. 85, p. 2623, 1989.
- [8] T. Dau, D. Puschel, and A. Kohlrausch, "A quantitative model of the effective signal processing in the auditory system. I. Model structure," *Journal of the Acoustical Society of America*, vol. 99, pp. 3615-3622, 1996.
- [9] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *The Journal of the Acoustical Society of America*, vol. 106, p. 2719, 1999.
- [10] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of Dynamic Spectra in Ferret Primary Auditory Cortex: I. Characteristics of single unit responses to moving ripple spectra," 1996.
- [11] Y.-N. Hung, "Speech Enhancement Method based on Auditory Perceptual Model " in *Communication Engineering*. vol. master Hsin-Chu, Taiwan: National Chiao Tung University, 2008, p. 50.
- [12] W. Kuansan and S. Shamma, "Self-normalization and noise-robustness in early auditory representations," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 421-435, 1994.
- [13] N. Kowalski, D. A. Depireux, and S. A. Shamma, "Analysis of Dynamic Spectra in Ferret Primary Auditory Cortex: II. Prediction of Unit Responses to

- Arbitrary Dynamic Spectra," 1996.
- [14] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, pp. 251-266, 1995.
- [15] H. Yi and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, pp. 334-341, 2003.
- [16] Zhujie and Y. L. Yu, "Face recognition with eigenfaces," in *Industrial Technology, 1994. Proceedings of the IEEE International Conference on*, 1994, pp. 434-438.
- [17] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR '91., IEEE Computer Society Conference on*, 1991, pp. 586-591.
- [18] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04). IEEE International Conference on*, 2004, pp. I-601-4 vol.1.
- [19] S. Haykin, *Adaptive filter theory*: Prentice-Hall, Inc. Upper Saddle River, NJ, USA, 1996.
- [20] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*: MIT press, 1990.
- [21] M. F. Dorman, P. C. Loizou, and D. Rainey, "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs," *The Journal of the Acoustical Society of America*, vol. 102, p. 2403, 1997.
- [22] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," 2000.