Two-channel noise reduction and post-processing for speech enhancement

Zhang, Xinxin

2008

Zhang, X. (2008). Two-channel noise reduction and post-processing for speech enhancement. Master’ s thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/3522

https://doi.org/10.32657/10356/3522

Two-Channel Noise Reduction and Post-Processing for Speech Enhancement

ZHANG XINXIN

SCHOOL OF ELECTRICAL & ELECTRONIC ENGINEERING

NANYANG TECHNOLOGICAL UNIVERSITY

2008

Two-Channel Noise Reduction and Post-Processing for Speech Enhancement

Zhang Xinxin

School of Electrical & Electronic Engineering

A thesis submitted to the Nanyang Technological University

in fulfillment of the requirement for the degree of

Master of Engineering

2008

ACKNOWLEGEMENTS

I would like to express my appreciation to Professor Koh Soo Ngee who supervises me on this project and provides me the convenience to access the resources along the way.

I would also like to thank Assoc. Professor Soon Ing Yann, Mr V. R. Reju, Mr Zhang Yi, Dr Liu Xiaobei and Mr You Changhuai for helping me coping with the difficulties I encountered. Thanks also go to the staff members in Communication Research Lab, Mr Lim Cheng Chye and Ms Zhang Huiyu, for their help and supports.

Last but not least, I would like to thank my family members and my friends who have given me lots of help and support.

SUMMARY

This thesis is focused on speech enhancement techniques based on short-time spectral amplitude (STSA) estimation. Some common STSA estimation methods and the β-order minimum mean-square error (MMSE) estimation technique incorporating auditory masking properties (β-masking in short) are reviewed. A word pair recognition accuracy test is done to assess the intelligibility of the β-masking enhanced speech signals. Though the β-masking technique outperforms most of the existing STSA estimation speech enhancement methods, some slight tonal distortion is audible in the processed speech signals when the background noise level is very high, because some speech spectral components are over-attenuated while some are not due to the enhancement algorithm. Two post-processing techniques are proposed to improve the quality of the single-channel β-masking enhanced speech signals. One technique involves non-linear high-frequency regeneration, which uses the lower-band spectral information to re-synthesize the upper-band spectral structure. The other technique involves re-synthesis of the weak spectral components using the autocorrelations of the strong spectral components.

With the increasing use of mobile communication, a two-channel speech enhancement method for communication in a car environment is also studied. To achieve a better performance, the single-channel β-masking method is incorporated within the two-channel enhancement system. The resulting output speech signals have low background noise and the distortion to the speech components is also very low, thus achieving an overall very satisfactory speech enhancement performance.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS AND SYMBOLS

Abbreviations

β-masking	β-order MMSE speech enhancement incorporated masking properties
DFT	Discrete Fourier Transform
ETSI	European Telecommunications Standard Institute
FTMS	fluctuation-tracking minima-search
HPF	high-pass filter
HMM	Hidden Markov Model
IDCT	Inverse Discrete Cosine Transform
IDFT	Inverse Discrete Fourier Transform
LPF	low-pass filter
LSA	log spectral amplitude
MMSE	minimum mean-square error
MOS	Mean Opinion Score
NHR	non-linear high-frequency regeneration
NTW	narrowband to wide band transformation
PDF	probability density function
PESQ	Perceptual evaluation of speech quality
PMD	psycho-acoustically motivated distortion
PSD	power spectral density
SNR	Signal-to-noise ratio
STSA	short-time spectral amplitude
VADNest	Voice Activity Detection used for Noise estimation
WF	Wiener Filter
ZCR	zero crossings rate

Symbols

A_k	amplitude of *S(k)*
α_{NHR}	Weighting factor of the NHR method
$E\{\cdot\}$	the expectation operator
ξ_k	*a priori* SNR
$H_{css}(k)$	the two-sensor filter gain for *k*-th spectral component
$G(k)$	suppression filter gain for *k*-th spectral component
$G_{\beta}(k)$	gain function of β-order MMSE estimator
$G_{\beta p}(k)$	gain function of β-order MMSE incorporating masking properties
$G_{MMSE}(k)$	gain function of MMSE estimator
γ_k	*a posteriori* SNR
$\Gamma(\cdot)$	the gamma function
J	cost function, distortion measure
$M(a;c;x)$	the confluent hypergeometric function
$N(k)$	*k*-th spectral component of a noise process
N_k	amplitude of *N(k)*
$n(t)$	noise at time *t*
R_k	amplitude of *X(k)*
$S(k)$	*k*-th spectral component of a clean speech signal
$\hat{S}(k)$	estimated *k*-th spectral component of a clean speech signal
$s(t)$	clean speech signal at time *t*
$\hat{s}(t)$	estimated clean speech signal at time *t*
$X(k)$	*k*-th spectral component of a noisy speech signal
$x(t)$	noisy speech signal at time *t*

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Motivation

In a typical communication system, speech may be corrupted by background noise which is likely to cause listening fatigue. It also invariably degrades the intelligibility of speech, resulting in misunderstanding in a conversation. Under some circumstances, uncorrupted speech signals are of great importance for certain applications. For example, when a pilot is communicating through a wireless device with an air traffic control tower, a high level of background noise could be distractive and annoying. If wrong information is interpreted due to the background noise, the consequence could be very serious. Therefore, it is essential or even vital to process noisy speech so as to enhance its perceptual quality before it is received by the intended listener. Besides improving the perceptual quality, speech enhancement may also be useful for automatic speech recognition systems.

Speech enhancement is a process that removes the noise components from noise-corrupted speech to obtain as clean the speech components as possible. As speech and noise are non-stationary, it is a big challenge to accurately estimate the noise level characteristics from a noisy speech signal. If the noise level in a noisy speech signal is over-estimated, the speech components might be over-attenuated through the enhancement process and hence causes distortions to the processed speech signals. On the other hand, if the noise level is under-estimated, the output speech might still sound

noisy after the enhancement process. Therefore, there is always a trade off between the removal of noise components and the retention of speech components. An optimal speech enhancement method should ideally attenuate the noise components as much as possible, while still keeping the underlying speech signals with as little distortion as possible. For decades, many speech enhancement methods have been developed [1-32] to improve the perceptual quality of the enhanced speech signals, or to enhance the speech signals so as to improve the performance of the ensuing stage of speech compression, recognition or authentication system, etc.

In this thesis, single-channel short time spectral amplitude estimation methods are reviewed. A robust single-channel speech enhancement technique, namely the β-order minimum mean-square error estimator incorporating masking properties [32] is examined in detail. Its strengths and weaknesses are identified and discussed. In a high background noise environment, a two-channel noise reduction technique is also studied [33]. The objective of our research is to analyze the shortcomings of the recently developed single-channel and two-channel speech enhancement systems and to develop new approaches to improve the objective as well as subjective performances of these systems. Various subjective and objective performance measurements will be used to ascertain the effectiveness of our proposed methods.

1.2 Major Contribution of the Thesis

This thesis examines various concepts and techniques of mainly frequency domain speech enhancement techniques. The well-known short-time spectral amplitude (STSA) estimation technique is first reviewed followed by the minimum mean-square

error (MMSE) estimator, and the β-order MMSE method that incorporates masking properties (β-masking).

To further enhance the performance of the β-masking method, two post processing approaches are proposed and their effectiveness is examined. The first post-processing technique is a non-linear high-frequency regeneration (NHR) approach to re-synthesize the upper-band signal using the lower-band spectral information. High-frequency regeneration technique is used because some of the high frequency spectral structures of the enhanced speech signal obtained from β-masking are distorted. The second post-processing technique involves the use of autocorrelation function. The autocorrelation of the spectrum is used to modify the β-masking filter gain and it results in the enhancement of the originally weak spectral components based on the information of the strong spectral components.

For the cases where the background noise level is high and time-varying, e.g., a mobile-communication application in a moving car interior, the performance curve of single-channel speech enhancement seems to plateau. Therefore, two-channel speech enhancement method is studied. One advantage of using a two-channel speech enhancement system over a single-channel system is that the cross-correlation between the two microphone inputs can be incorporated into the filter gains. A two-sensor noise reduction system for hands-free car kit is reviewed and the performance is examined in this thesis. The two-sensor approach is good in retaining the speech signal; however, the background noise of the output speech signals remains high in general. In contrast, under the same noise condition, the signals enhanced by the single-channel β-masking method have very low background noise, but with slight tonal distortion. Therefore, it is proposed

in this thesis that the β-masking technique be incorporated into the filter block of the two-sensor enhancement system so as to achieve a better enhancement performance.

As the original clean speech is not provided in most of the existing two-channel speech databases, it is impossible to perform objective quality assessments for the enhanced speech. Therefore, a three-channel speech recording experiment was set up to record noisy speech signals (two microphones) and clean speech signal (one microphone) under the same environment, which is in a moving vehicle. The recorded signals are then used in our simulation study.

Comparisons of the proposed techniques and the β-masking technique are performed. Besides the commonly used signal-to-noise ratio (SNR) measurement, a psycho-acoustically motivated distortion (PMD) measurement is also used as an objective measurement. Besides objective measurements, subjective comparisons are also performed. Both subjective and objective comparisons show that the proposed single-channel post-processing techniques and the hybrid two-channel speech enhancement technique outperform many existing speech enhancement methods.

1.3 Organization of the Thesis

This thesis consists of five chapters.

Chapter 1 introduces the motivation and objectives of the research project, major contributions made and organization of the thesis.

Chapter 2 presents a brief literature review and an overview of some commonly used frequency domain speech enhancement techniques. Various short-time spectral

amplitude estimation techniques are also outlined. In addition, the β-order MMSE method is studied in detail.

Chapter 3 presents two post-processing methods for the β-order MMSE enhancement system, namely the high-frequency regeneration method and the autocorrelation method. The effectiveness of these two proposed methods in improving the β-order MMSE enhanced speech is measured in terms of objective and subjective criteria.

Chapter 4 studies a two-sensor noise reduction technique for a hands-free car kit for a mobile communication system. A hybrid technique is proposed to combine the single channel speech enhancement technique with a two-channel configuration to form the final enhancement system. Performance comparisons were done between the two-sensor method, the single-channel β-order MMSE method and the hybrid system.

Chapter 5 provides a summary of the thesis with conclusions and recommendations for future research.

The Appendix provides the list and information of the sample speech files attached in the CD.

CHAPTER 2

REVIEW OF SPEECH ENHANCEMENT TECHNIQUES

In parallel with the development of many speech processing techniques, speech enhancement has been studied by many researchers for more than twenty years. Speech enhancement plays an important role in a communication system because it can improve speech quality to reduce listener's fatigue, and it can also improve speech intelligibility in a tele-conversation. The challenge of a speech enhancement process is to remove the corrupting noise components as much as possible in order to improve the quality of the processed speech. The performance of a speech enhancement system depends critically on the accuracy of the estimation of the noise and speech components given a noisy speech signal, and many research works in this area have been reported. In 1979, J. S. Lim [1] gave an overall review of various single-channel speech enhancement techniques and categorized them based on different approaches to estimate the noise and speech information.

Some well-known speech enhancement techniques are based on short-time spectral amplitude (STSA) estimation. There are two main reasons for using the STSA estimation technique. One reason is the non-stationarity of the speech characteristics, as speech signals are time-varying in nature. Therefore a speech signal is usually processed

in short-time windowed frames. The other reason is that the spectral amplitude is more important than phase for speech enhancement [1]. Some of the well-known techniques based on STSA estimations include: (1) direct estimation of STSA, such as power spectral subtraction (also known as correlation subtraction) originated by Weiss *et al.* [2] and modified by Boll, Lim, Berouti, and McAulay *et al.* [3-9]; and (2) Wiener filtering [10-14].

Some speech enhancement methods are based on the periodicity of voiced speech. The time domain speech waveform of a voiced speech signal is periodic, and the pitch period is the inverse of the fundamental frequency. Therefore, the following techniques were introduced using this feature: (1) adaptive comb filtering [15]; (2) harmonic selection [16]; and (3) adaptive noise cancelling techniques [17, 18].

Some speech enhancement techniques are based on a speech production model which could be used to re-produce speech. Voiced speech is excited by regular pulse trains and unvoiced speech is excited by random noise. When the parameters for the excitation sources are estimated, speech can be regenerated and used for speech enhancement. Some common techniques based on a speech model are: (1) all-pole model of speech [19-21]; (2) pole-zero model of speech [22-24]; and (3) nonparametric model of speech [25, 26].

In this thesis, some STSA estimation methods are revisited and studied. In 1984, Y. Ephraim and D. Malah derived a minimum mean-square error (MMSE) STSA estimator [27] and log spectral amplitude (LSA) estimator [28] for speech enhancement. The former is based on modeling speech and noise spectral components as statistically

independent Gaussian random variables. The estimated spectral amplitudes were derived directly from the noisy observations in order to make the estimation optimal. Later, C. H. You *et al* proposed the adaptive β-order MMSE speech enhancement algorithm [29-31] which incorporated masking properties ("β-masking" in short) [32].

In this chapter, time domain and frequency domain speech enhancement concepts are reviewed and the salient features of some speech enhancement systems are described.

2.1. Time Domain and Frequency Domain Speech Enhancement methods

Speech enhancement procedures can be performed in either the time domain or the frequency domain. In the time domain, when a stationary random signal *s(t)* is degraded by uncorrelated additive noise *n(t),* the noisy speech *x(t)* can be expressed as:

$$x(t) = s(t) + n(t). \tag{2.1}$$

There are various types of noise which might include the following:

1) Babble noise. Many people talking in pubic produces babble noise. Thus individual voices are audible to a certain extent. This will cause communication interference.

2) Vehicle interior noise. Inside a car, the undesirable noise might come from the car engine, wind, rain, the contact of wheels with road, and the noise from other cars passing by.

3) F16 Noise. It is often used by researchers as a typical fighter aircraft noise. Like other noises produced by a running engine, the noise spectrum produced by a flying F-16 has a strong peak at the frequency at which the engine operates.

4) Pink noise. Pink noise is also known as "1/f noise" because the power spectral density is proportional to the reciprocal of the frequency. This is equivalent to say that the power density will drop by 3dB per octave and the energy is equal in all octaves. At high frequencies, pink noise will not be dominant. Pink noise patterns have been found in music melodies, semiconductors and atomic clocks. They are also found in nature, including the sounds of wind and waterfalls.

5) White Noise. White noise exhibits equal energy per Hz. A white noise signal is random and all of the frequency components have an average uniform power level. In this thesis, white noise is used as an additive noise to generate noisy speech signals from clean speech signals.

Some commonly used time domain speech enhancement techniques are comb filtering, Kalman filtering, Hidden Markov Models (HMM), neural networks, etc. For comb filtering process, an FIR filter is used with coefficients determined by the pitch period. For Kalman filtering process, speech and noise production models are used whose AR parameters are estimated from the noisy speech data. In an HMM-based enhancement scheme, hidden Markov models are used to model the statistics of speech and noise, based on state-dependent probability distributions (PD). These composite source models provide the estimated speech information for further noise suppression. In a neural network system, multi-layered perceptrons are used and they function like a non-linear

filter to suppress noise. The weights of the perceptrons are trained using the error back propagation algorithm.

If the additive model is transformed to the frequency domain, it is also additive. For a series of N random samples $x(n)$, $n=0, 1, 2,...N-1$, the Discrete Fourier Transform (DFT) is defined as:

$$X(k) = \sum_{n=0}^{N-1} x(n) \cdot e^{-j\frac{2\pi}{N}kn}, \; k = 0, 1, 2, \ldots, N\text{-}1 \tag{2.2}$$

and the Inverse Discrete Fourier Transform (IDFT) is given by:

$$x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) \cdot e^{j\frac{2\pi}{N}kn}, \; n = 0, 1, 2, \ldots, N\text{-}1 \tag{2.3}$$

It is widely accepted in most speech enhancement research papers that it is easier to estimate the spectral amplitude of a speech signal than to estimate both its amplitude and phase [1]. In fact, most of the frequency domain speech enhancement techniques only modify the spectral amplitudes, leaving the phase untouched. In the frequency domain, the noisy signal can be expressed as:

$$X(k) = S(k) + N(k) \tag{2.4}$$

where $S(k)$, $X(k)$ and $N(k)$ denote the k-th spectral components of the clean speech, noisy speech and noise signals, respectively. Let A_k, R_k, and N_k denote the respective amplitudes, the above equation is then equivalent to:

$$R_k \cdot \exp(j\vartheta_k) = A_k \cdot \exp(j\alpha_k) + N_k \cdot \exp(j\varphi_k). \tag{2.5}$$

The overall speech enhancement procedure for a transform-based technique (DFT is used here) is shown in Figure 2.1. Firstly, the noisy speech signal in the time domain $x(t)$ is windowed and the Hanning function or the Hamming function is commonly

applied for this purpose. After windowing, the signal is partitioned into frames. Each frame of samples is transformed into the frequency domain to obtain the spectral amplitude R_k and phase ϑ_k. R_k is then filtered to obtain the estimated clean spectral amplitude $\hat{A}(k)$. With the phase ϑ_k unchanged, the amplitude and phase are inverse-transformed to obtain the time domain signal. After overlapping and adding, we finally obtain the enhanced speech signal $\hat{s}(t)$. The filtering process is the major research topic for the frequency domain speech enhancement approach for many years. Many methods are well developed and some of them are regarded as the milestones in this area; for example the spectral subtraction technique, Wiener filtering and MMSE estimator, etc, are often used as a performance benchmark for new algorithms.

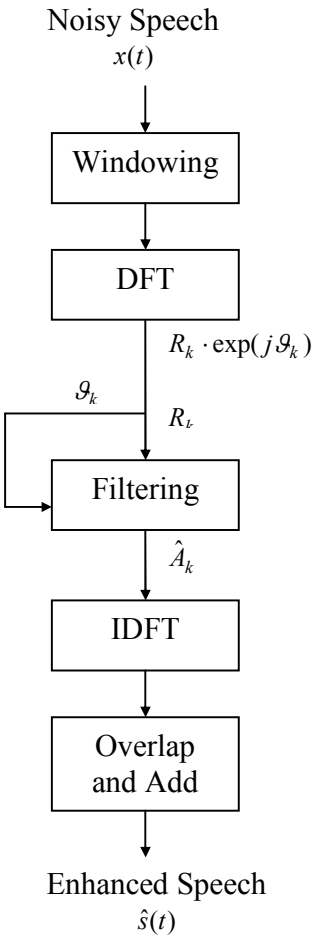

Figure 2.1: Block diagram of a frequency domain speech enhancement process

2.2. Frequency Domain Speech Enhancement Methods -- The Short-Time Spectral Amplitude (STSA) Estimator

As natural speech signals are non-stationary, the STSA concept is often used. Therefore the time domain speech signals are sectioned and windowed into small time frames with a limited sample size. The signal within a small time frame is relatively stationary. Next they are transformed into frequency spectra. The usual window size used is 256 (for 8 KHz sampling frequency) or 512 (for 16 KHz sampling frequency) samples with an overlapping ratio of 0.5 or 0.75. The STSA estimation method has been studied for many decades, and some very well-known algorithms will be introduced in this section; they range from the very simple spectral subtraction method to the more sophisticated yet robust β-order MMSE method.

2.2.1 Spectral Subtraction

The magnitude spectral subtraction method introduced by S.F. Boll [3] is a very simple speech enhancement method. It assumes that the noise signal is additive and stationary within a small time interval and it varies very slowly. Given the spectrum of the noisy speech $X(k) = S(k) + N(k)$, we have the magnitude of the estimated speech spectrum given as follows:

$$\left|\hat{S}(k)\right| = \left|X(k)\right| - E\{\left|N\right|\}$$

$$\left|\hat{S}(k)\right| = \left|X(k)\right| \cdot \left(1 - \frac{E\{\left|N\right|\}}{\left|X(k)\right|}\right)$$

where $E\{\cdot\}$ denotes the expectation operator. The suppression filter gain $G(k)$ for the simple spectral subtraction method can be obtained as:

$$\left|\hat{S}(k)\right| = G(k)\cdot\left|X(k)\right|,\ G(k) = 1 - \frac{E\{|N|\}}{|X(k)|} \tag{2.6}$$

To avoid abrupt fluctuation in the output speech, a three-frame averaging technique is used to smoothen $G(k)$. This simple filtering process invariably leads to a high level of residual noise in the enhanced speech. Furthermore, the residual noise manifests itself in the form of musical tones.

Subsequent to the magnitude spectral subtraction method, a spectral power subtraction technique was proposed by Scalart [5] which makes use of the power level instead of the magnitude of the speech spectrum, i.e.,

$$|X(k)|^2 = |S(k)|^2 + |N(k)|^2 \tag{2.7}$$

The filter gain $G(k)$ can be shown to be given by:

$$G(k) = \sqrt{1 - \frac{E\{|N|^2\}}{|X(k)|^2}} \tag{2.8}$$

The power spectral subtraction method has been found to perform better with less musical tones in the enhanced speech as compared to the spectral subtraction approach. However, the overall residual noise level is still high.

2.2.2 Wiener Filtering

Following the development of the simple and "direct" spectral subtraction method, more advanced and effective filtering techniques were proposed. From

$X(k) = S(k) + N(k)$, the non-causal Wiener filter with frequency response of *G(k)* was proposed, i.e.,

$$\hat{S}(k) = G(k) \cdot X(k), \text{ where } G(k) = \frac{|S(k)|^2}{|S(k)|^2 + |N(k)|^2} \tag{2.9}$$

As the clean speech spectrum *S(k)* and the noise spectrum *N(k)* are non-stationary and not accurately predictable, an approximation to the Wiener filter is given below:

$$G(k) = \left[\frac{E\{|S(k)|^2\}}{E\{|S(k)|^2\} + \alpha \cdot E\{|N(k)|^2\}} \right]^{\beta} \tag{2.10}$$

where α and β are some parametric constants [1]. It had been shown that Wiener filter could reduce the background noise to a remarkably low level.

2.2.3 Minimum Mean-Square Error (MMSE) STSA Estimator

In 1984, Y. Ephraim and D. Malah derived a minimum mean-square error (MMSE) STSA estimator, based on modeling speech and noise spectral components as statistically independent Gaussian random variables [27]. The estimator was based on minimizing the mean-square error between the enhanced speech and the clean speech. Both noise and speech are modeled using the complex Gaussian model.

Given the spectrum of a noisy speech signal $X(k) = S(k) + N(k)$, where *S(k)*, *X(k)* and *N(k)* denote the *k*-th spectral components of the clean speech, noisy speech and noise, respectively, the distortion measure between the clean and estimated speech is the mean-square error defined as:

$$J_{MMSE} = E\{(A_k - \hat{A}_k)^2\} \tag{2.11}$$

where A_k and $\hat{A}_k$ are the amplitudes of the *k*-th spectral components of the clean speech and estimated speech, i.e., $A_k = |S(k)|$, $\hat{A}_k = |\hat{S}(k)|$. By minimizing the distortion measure J_{MMSE}, which is to take the first-order derivative of J_{MMSE} and set to zero, the estimator is given by:

$$\hat{A}_k = E\{A_k \mid x(t), \quad 0 \le t \le T\} \tag{2.12}$$

Assuming Gaussian model is used, the estimator can be shown to be given as follows:

$$\hat{A}_k = E\{A_k \mid X_k\} = \frac{\int_0^{\infty}\int_0^{2\pi} a_k\, p(X_k \mid a_k, \alpha_k) d\alpha_k da_k}{\int_0^{\infty}\int_0^{2\pi} p(X_k \mid a_k, \alpha_k) d\alpha_k da_k} \tag{2.13}$$

where a_k is a possible value of the amplitude, α_k is a possible value of the phase, and $p(\cdot)$ is the probability density function (PDF). The conditional PDFs are given by:

$$p(X_k \mid a_k, \alpha_k) = \frac{1}{\pi \cdot E\{{N_k}^2\}} \cdot \exp\left(-\frac{|X_k - a_k \cdot \exp(j\alpha_k)|^2}{E\{{N_k}^2\}} \right) \tag{2.14}$$

$$p(a_k, \alpha_k) = \frac{a_k}{\pi \cdot E\{{A_k}^2\}} \cdot \exp\left(-\frac{{a_k}^2}{E\{{A_k}^2\}} \right) \tag{2.15}$$

The multiplicative nonlinear gain function can then be derived as:

$$G_{MMSE}(\xi_k, \gamma_k) = \Gamma(1.5) \frac{\sqrt{v_k}}{\gamma_k} M(-0.5;1;-v_k) \tag{2.16}$$

where $\Gamma(\cdot)$ denotes the gamma function defined as:

$$\begin{aligned} \Gamma(x) &= (x-1)! \\ &= (x-1) \cdot \Gamma(x-1) \\ &= \int_0^{\infty} t^{x-1} e^{-t} dt \end{aligned} \tag{2.17}$$

and *M(a;c;x)* is the confluent hypergeometric function [34]:

$$M(a,c,x) = \sum_{r=0}^{\infty} \frac{(a)_r}{(c)_r} \frac{x^r}{r!} = 1 + \frac{a}{c}x + \frac{a(a+1)}{c(c+1)}\frac{x^2}{2!} + \frac{a(a+1)(a+2)}{c(c+1)(c+2)}\frac{x^3}{3!} + \ldots \tag{2.18}$$

v_k is defined by:

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k \tag{2.19}$$

where ξ_k and γ_k are interpreted as the *a priori* and *a posteriori* SNR, respectively:

$$\xi_k = \frac{E\{A_k{}^2\}}{E\{N_k{}^2\}}, \quad \gamma_k = \frac{R_k{}^2}{E\{N_k{}^2\}} \tag{2.20}$$

and A_k, R_k and N_k stand for the amplitudes of the k-th spectral components of the clean speech *S(k)*, noisy speech *X(k)* and noise *N(k)*, respectively. The authors derived the MMSE STSA estimator, which takes into account the uncertainty of signal presence in the noisy spectral components. When the noisy speech level is well above the noise level, the estimation of *a priori* SNR is more dynamic so as to accurately adapt the filter gain to the non-stationary noisy speech signal. When the noisy speech level is low, the estimation of *a priori* SNR is smoothed so as to reduce the musical residue of the output speech signals.

Besides MMSE, Ephraim also proposed the Log MMSE estimator, where the log of the mean-square error as distortion measurement is used [28]. The authors showed that the MMSE STSA estimator performs better than other known estimators (Wiener estimator, etc.) by achieving low MSE and reducing musical tones to a certain extent [27, 28].

2.2.4 Adaptive β-Order MMSE STSA Estimator Incorporating Masking Properties

As a generalization, C. H. You *et al* derived the adaptive β-order MMSE STSA estimator based on Ephraim and Malah's MMSE STSA estimator. Instead of evaluating the mean-square error of the spectral amplitude, the mean-square error of the β-order spectral amplitude is used as the distortion measure [29], i.e.,

$$J_\beta = E\{(A_k^\beta - \hat{A}_k^\beta)^2\} \tag{2.21}$$

Again, by minimizing J_β, the estimation of the speech spectral component is expressed as:

$$\hat{A}_k = [E\{A_k^\beta \mid X_k\}]^{1/\beta} = G_\beta(\xi_k, \gamma_k) \cdot |X_k| \tag{2.22}$$

The estimation of $E\{A_k^\beta \mid X_k\}$ is based on the conditional power density functions (PDFs) of the observed signal. Finally, the gain function G_β was derived for different values of β as follows:

$$G_\beta(\xi_k, \gamma_k) = \frac{\sqrt{v_k}}{\gamma_k}\left[\Gamma\left(\frac{\beta}{2}+1\right) M\left(-\frac{\beta}{2};1;-v_k\right)\right]^{1/\beta} \tag{2.23}$$

From the expression of G_β, it is apparent that Ephraim and Malah's estimator, G_{MMSE}, is a special case of β=1. Here $\Gamma(\cdot)$ denotes the gamma function, *M(a;c;x)* is the confluent hypergeometric function as in the previous section, and v_k ξ_k and γ_k also have the same interpretation. Figure 2.2 shows the gain curves of the β-order MMSE estimator, G_β, in comparison with the Wiener estimator, as a function of ξ_k and γ_k.

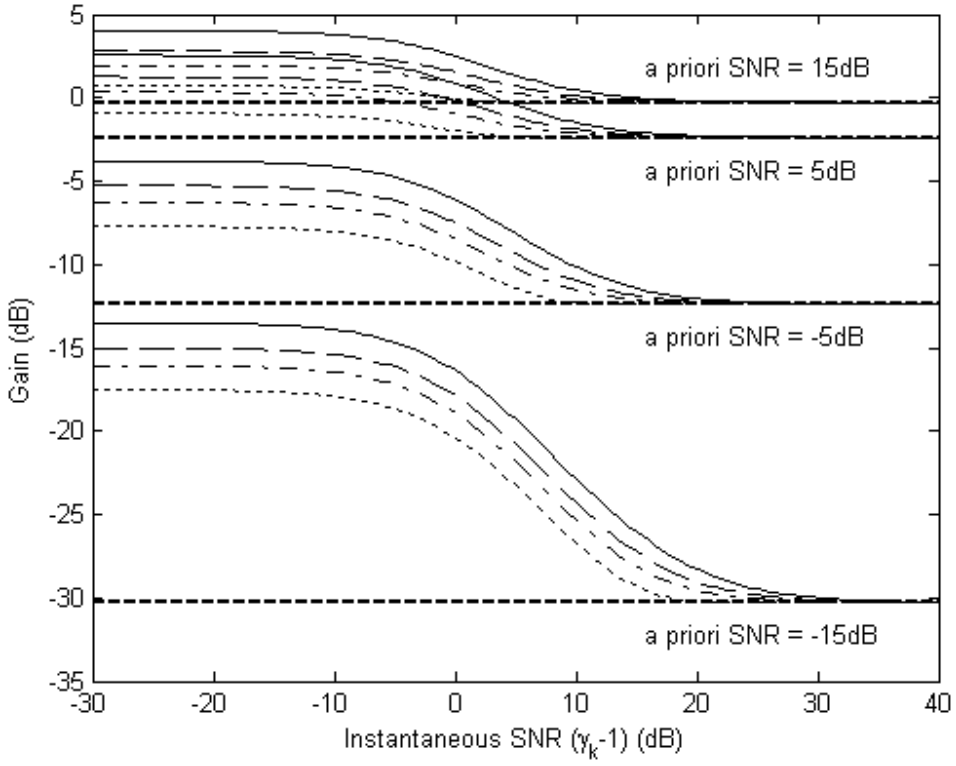

Figure 2.2: Gain Vs instantaneous SNR (γ_k-1) in comparison with the Wiener gain (bold-dashed) for β=0.01 (dotted), 1.0 (dot-dashed), 2.0 (dashed) and 4.0 (solid); and a-priori SNR ξk=-15, -5, 5, 15dB. [29]

The β-order MMSE algorithm makes the gain G_β adaptive by adjusting β to a proper value in order not to over-attenuate the weak spectral components. G_β increases with the *a-priori* SNR ξ_k because the presence of speech is expected. For a given *a-priori* SNR ξ_k, G_β increases as the instantaneous SNR (γ_k -1) decreases. This is done so that weak spectral components with low SNR will be appropriately enhanced. Compared to the Wiener gain (bold-dash), G_β is better for recovery of weak speech spectral components for low SNR (γ_k - 1), and it converges to the Wiener gain at high SNR (γ_k -1), which means that the gains are nearly the same for both schemes for strong speech

spectral components. It should be noted that when β is closed to 0, the gain curve is closed to the E-M LSA gain curve [28]; when β=1 (the dot-dashed curve), the gain is exactly the same as E-M STSA-MMSE gain [27].

For further improvement in performance, the auditory masking properties are also incorporated into the gain [35] which is named β-masking gain in short. The β-masking gain G_β applies only when the *a posteriori* SNR γ_k is above the masking threshold. When γ_k is below the masking threshold, the particular spectral component (regardless of noise or speech) is unlikely to be audible. Therefore, the gain is re-adjusted so as to reduce the spectral component to a very low level. Thus the overall gain $G_{\beta p}$ becomes:

$$G_{\beta p}(\xi_k,\gamma_k)=\begin{cases}\dfrac{\sqrt{v_k}}{\gamma_k}\left[\Gamma\left(\dfrac{\beta}{2}+1\right)M\left(-\dfrac{\beta}{2};1;-v_k\right)\right]^{1/\beta}, & if \quad \gamma_k > \rho_1(k)+\rho_2(k)\\ \sqrt{\rho_2(k)/\gamma_k}, & otherwise\end{cases} \tag{2.24}$$

Here $\rho_1(k)$ and $\rho_2(k)$ are determined by the noise masking threshold, which is obtained through modeling the frequency selectivity of the human auditory system and its masking properties [35, 36]. It is estimated by critical Bark band analysis, a spreading function, the noise masking threshold and the absolute auditory threshold.

The block diagram of the β-masking enhancement method is depicted in Figure 2.3. The β-masking method estimates the noise level throughout all the time frames of the entire speech utterance. C. H. You *et al* used a fluctuation-tracking minima-search (FTMS) noise estimation method. When the pre-SNR of a particular frame is below a certain threshold, its value is copied into the periodogram bin, from which the noise information is obtained. As the method analyses noise frame by frame, the first few

frames of the output speech may be noisy because the noise information has not been fully established.

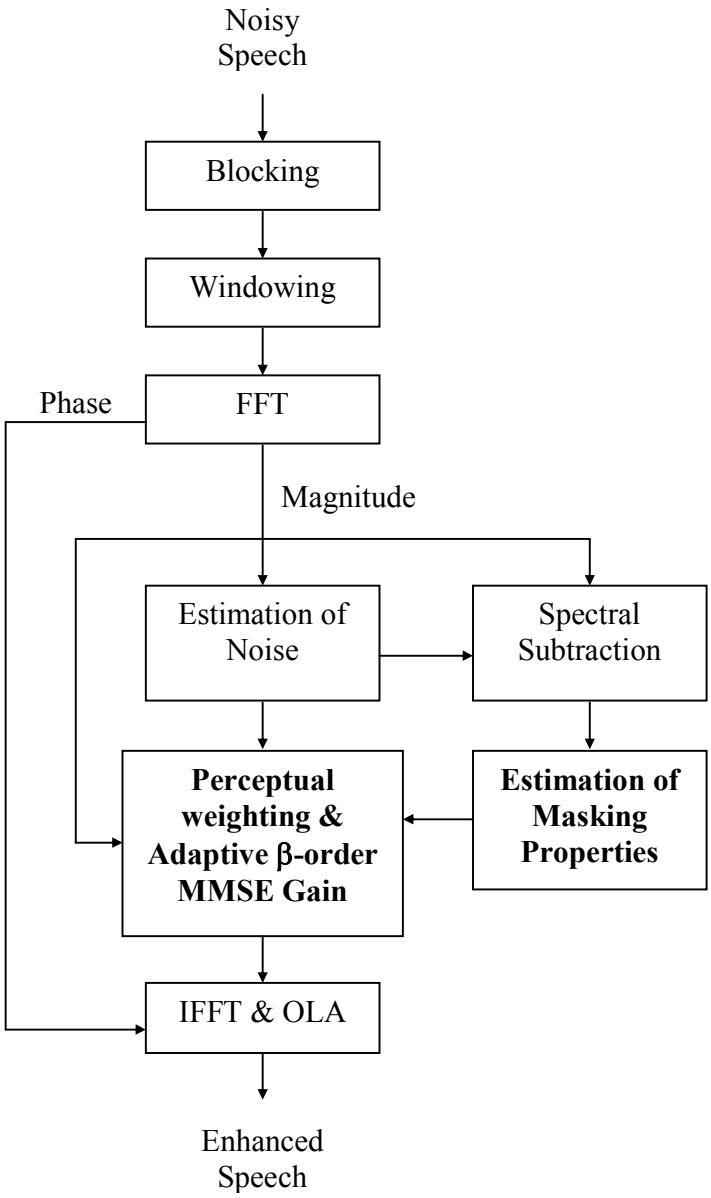

Figure 2.3: Structure of masking-based adaptive β-order MMSE Method

In [32], the authors showed that β-masking method always outperforms the traditional methods such as Spectral Subtraction, E-M LSA, E-M STSA-MMSE and OM-LSA. In terms of objective measurement, the speech signals enhanced by the β-masking method have an average improvement of more than 1dB on segmental SNR as compared to the other methods mentioned earlier. In terms of subjective evaluation, the β-masking method has a much higher Mean Opinion Score (MOS).

As the β-masking method for speech enhancement was assessed in terms of objective quality measures such as segmental SNR improvement and informal listening tests, an intelligibility test was performed in our study to evaluate the recognition rate of speech signals enhanced by the method. There are eight groups of word pairs with different groups of words spoken by different persons. Each group consists of 63x2 pairs of words. For example, one word pair is “bank-dank” or “dank-bank”, and there are 63 noisy word pairs, and 63 pairs which were enhanced by the β-masking method. Listeners were asked to select one or more random groups and listen to the 63x2 pairs of word. Next, they were asked to decide and select whether what they heard was for example, “bank-dank”, “dank-bank” or “not distinguishable”. The test sequences were randomized.

The interface of the word-pair intelligibility test is as shown in Figure 2.4. In the figure, the bold “Noisy” indicates noisy word pairs and the bold “Clean” indicates enhanced word pairs. Besides removing background noise, it is also important for a good speech enhancement method to retain the intelligibility of speech. For most cases, vowels (voiced speech) can be easily identified; however, consonants (mostly unvoiced speech)

could be easily masked by noise and are difficult to be recovered. In this test, each pair of words have the same vowel but different consonants (beginning consonants or terminating consonants).

The result of the intelligibility test is as shown in Table 2.1. There are two word pair recognition accuracies for each group. One is the recognition accuracy of the noisy speech and the other is that of the β-masking enhanced speech. It is clear that the β-masking method can effectively increase the word pair recognition accuracy. The high recognition accuracy shows the effectiveness of this speech enhancement approach, in addition to improving the perceptual quality of the enhanced speech.

Group	Word pair recognition accuracies (Noisy)	Word pair recognition accuracies (β-masking)
F1	90.48	95.77
F2	90.87	96.83
F3	82.54	96.83
F4	89.42	95.24
M1	96.03	98.41
M2	91.27	95.24
M3	92.06	95.24
M4	92.06	95.24

Table 2.1: Word pair recognition accuracies for noisy and β-masking enhanced speech

Figure 2.4: The word-pair intelligibility test

2.2.5 Post processing methods of β-masking-enhanced speech signals

Past simulation results show that β-masking is very good in background noise reduction, giving a quiet output. However, if the original background noise is too high, some weak spectral components, mostly in high frequency bands, are completely masked by noise, as shown in Figure 2.5. These components are unlikely to be recovered by conventional speech enhancement methods. When the background noise is high, it is very likely that some perceptually important speech spectral components are over-attenuated because they are very difficult to be distinguished from the noise components. For example, after noise removal using the β-masking method, they are over-attenuated, as shown in Figure 2.6. For some instances, high frequency musical tone was observed in the β-masking enhanced speech signals, which causes perceptual distortions in the enhanced speech signals, and it will also affect the objective measurement results.

For some low SNR speech signals, harmonic regeneration [37] is necessary to artificially synthesize the lost information by exploiting the preserved spectral information and using some assumptions. In order to reconstruct the high frequency spectrum, two post-processing methods are proposed, namely, the non-linear high-frequency regeneration method and the cepstrum-aided autocorrelation method. These two methods will be discussed in Chapter 3.

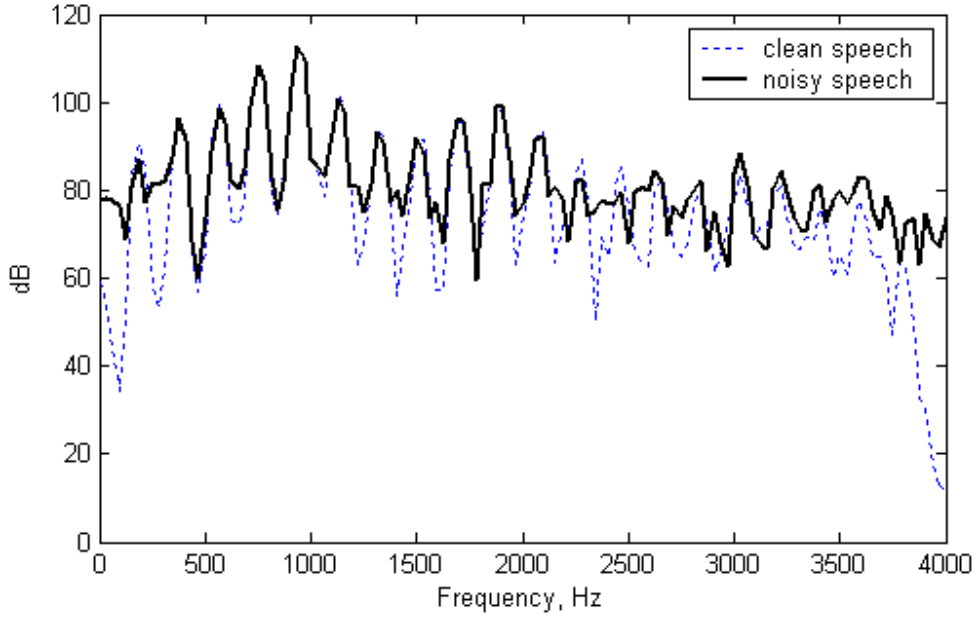

Figure 2. 5: Frequency components in noisy speech signals with some components completely masked by noise

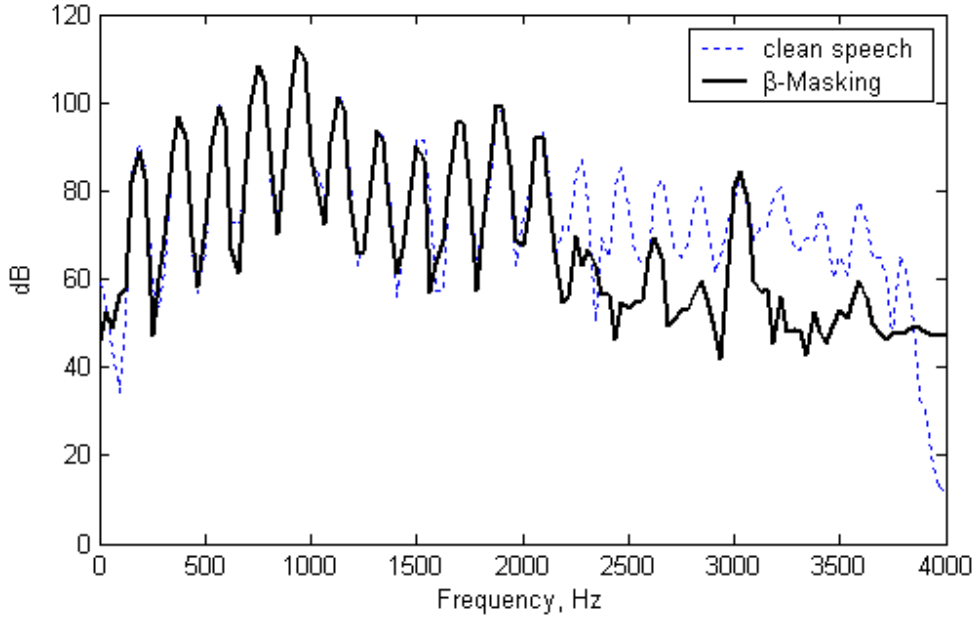

Figure 2. 6: Frequency components in β-masking enhanced speech signals with some components unable to be recovered

2.3 Conclusions

In this chapter, some commonly used frequency domain STSA estimation techniques are studied. Spectral/power subtraction is a very simple and direct STSA estimation technique. Wiener filter is a very well-known frequency-weighted filter used for enhancement given a noisy speech signal. MMSE is an optimum filtering technique where statistical analysis is used for minimizing the MSE distortion measure. A recent very robust single channel speech enhancement algorithm, β-masking, is also reviewed. This algorithm derives the MMSE estimator using adaptive β-order cost function, with the incorporation of auditory masking effects. The β-masking method outperforms the traditional STSA methods and the value of β is adapted empirically. Although β-masking method is good at noise removal, some spectral components are over-attenuated in the enhanced speech signals when the background noise level is high in the noisy speech signals. Therefore, two post processing methods are proposed in the next chapter in an attempt to recover the weak spectral components swamped by the noise signals.

CHAPTER 3

POST PROCESSING OF β-MASKING ENHANCED SPEECH

In order to overcome some of the weaknesses of the β-masking method, two post-processing algorithms for single-channel speech enhancement techniques are proposed to further reduce the distortions in the β-masking-enhanced speech signals: a non-linear high-frequency regeneration method and a cepstrum-aided autocorrelation method. Simulation results show that these two post-processing methods are able to re-synthesize the over-attenuated spectral components, by exploiting the speech formant and pitch information.

3.1 Non-linear high-frequency regeneration (NHR)

The idea of non-linear high-frequency regeneration technique originates from the narrowband to wideband (NTW) transformation technique proposed by I. Y. Soon *et al* [38]. In the original NTW process, 4 kHz narrowband speech signals are transformed into 8 kHz wideband speech signals and the transformation is done in the time domain. Firstly, the 4 kHz narrowband signal $x_{4k}(t)$ is up-sampled and low-pass filtered (LPF) to obtain an 8 kHz wideband signal $x_{8k}(t)$. Although this wideband signal has a bandwidth of 8 kHz, the energy resides only below 4 kHz. This wideband signal is passed through a full wave

rectifier to obtain an initial estimate of the upper band signal $y_{8k}(t)$, which is then high-pass filtered (HPF) and conditionally added to the wideband signal $x_{8k}(t)$ to obtain a wideband signal. The amount of addition depends on the type of speech. In unvoiced speech frames, the energy distributed in the upper-band is relatively higher than in voiced frames and thus the amount of addition is higher. Zero crossings rate (ZCR) is used for voiced/unvoiced detection, as unvoiced speech signals usually have a higher ZCR and voiced speech signals almost always have a lower ZCR. The original NTW transformation process is carried out in the time domain.

In this thesis, a post-processing NHR stage makes use of the similar idea as that proposed in [38] to transform a 2 kHz signal into a 4 kHz signal. As discussed in Chapter 2, for a given noise corrupted speech signal, weak spectral components which are completely swamped by noise are very difficult to recover. The use of adaptive β value avoids over-attenuation of the weak spectral components to a certain extent. However, when the parameters controlling the gain function are not sufficiently accurately estimated, the weak spectral components could be suppressed to a very low level. Moreover, the weak speech spectral components are mixed with noise, thus making it difficult to recover the original harmonic structure. It is observed that for voiced frames, over-attenuation usually occurs in the high frequency bands, where the speech spectral components are relatively weak. Instead of recovering the original harmonic structure, slight tonal distortions are observed in the β-masking enhanced speech signals. In order to improve the spectral quality of the enhanced speech, the NHR post-processing method is used. For β-masking enhanced speech signals with a bandwidth of 4 kHz, the NHR technique is used to synthesize the upper-band (2 to 4 kHz) information based on the

lower-band (0 to 2 kHz) information. As the β-masking technique is a frequency domain process, the proposed NHR method is also performed in the frequency domain for easy implementation.

The block diagram of the NHR process is as shown in Figure 3.1. The β-masking enhanced speech, $x(t)$, has an effective bandwidth of 4 kHz (i.e. 8 kHz sampling rate). The spectrum is denoted by $X(k)$. For each frame of speech, voiced/unvoiced/silence decision is made. The voiced/unvoiced/silence decision is a function of the estimated SNR for the k-th spectral component, and it is determined by $X(k)$. Firstly, noise level is estimated from the spectra of the first 5-6 frames, and this noise information is stored in a buffer bin; during the enhancement process, the buffer bin keeps updating the noise information. Next, segmental SNR is estimated to decide whether the current frame contains noise or speech. If an unvoiced or a silence frame is expected, no NHR operation will be applied. If a voiced frame is detected, the NHR technique will be applied.

For voiced frames, $X(k)$ is firstly low-pass-filtered with the 3dB cutoff frequency at 2 kHz to remove the residual tonal distortions in the high-frequency band. The output spectrum is noted as $Y_1(k)$ in Figure 3.1. $Y_1(k)$ has an effective bandwidth of 2 kHz, and its corresponding time domain signal is denoted as $y_1(t)$. Next, the bandwidth of $y_1(t)$ is doubled by using a full wave rectifier (absolute function in the time domain). We thus obtain a time domain signal $y_2(t)$, which gives an initial estimate of the upper-band (2-4 kHz) signal. In order to avoid overlapping with the lower-band signal, $y_2(t)$ is then high-pass filtered in the frequency domain to obtain the upper-band spectrum $Y_3(k)$. The lower-band spectral components of $Y_3(k)$ are the same as those of $X(k)$, and the upper-band

spectral components are initially estimated. We can visualize this by imagining that the upper-band information is the image of the lower-band, and they are symmetrical about the 2 kHz line. The high-pass and low-pass filters correspond to linear-phase FIR filters with 21 coefficients each and 3dB points at 2 kHz. In the frequency domain, the spectral amplitudes of the filters can be obtained from these coefficients. The lower-band spectrum $Y_1(k)$ and the upper-band spectrum $Y_3(k)$ are then added, resulting in the synthesized spectrum $Y_4(k)$. Next, in order to shape the spectral energy, the refined spectrum $Y_5(k)$ is obtained by normalizing the energy level of $Y_4(k)$ to that of the general speech spectral envelope. $Y_5(k)$ thus inherits the harmonic structure of $Y_4(k)$, while maintaining the spectral envelope (or the energy level) of $X(k)$, i.e.,

$$Y_5(k) = Y_4(k)\frac{envlp(X(k))}{envlp(Y_4(k))} \tag{3.1}$$

Here the "*envlp* function" is as shown in Figure 3.2. The expression of Eqn (3.1) implies that $Y_5(k)$ inherits the quasi-periodicity of $Y_4(k)$, while keeping the spectral envelope of $X(k)$.

In order to examine any possibly negative effects introduced by the NHR scheme, a weighting factor is used. Thus the final output spectrum is given by:

$$Y_{out}(k) = \alpha_{NHR} \cdot Y_5(k) + (1 - \alpha_{NHR}) \cdot X(k) \tag{3.2}$$

where α_{NHR} is the weighting factor ranging from 0 to 1. In order to obtain the optimized value for α_{NHR}, eight speech sentences degraded by white Gaussian noise were processed by the β-masking-NHR method using different weighting factors, and the average psycho-acoustically motivated distortion (PMD) measurements are plotted in Figure 3.3. (The details of the PMD measurement will be introduced in section 3.3.5.). It is found

that the NHR post-processing method can reduce the distortion compared to β-masking method alone (where $\alpha_{NHR} = 0$). The optimum value for α_{NHR} is 0.8 as it gives the minimum PMD value. Therefore, $\alpha_{NHR} = 0.8$ is used for performance evaluation for the NHR algorithm in Chapter 4.

In summary, the NHR algorithm attempts to re-synthesize the upper-band by repeating the lower-band information, assuming that the frequency content of the voiced speech is perfectly periodic. Therefore, by repeating this periodic information, and restricting the magnitude of reconstructed spectrum to the general spectral shape, we can successfully remove the noise-like upper-band spectrum and the residual musical tones.

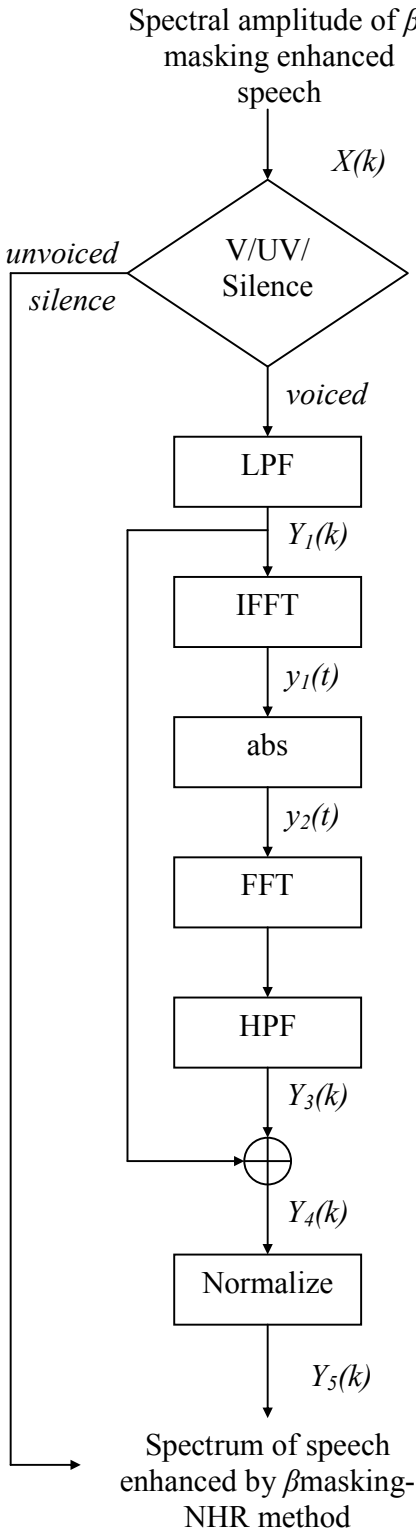

Figure 3.1: Block diagram of the NHR method

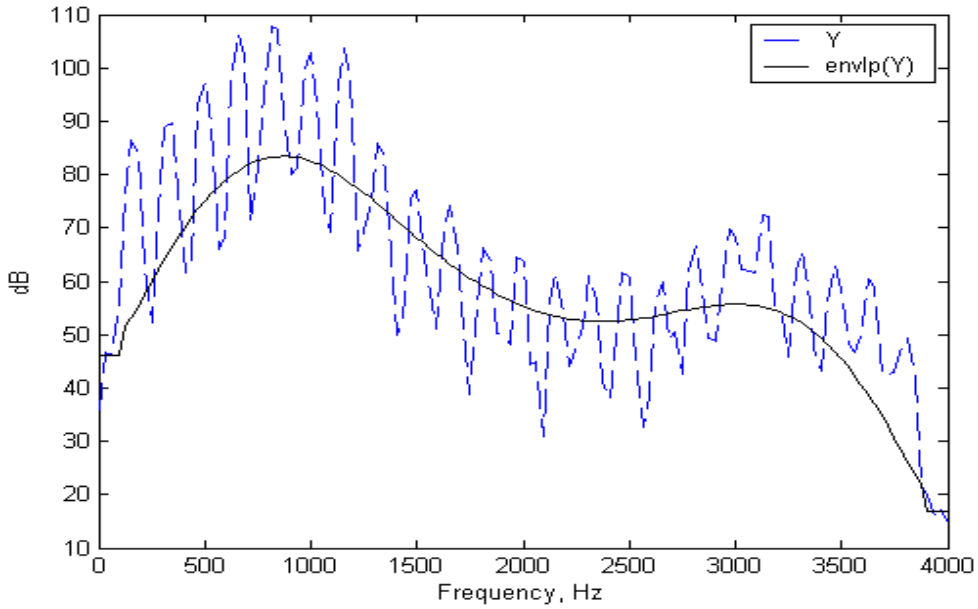

Figure 3.2: An envelope function

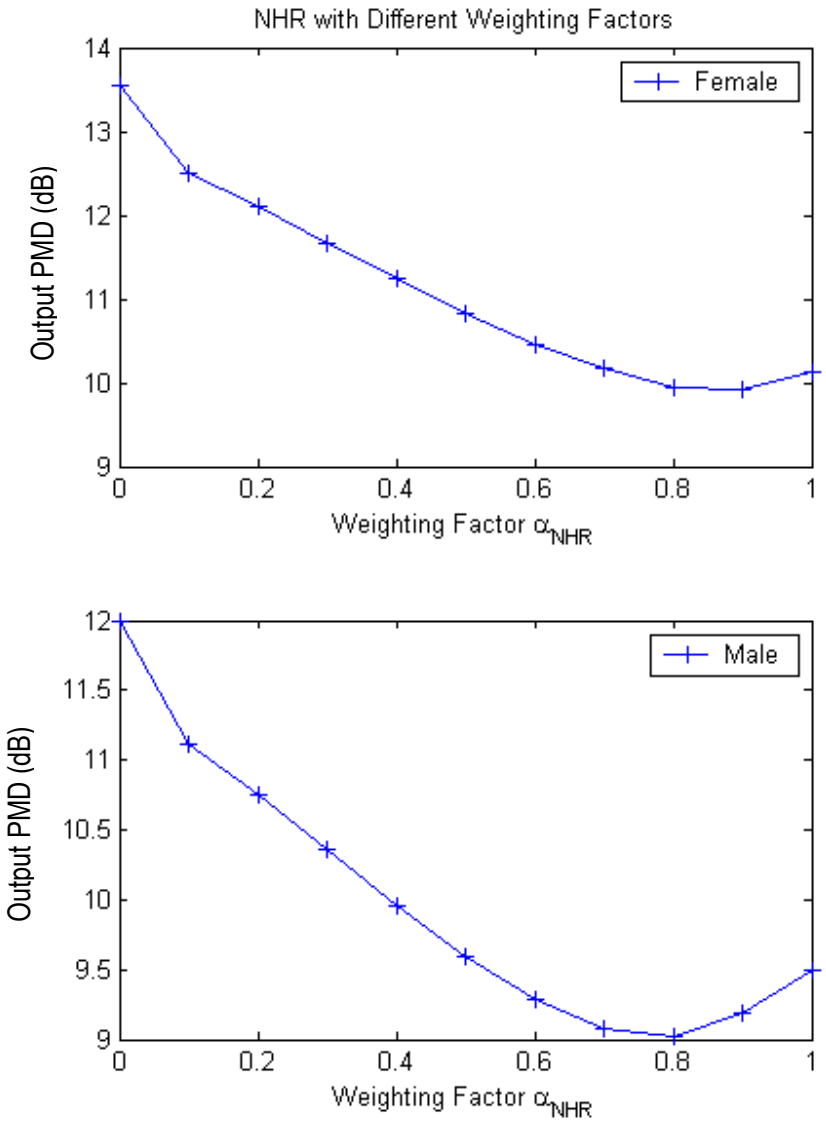

Figure 3.3: PMD measurements for different NHR weighting factors

3.2. Cepstrum-based autocorrelation (Corr)

Speech distortion observed in the β-masking enhanced speech signals is mostly due to the over attenuation of the weak spectral components, which mainly occurs in the high frequency band. Besides the non-linear high-frequency regeneration proposed in the previous section, another possible approach is the regeneration of the over-attenuated weak spectral components by using the autocorrelations of the strong spectral components. Before the re-synthesis method is proposed, the autocorrelation properties of the voiced speech signals will be examined first.

3.2.1 Autocorrelation properties of the voiced speech signals

Speech is the result of the excitation of the vocal tract by a quasi-periodic glottal pulse train and/or frication caused by a constriction of the airflow in the front cavity of the vocal tract. The signal components, which are the result of excitation by glottal pulses, are called voiced, and they are pitch-related, since pitch is the frequency of the glottal pulses.

In the frequency domain, for voiced speech signals, the structure can be interpreted as the product of the pitch spectrum and the formant spectrum. The pitch spectrum consists of pitch bars, which are spectral peaks at the pitch harmonics. The formant spectrum consists of several fairly broad spectral peaks. The resulting spectral structure of voiced speech is a formant spectrum, which is "sampled" at the pitch harmonics. Figure 3.4 shows the spectral components of a frame of voiced speech signal, which is normalized to have maximum amplitude of unity. The black dotted curve in

Figure 3.4 shows the normalized frequency response of the inverse LPC filter generated using the clean speech signal, and this curve shows the formant structure of this frame of speech signal.

During the β-masking speech enhancement process, when the noise level is so high that some spectral components are totally swamped by noise, some of the weak spectral components are not recovered. Figure 3.5 shows the spectral components of the same frame of voiced speech signal degraded by additive white noise and then enhanced by the β-masking method. As compared to the clean spectral components (Figure 3.4), most of the components of the processed speech (Figure 3.5) are suppressed to a very low level. The corresponding β-masking gain function G_β that produces this output is plotted in Figure 3.6. The over-attenuation of the output spectrum is due to the small value of the G_β, which has lost its comb-structure in the high frequency band, and only a small portion of the spectrum retains the periodicity. The autocorrelation of the spectrum may help to obtain the periodicity information for the voiced spectrum [39, 40]. Knowing the periodicity, the weak spectral components can be artificially yet reasonably well re-synthesized. The re-synthesized spectrum can be used to obtain a new β-masking gain $G_{\beta C}$, which helps to reconstruct the spectral comb-structure.

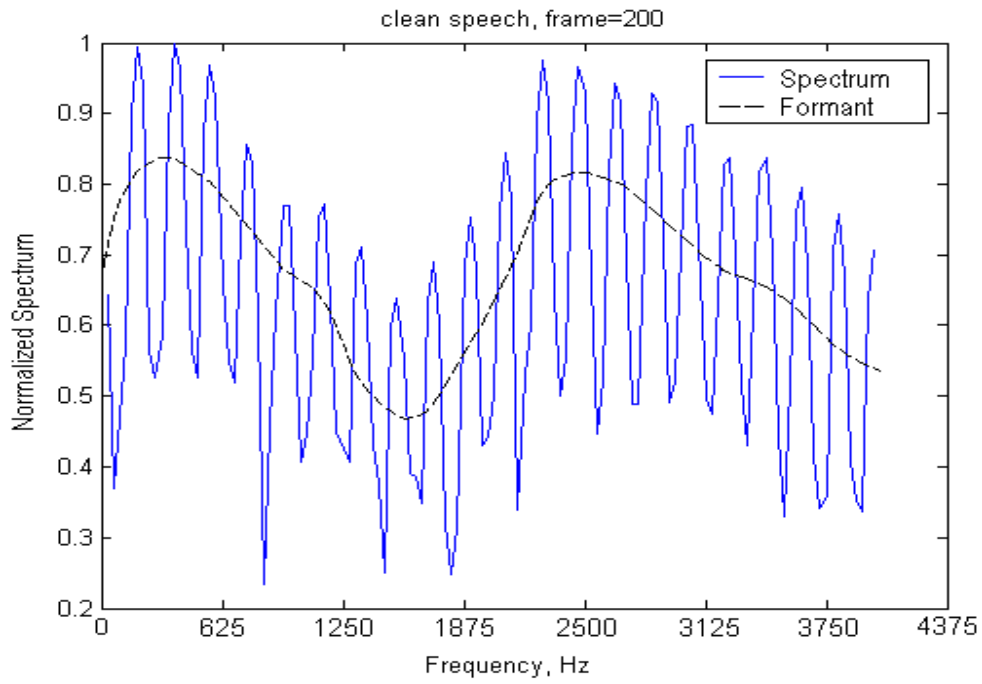

Figure 3.4: Spectrum of a frame of voiced speech

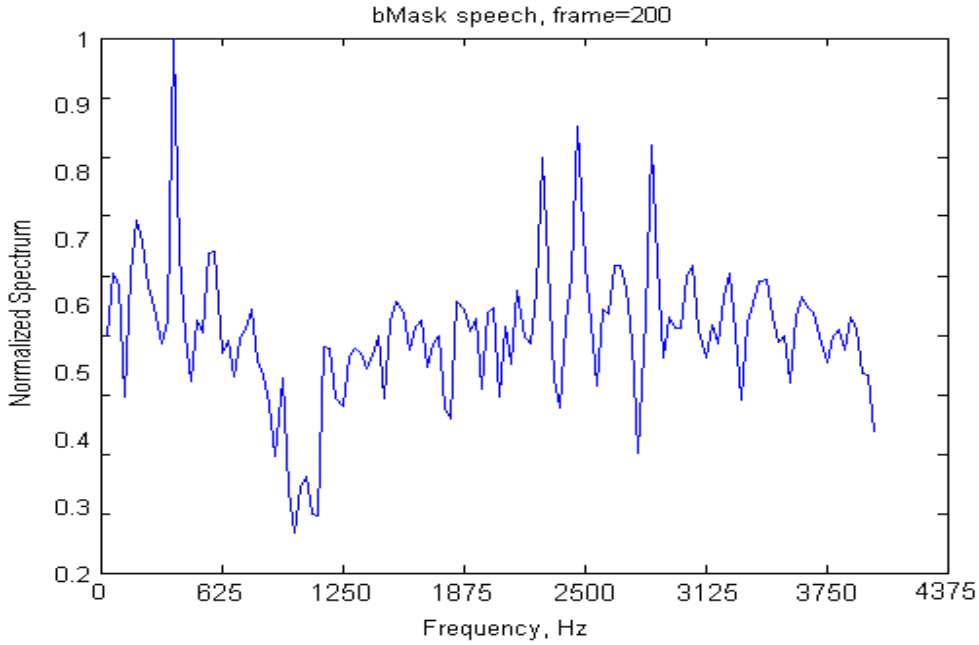

Figure 3.5: Spectrum of a frame of voiced speech (β-masking-enhanced speech)

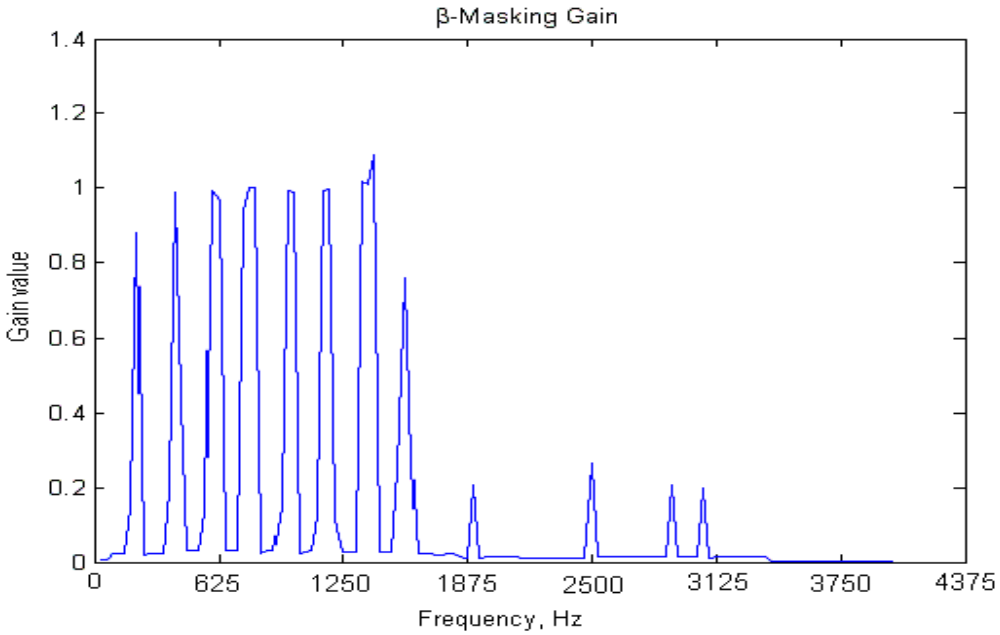

Figure 3.6: β-masking gain G_β of the same frame, which produces the spectrum shown in Figure 3.5

The autocorrelation function of the spectral components for a particular frame of speech signal is defined as:

$$r(n) = \frac{\sum_{i=1}^{M} X(i)X(i+n)}{\sum_{i=1}^{M} X^2(i)} \qquad for \quad 1 \le n \le M \tag{3.2}$$

with $X(i+n) = X(i+n-M)$ $\quad for \quad i+n > M$, and M is half of the window length. For example, if a window size of 256 samples is used, the effective number of spectral components $M=128$, excluding the dc component.

The autocorrelations of the clean speech signals and the β-masking enhanced speech signal of Figures 3.4 and 3.5 are as shown in Figures 3.7 & 3.8, respectively. Figure 3.7 shows that the correlation at the harmonic frequencies is more than 0.95, and it

drops periodically at the spectral valleys. The "comb" shape of the correlation plot indicates that the spectral components of the clean speech are well correlated due to their periodicity. For β-masking enhanced speech, the correlation is much lower, but the periodicity at the lower frequency band is still well retained. This is the same for the β-masking gain G_β, and in fact, it is the loss of much periodicity in G_β that causes the loss of periodicity in the β-masking enhanced speech. Next, we propose a way to make the gain function, G_β, periodic by shifting and adding the spectrum of the β-masking enhanced speech by using the autocorrelation method.

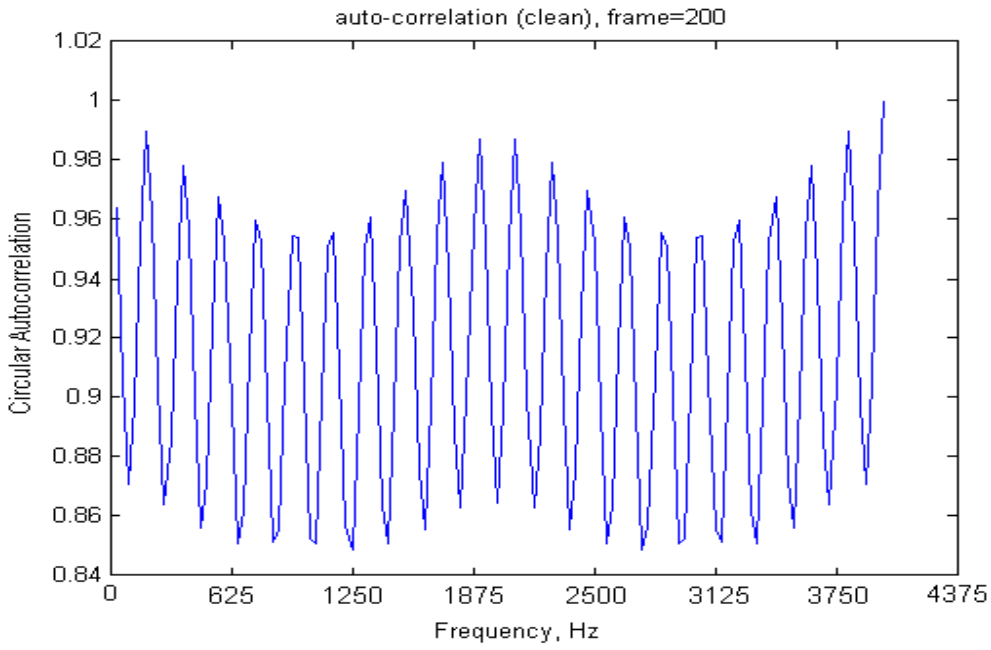

Figure 3.7: Autocorrelation of the spectrum of the voiced speech shown in Figure 3.4

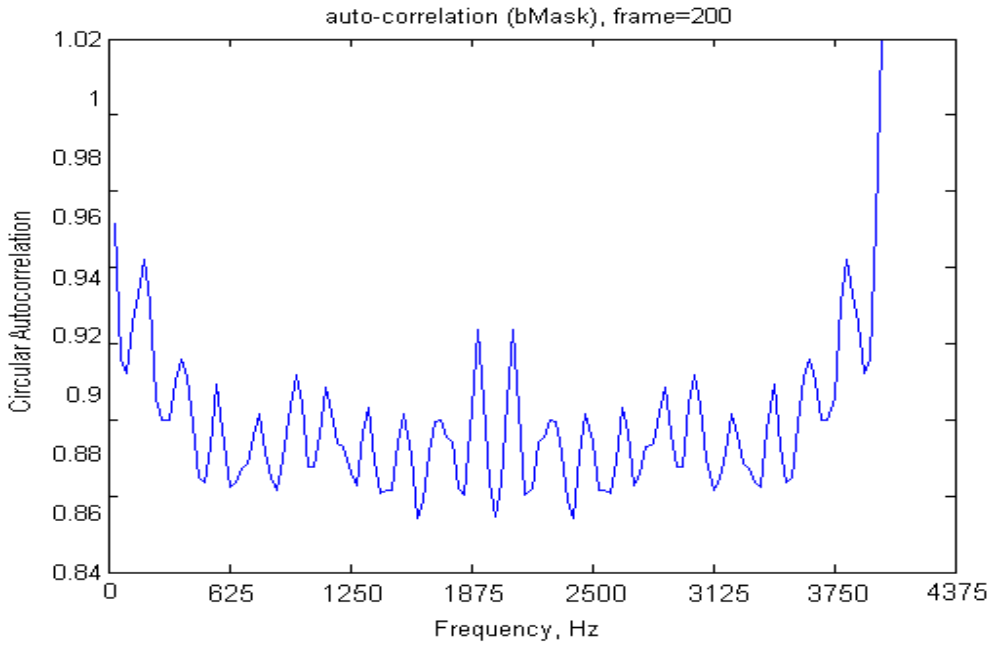

Figure 3.8: Autocorrelation of the spectrum of the voiced speech shown in Figure 3.5

Given a frame of β-masking enhanced speech $X(i)$, $1 \le i \le 128$ (e.g. the spectrum as shown in Figure 3.5), we shift and add the spectral components in the following way:

$$X_1(i) = X(i) + \sum_j X(i+j) \cdot r(i) \tag{3.3}$$

where j is the index for the local maxima of r. Here X_1 is the re-synthesized spectrum of X, as shown in Figure 3.9. Ideally, $j=n \times p$, where n is an integer and p is the spectral interval between the harmonics. Practically, j is the index of the local maxima of the correlation function. The spectrum is shifted and added to produce the periodicity. Compared to the spectrum in Figure 3.5, the re-synthesized spectrum is periodic. Next, this re-synthesized periodic spectrum is taken as the noisy speech to calculate a new set of β-masking gain $G_{\beta C}$ as shown in Figure 3.10, which is more periodic. With this

periodic gain $G_{\beta C}$ applied on the original noisy speech, we expect the output speech to be more periodic than the original β-masking enhanced speech signals.

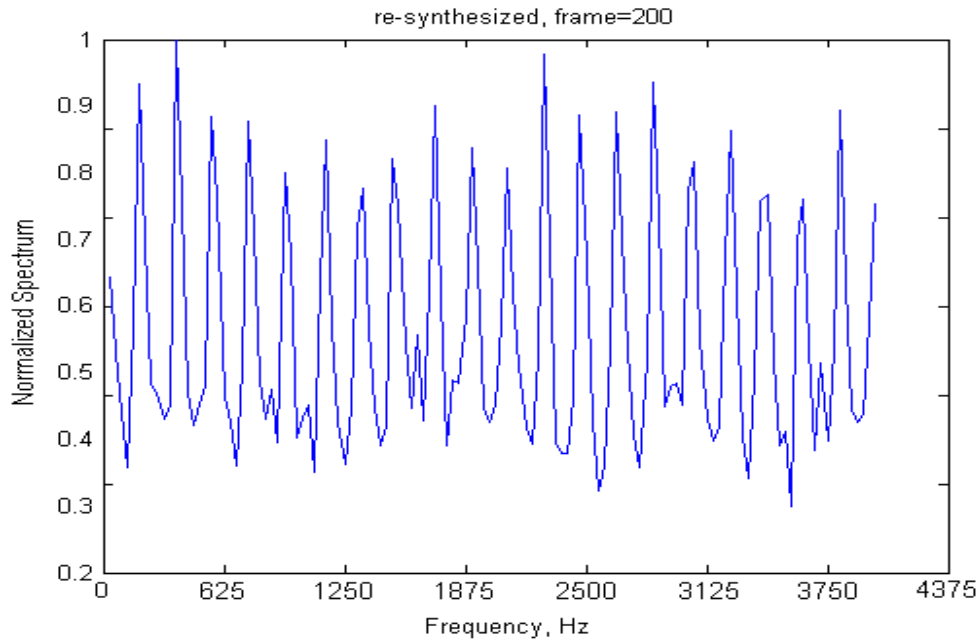

Figure 3.9: Re-synthesized spectrum (β-masking enhanced speech)

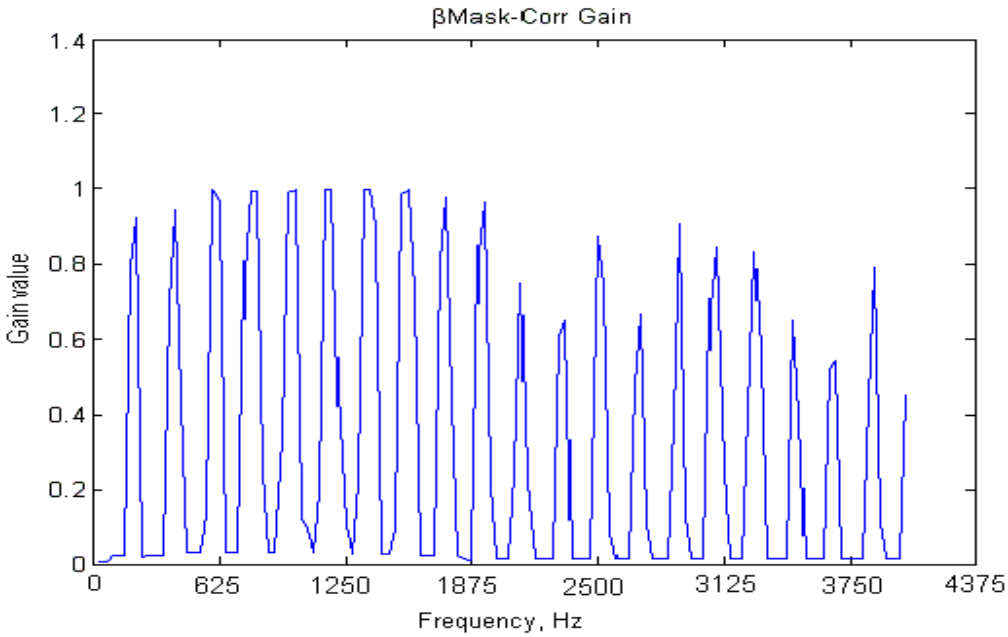

Figure 3.10: New β-masking gain $G_{\beta C}$

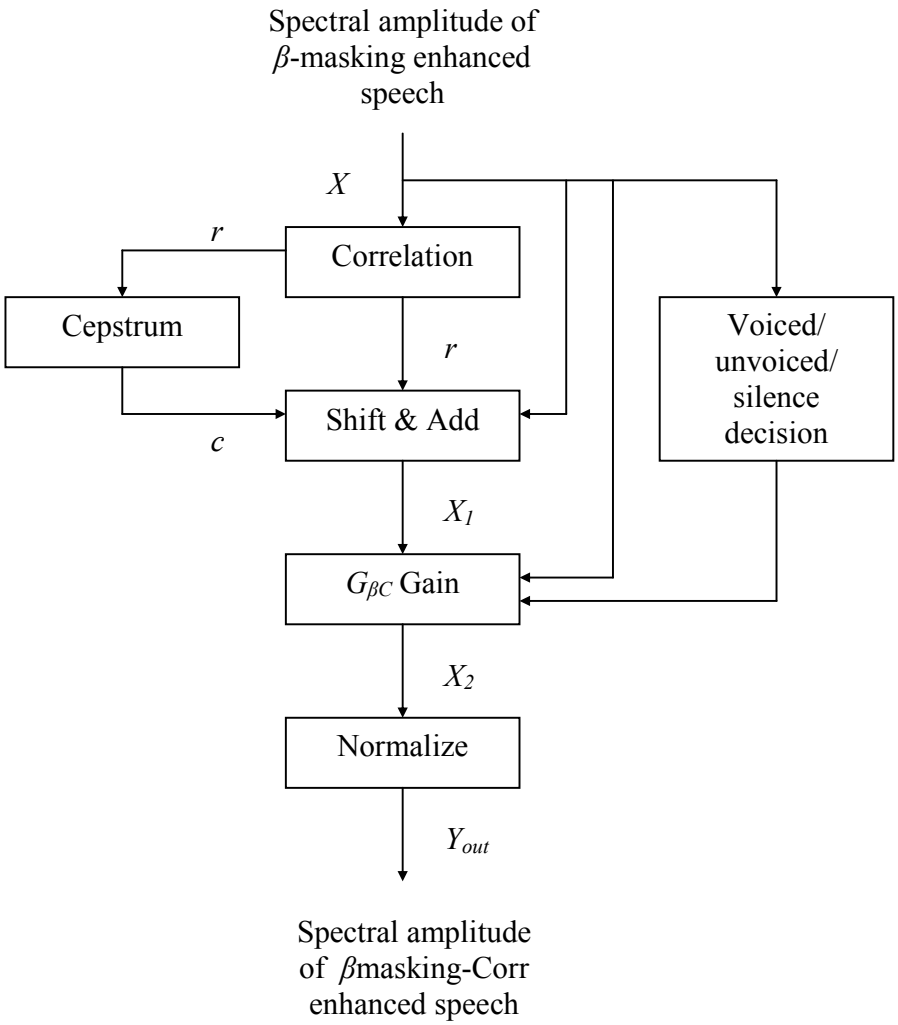

Figure 3.11: Block diagram of the cepstrum-aided correlation method

The block diagram of the correlation method is given in Figure 3.11. First the spectrum of the β-masking enhanced speech signals X is normalized and its autocorrelation, r, is obtained. Next X is shifted and added according to equation (3.3), with the additional use of cepstrum which will be discussed in the following section. The re-synthesized spectrum X_1 is periodic, and therefore, the gain $G_{\beta C}$ is periodic. The gain

$G_{\beta C}$ developed here is then applied to the original noisy speech (which is totally unprocessed) to obtain a new output Y_{out}. The function "*Normalize*" is the same as that of Equation (3.1), i.e.,

$$Y_{out}(k) = X_2(k)\frac{envlp(X(k))}{envlp(X_2(k))}. \tag{3.4}$$

With a gain that restores the periodicity, the weak spectrum that was masked by noise is now restored.

The correlation shown in Figure 3.9 is useful because it retains the periodicity of the original speech. The local maxima of the correlation indicate where the spectral peaks are. However for some other speech frames, the local maxima of the correlation may not be exactly equal to the interval of the harmonics, as shown in figure 3.12. Thus the normalized re-synthesized spectra obtained as shown in Figure 3.13 are obviously not what they are intended to be. In order to remove the interfering peaks, the use of cepstrum as an additional piece of information is considered.

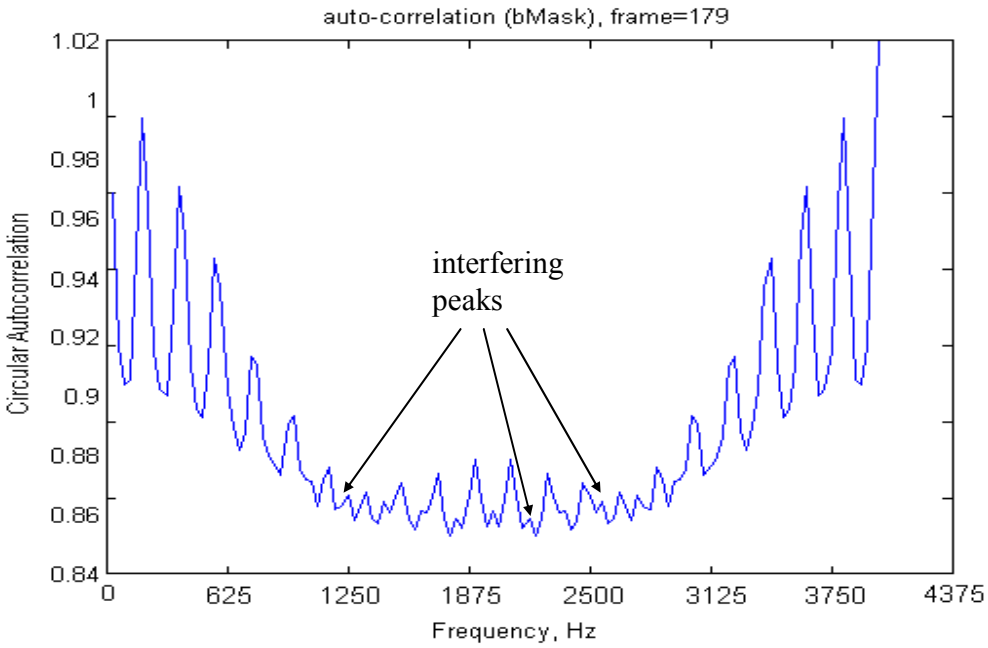

Figure 3.12: Autocorrelation of a frame of voiced β-masking-enhanced speech, where the local maxima do not indicate periodicity correctly

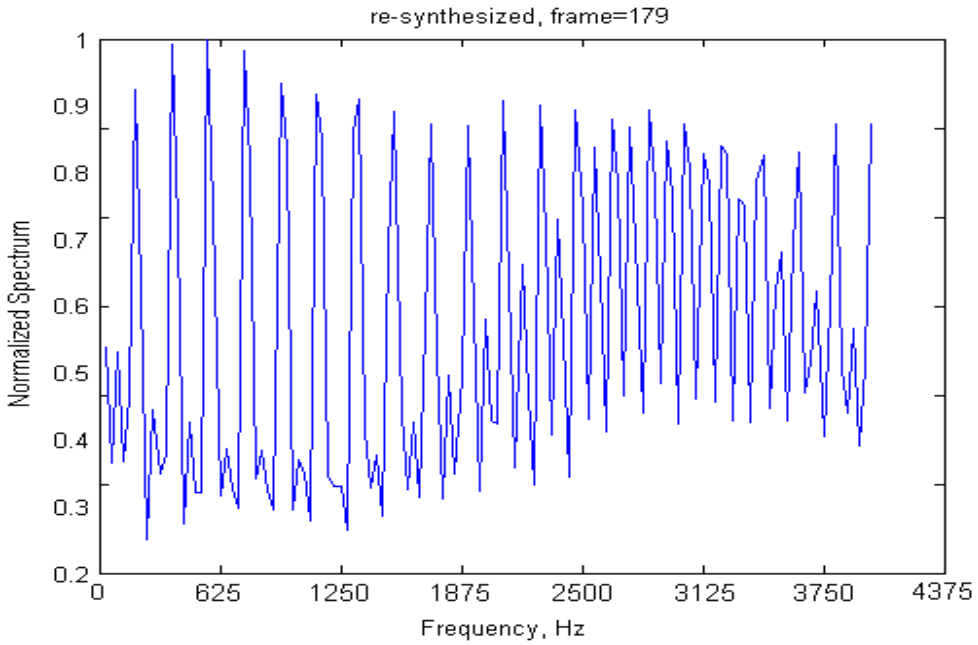

Figure 3.13: Re-synthesized spectrum, when the local maxima of the autocorrelation do not indicate periodicity correctly

3.2.2. The Cepstrum

The cepstrum of a signal is the Fourier analysis of the logarithmic amplitude spectrum of the signal [41]:

$$c(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \log\left|X(e^{j\omega})\right| e^{j\omega \cdot n} d\omega \qquad (3.5)$$

If the log amplitude spectrum contains many regularly spaced harmonics, then the Fourier analysis of the spectrum will show a peak at the position corresponding to the spacing between the harmonics, i.e., the fundamental frequency, which is depicted in Figure 3.14:

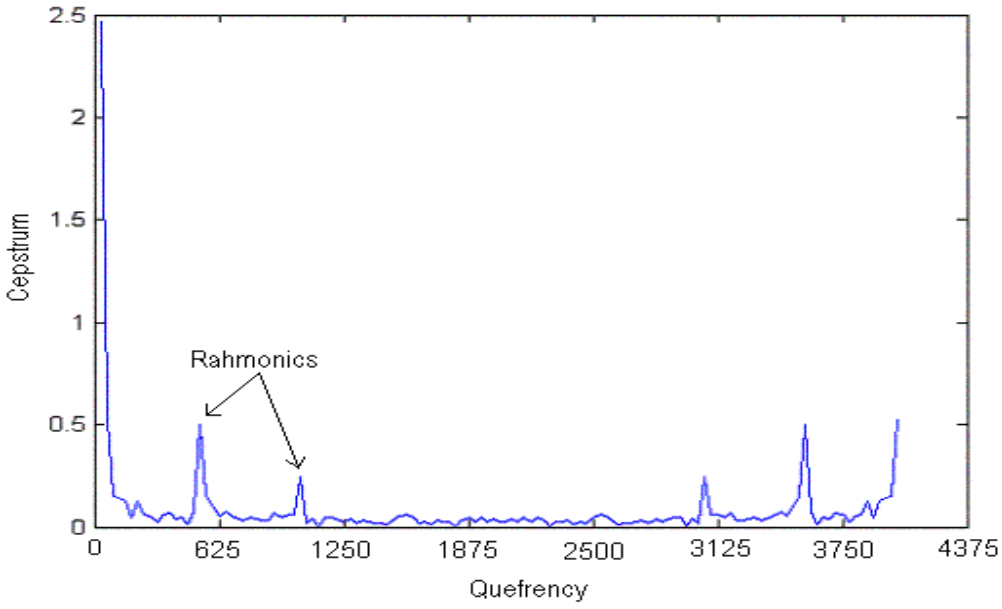

Figure 3.14: Cepstrum of a frame of voiced speech

The word "*cepstrum*" is so-called because it turns the word "*spectrum*" inside-out. The *x*-axis of the cepstrum has the unit of *quefrency* (which turns the word

"*frequency*" inside-out), and peaks in the cepstrum, which relate to periodicities in the spectrum, are called *rahmonics* (which turns the word "*harmonics*" inside-out).

Given the cepstrum of a frame of voiced speech, the rahmonics can be easily distinguished. By using the interval between two adjacent rahmonics, setting all other values to zero and converting the cepstrum back into the spectrum domain, we can obtain a pseudo spectrum which is smooth with perfect harmonics, as shown in Figure 3.15. The location of the peaks can be used to decide where the wanted and unwanted peaks are located. In Figure 3.15, the information of the frequencies corresponding to the peaks is saved. Peaks of Figure 3.13 located at these frequencies are considered as wanted peaks and others are unwanted. With the aid of the correctly constructed harmonics positions obtained from the cepstrum, the effect of the interfering peaks (i.e., the unwanted peaks in Figure 3.13) can be removed because the shift and add operation will be skipped when the unwanted peaks are detected. The spectrum can be re-synthesized and become periodic as shown in Figure 3.16, which can be used to help increase the periodicity of the β-masking gain.

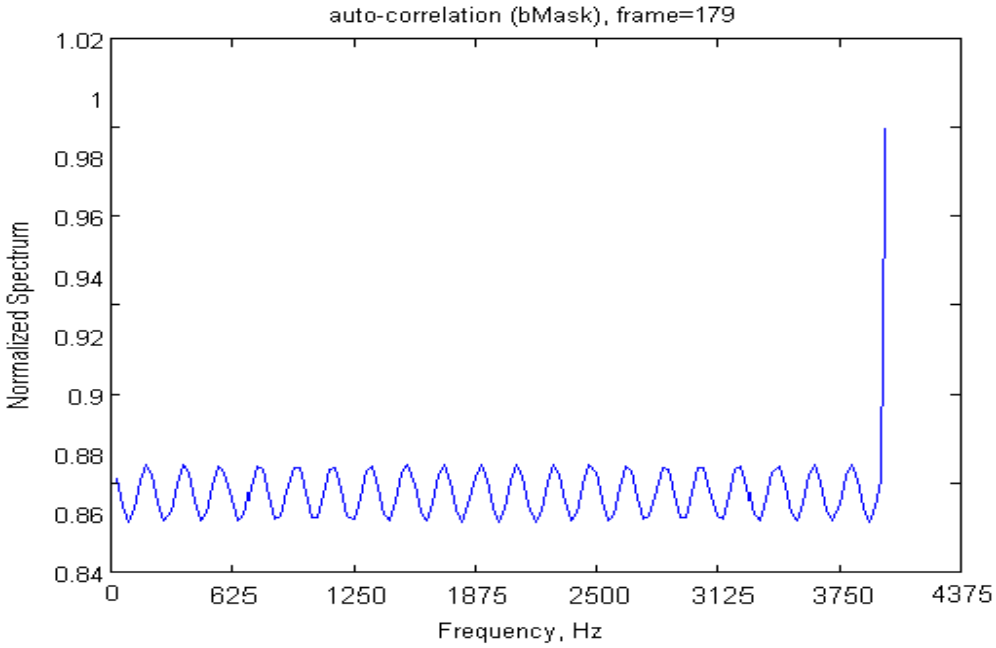

Figure 3.15: Cepstrum-smoothed pseudo-spectrum

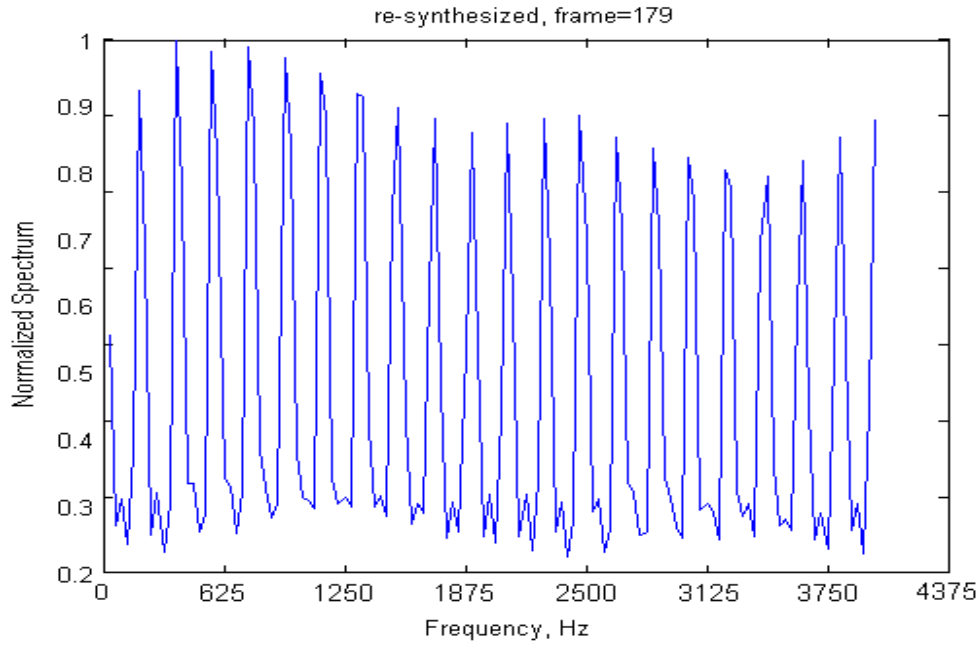

Figure 3.16: Re-synthesized spectrum

3.3. Simulation Results, Comparisons & Discussions

3.3.1. Noisy & Enhanced Speech signals

Sixteen different speech utterances from the TIMIT database were used for simulation. Eight sentences are spoken by females and the other eight are by males:

-- Utterances from females:

FA.pcm ("Jane may earn more money by working hard.")

FB.pcm ("She had your dark suit in greasy wash water all year.")

FC.pcm ("Serve the coleslaw after I add the oil.")

FD.pcm ("Steve collects rare and novel coins.")

FE.pcm ("Dolphins are intelligent marine mammals.")

FF.pcm ("Don't ask me to carry an oily rag like that.")

FG.pcm ("Cheap stockings run the first time they're worn.")

FH.pcm ("The rides were tame enough; mostly we talked.")

-- Utterances from males:

MA.pcm ("A muscular abdomen is good for your back.")

MB.pcm ("The causeway ended abruptly at the shore.")

MC.pcm ("To many experts, this trend was inevitable.")

MD.pcm ("The big dog loved to chew on the old rag doll.")

ME.pcm ("We saw eight tiny icicles below our roof.")

MF.pcm ("The paper boy bought two apples and three ices.")

MG.pcm ("Penguins live near the icy Antarctic.")

MH.pcm ("Don't ask me to carry an oily rag like that.")

These utterances were sampled at 8 kHz and linearly quantized at 16 bits. The sixteen sets of sampled data were then corrupted to different segmental SNRs (-8.6dB, -4.3dB, 0dB & +4.3dB) by additive white noise. There were altogether 64 noisy speech files (e.g., *fa_n8.pcm, fa_n4.pcm, fa_p0.pcm, fa_p4.pcm, etc*) to be processed by β-masking method, and the β-masking enhanced speech signals are then post-processed by the NHR technique (βmasking-NHR enhanced) and the autocorrelation technique (βmasking-Corr enhanced), resulting in 192 enhanced speech utterances.

3.3.2. Comparisons of Spectral Components

We choose the clean speech utterance *fa.pcm:* "Jane may earn more money by working hard" in this thesis as an example. The speech was added with Gaussian white noise to obtain the noisy speech *fa_p0.pcm* whose segmental SNR is zero dB. Figure 3.17(1-5) shows the spectrograms of the clean speech, noisy speech, β-masking enhanced speech, βmasking-NHR enhanced speech and βmasking-Corr enhanced speech, respectively.

We can observe from the spectrogram of the unprocessed noisy speech signal (Figure 3.17.2) that the original speech signal is degraded very badly, so mush so that most of the weak spectral components are totally masked. These components could hardly be distinguished from the corrupting noise. The enhanced speech signals (Figure 3.17.3-5) are apparently spectrally cleaner and perceptually enhanced for listeners to comprehend. Figure 3.17.3 shows the spectrogram of the β-masking enhanced speech. Musical tone and white noise can still be perceived in the enhanced speech signals. For

some instances, the periodicity of the harmonics is lost in the high frequency band. By folding the lower-band information to re-synthesize the higher-band harmonics, we obtain the βmasking-NHR enhanced speech, as shown in figure 3.17.4. By using the autocorrelation post-processing method with the additional use of cepstrum, we obtain the βmasking-Corr enhanced speech, as shown in figure 3.17.5. Both post-processing methods are able to improve the upper-band harmonics to a certain extend. They are also effective in removing some of the high frequency music-like residues in the β-masking enhanced speech, because the residual musical tones are replaced by the re-constructed speech harmonics.

We can also have a closer look at the voiced frames as in Figure 3.18. The spectral components of the clean, noisy and enhanced speech signals are also plotted in Figures 3.19-3.22. These figures show that the NHR and autocorrelation post-processing methods are able to recover some of the high frequency harmonics of the voiced sounds. However, the above discussions are purely qualitative. Next, we will examine the performance of the proposed methods by using objective assessments which include average segmental SNR, perceptual evaluation of speech quality (PESQ) and psycho-acoustically motivated distortion (PMD) measure.

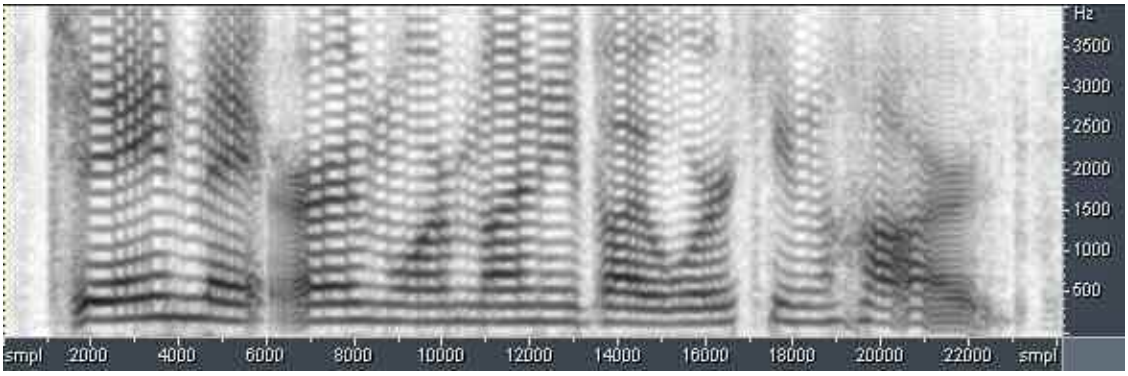

Figure 3.17 (1): Spectrogram of the clean speech (FA.pcm)

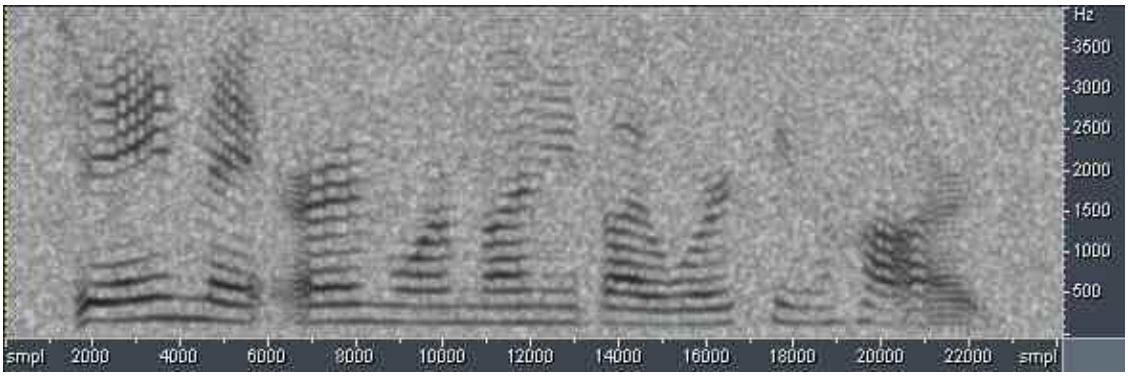

Figure 3.17 (2): Spectrogram of the noisy speech (SNR = 0dB)

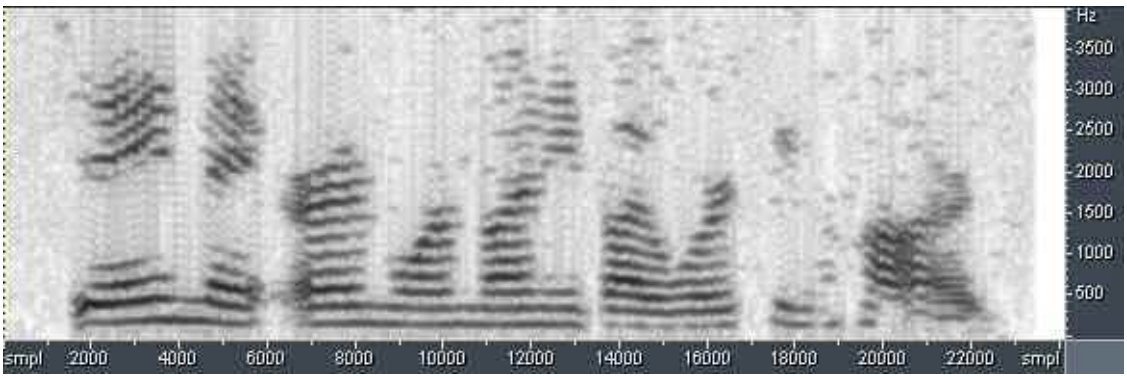

Figure 3.17 (3): Spectrogram of the speech enhanced by β-masking approach

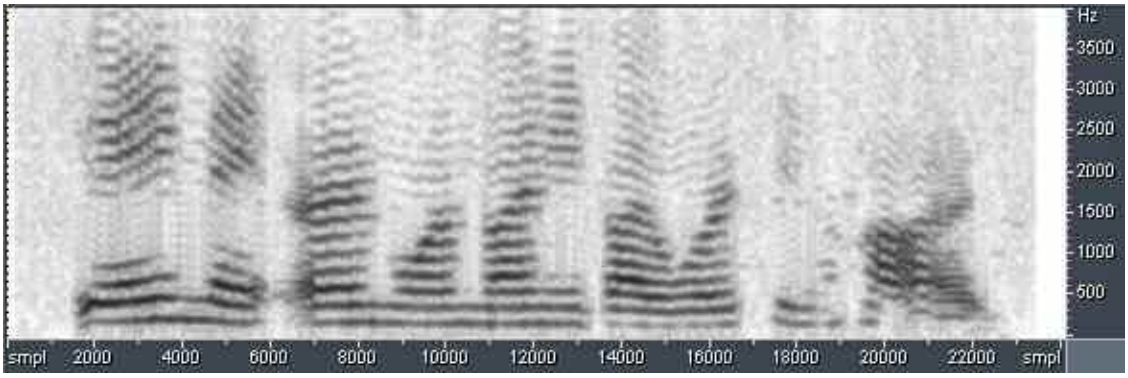

Figure 3.17 (4): Spectrogram of speech enhanced by βmasking-NHR approach

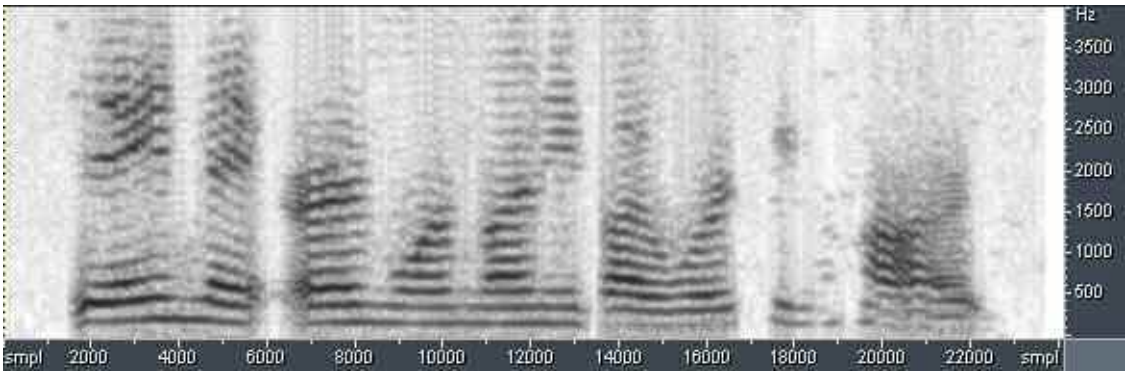

Figure 3.17 (5): Spectrogram of speech enhanced by βmasking-Corr approach

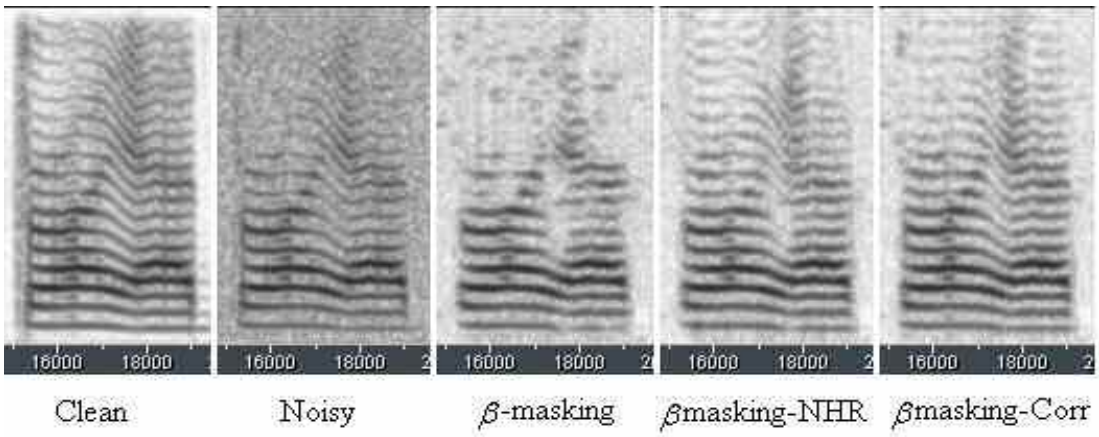

Figure 3.18: Spectrogram of a particular voiced frame

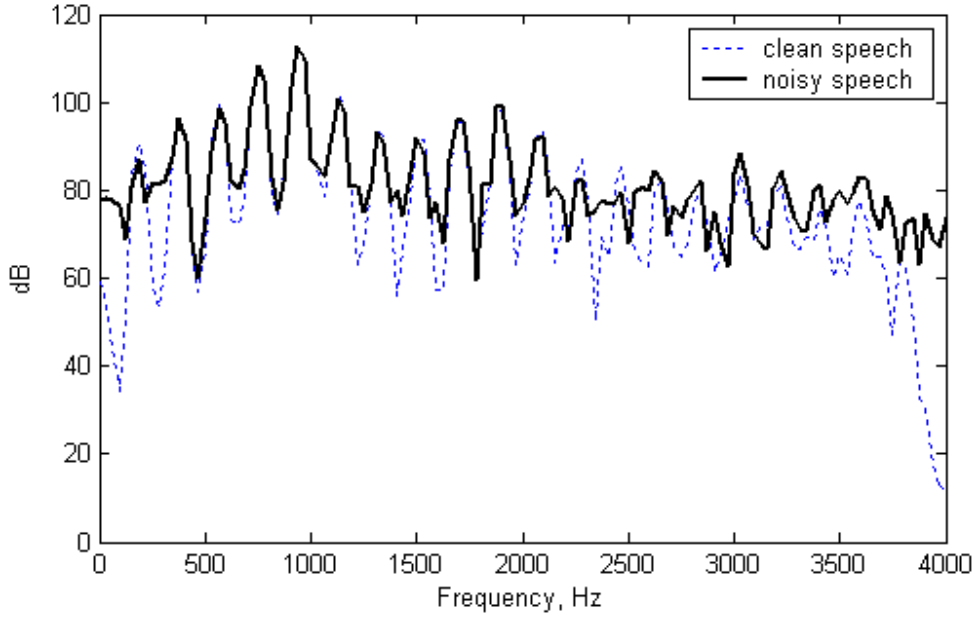

Figure 3.19: Frequency components of a white-noise-corrupted speech signal with some components completely masked by noise

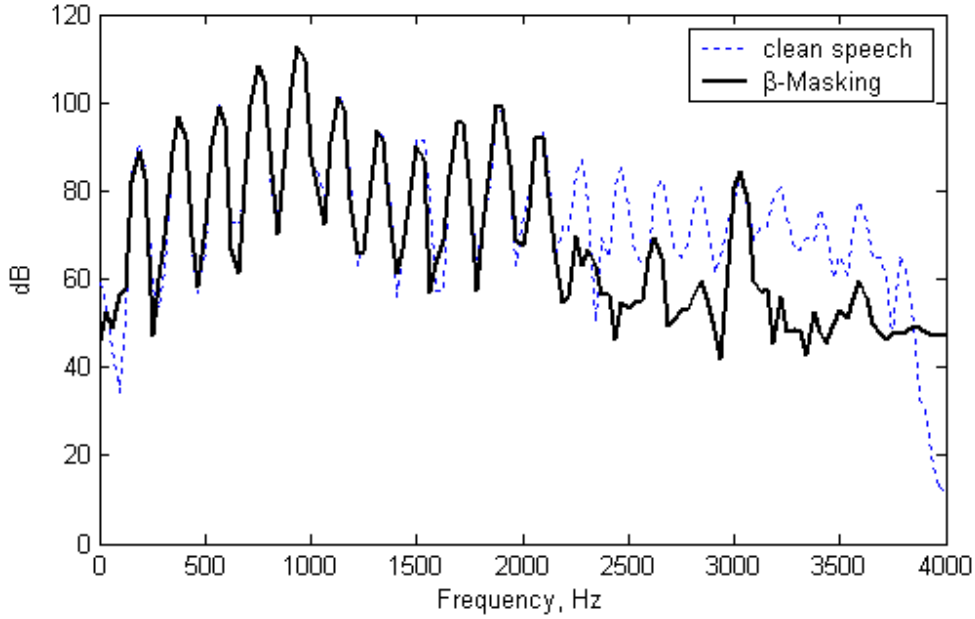

Figure 3.20: Frequency components of the β-masking enhanced speech signal with some components failed to be recovered

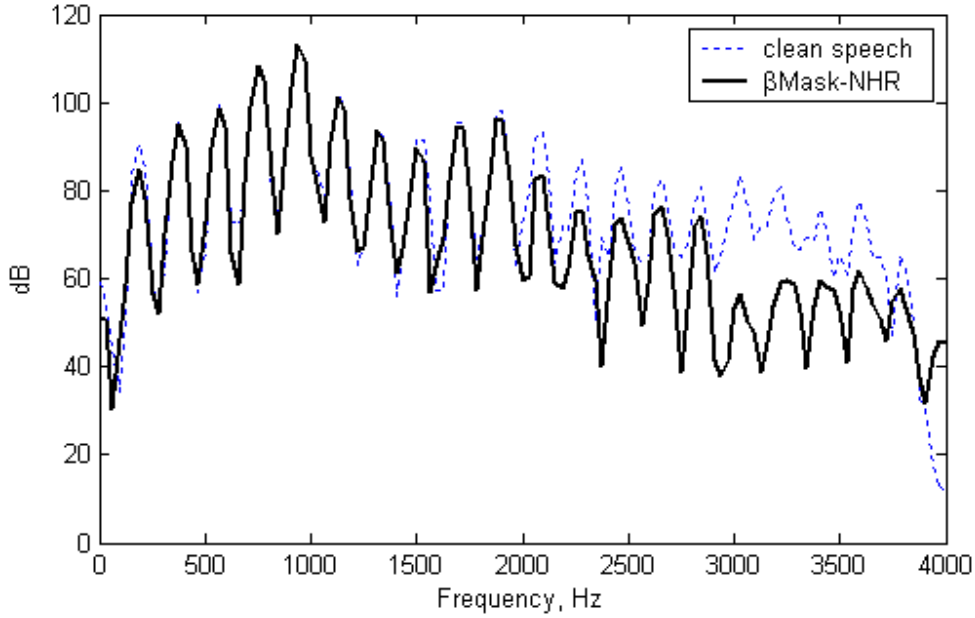

Figure 3.21: Frequency components of the βmasking-NHR enhanced speech signal

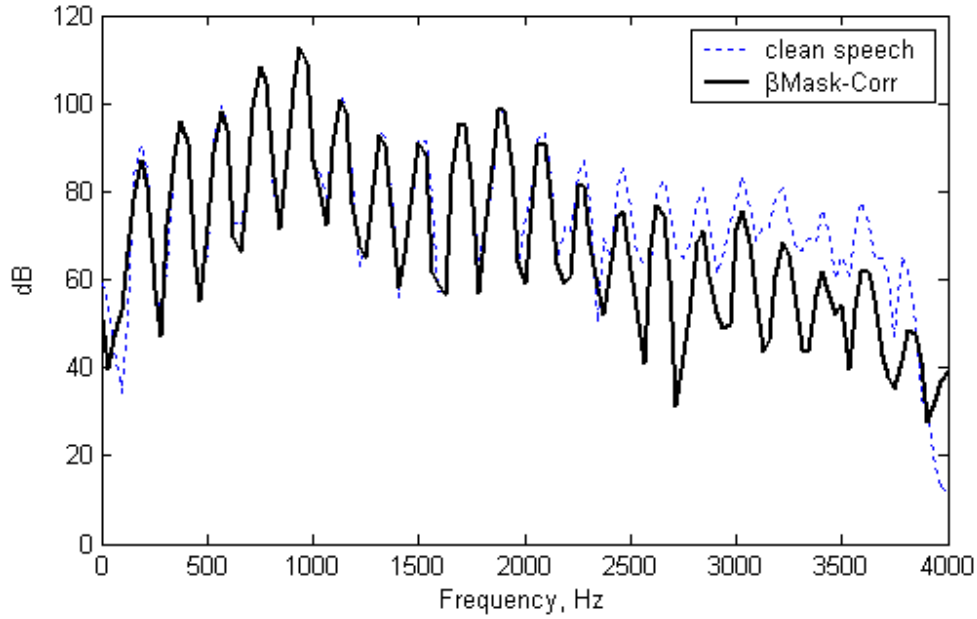

Figure 3.22: Frequency components of the βmasking-Corr enhanced speech signal

3.3.3. Segmental SNR Evaluations

One of the characteristics of speech signals is its time-varying nature, which results in some speech segments with high energy and other segments with low energy. The perceptual effects of noise in the regions of lower energies are more severe. In order to take this into consideration, segmental signal-to-noise ratio (SNR) is used instead of global SNR. The average segmental SNR is a meaningful objective indication; it is measured by computing the SNR of each speech frame and averaging these measures over the entire signal. It is defined as

$$\Xi_{seg} = \frac{1}{L_s}\sum_{l=0}^{Ls-1}\Xi_{block}(l) \tag{3.6}$$

where the block SNR $\Xi_{block}(l)$ is defined as

$$\Xi_{block}(l) = 10\log_{10}\left(\frac{\sum_{j=0}^{Ms-1} s^2(M_s l - j)}{\sum_{j=0}^{Ms-1}[x(M_s l - j) - s(M_s l - j)]^2}\right) \tag{3.7}$$

s denotes the clean speech magnitude, and x denotes the processed speech (noisy or enhanced) magnitude. M_s denotes the number of samples per block and L_s is the number of blocks in a given test utterance.

Figure 3.23 shows the output segmental SNR versus the input segmental SNR of the noisy and the three types of enhanced speech signals. Higher value of SNR indicates better speech quality. To avoid massive plotting, each data point in the figure is the average SNR of eight utterances. For example, the "sample point 1" shown in Figure 3.23 means that the eight clean speech utterances MA to MH are degraded to obtain the –8.6 dB noisy speech signals; after the βmasking-NHR enhancement process, these eight utterances have an average segmental SNR of 3.8dB, which is to say, the average

segmental SNR improvement of the *β*masking-NHR method applied on the –8.6dB (male) signals is 12.4dB. This SNR improvement is also given in Table 3.1. Besides figure 3.23, the data points in figure 3.24 and 3.25 are also the average values of eight measurements.

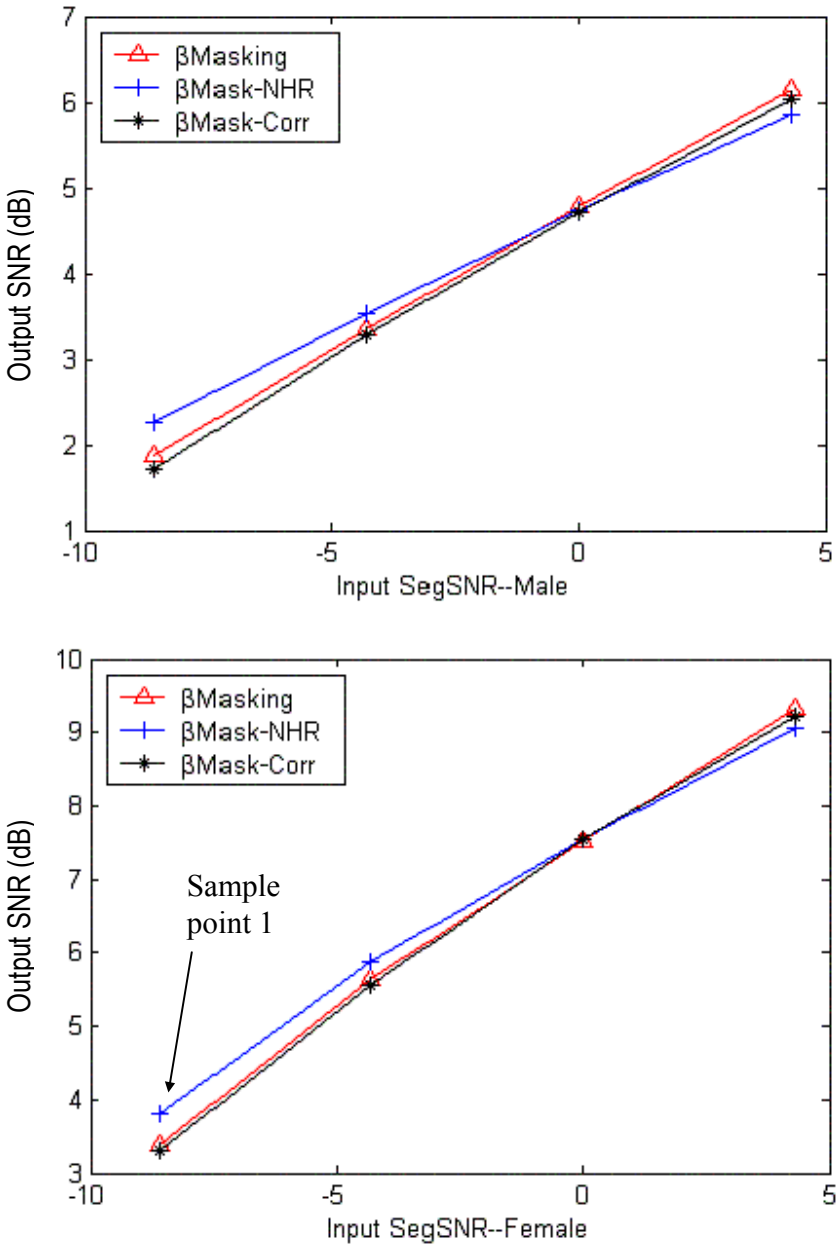

Figure 3.23: Output Segmental SNR versus Input Segmental SNR

SegSNR (dB) of the unprocessed speech signals		SegSNR (dB) improvement of the β-masking enhanced speech signals	SegSNR (dB) improvement of the βmasking-NHR enhanced speech signals	SegSNR (dB) improvement of the βmasking-Corr enhanced speech signals
Female	-8.6	10.40	10.90	10.25
	-4.3	7.45	7.75	7.44
	0	4.70	4.69	4.69
	4.3	1.80	1.60	1.75
Male	-8.6	11.90	12.40 (Sample point 1)	11.88
	-4.3	9.90	10.20	9.88
	0	7.52	7.52	7.52
	4.3	4.90	4.72	4.85

Table 3. 1: Segmental SNR improvement (average of 8 utterances) of speech signals enhanced by the three methods

All of the three enhancement methods achieve a large improvement in segmental SNR as compared to the noisy speech signals; however, they have nearly the same segmental SNR results when they are compared with each other. As segmental SNR provides only the time domain magnitude difference between the clean and processed speech signals, it may not indicate truly the perceptual effect. Therefore, PESQ and PMD measurements are used in later sections.

3.3.4. PESQ Evaluations

Perceptual evaluation of speech quality (PESQ) is the ITU-T recommendation for predicting the quality of 3.1 kHz narrow-band speech codecs objectively [42]. PESQ compares an original speech signal *s(t)* with its processed speech signal *x(t)*. The highest score of PESQ that a speech can take is 4.5, i.e., when *s(t)* is compared to *s(t)* itself. Figure 3.24 shows the output PESQ Mean Opinion Score (MOS) versus input PESQMOS. As PESQ indicates "how much likeness a processed speech signal compared to its original version", it takes into account both perceptual frequency (in Bark) and loudness (in Sone). Figure 3.24 shows the PESQ scores of the three types of enhanced speech signals and Table 3.2 shows the PESQ improvement accordingly. All of these signals processed by the three methods have a higher PESQ score as compared to the unprocessed noisy speech utterances. For male speech signals, the non-linear high-frequency regeneration and the cepstrum-aided autocorrelation methods give better PESQ scores than β-masking alone. For female speech signals, the correlation method gives better PESQ score, and the other two methods have almost the same scores.

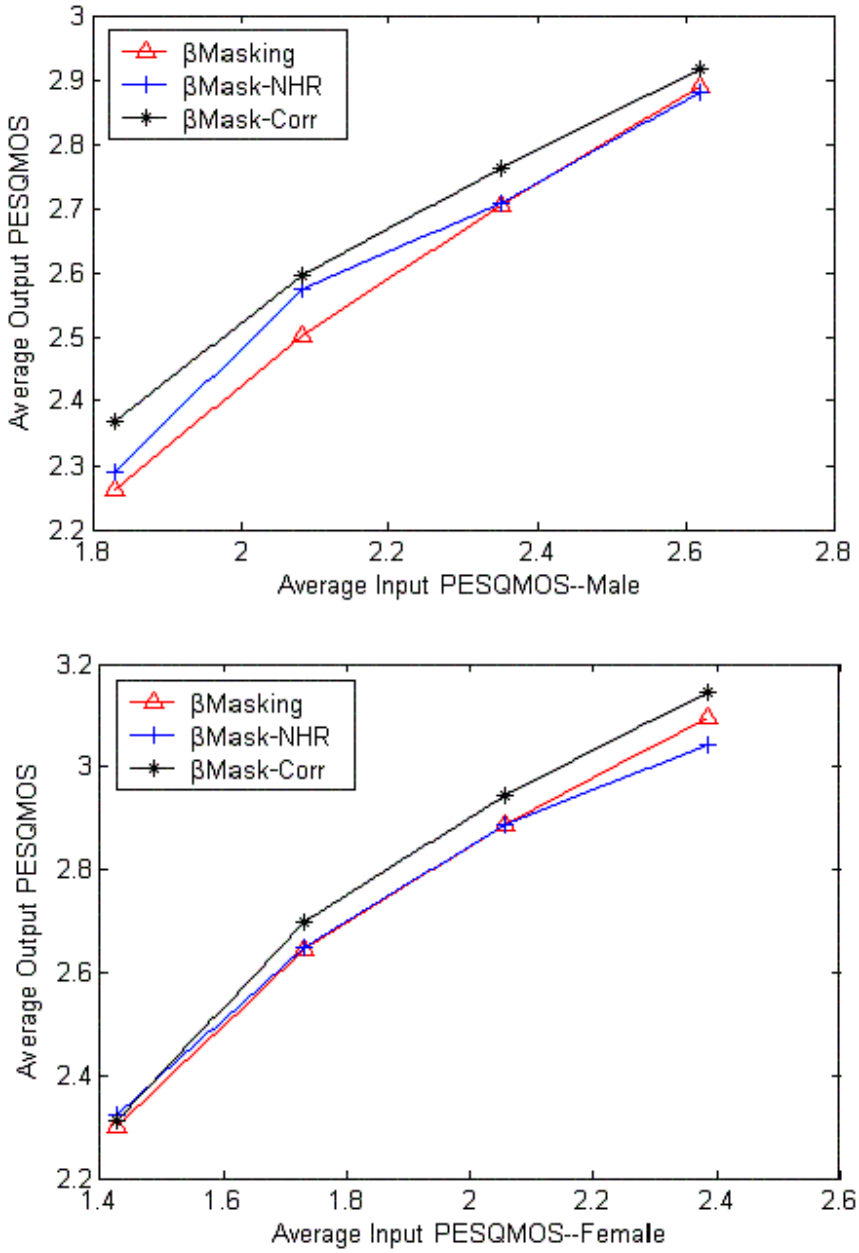

Figure 3.24: PESQ measurements of the three enhancement methods

PESQMOS of the unprocessed speech signals		PESQMOS improvement of the β-masking enhanced speech signals	PESQMOS improvement of the βmasking-NHR enhanced speech signals	PESQMOS improvement of the βmasking-Corr enhanced speech signals
Female	1.83	0.43	0.465	0.53
	2.09	0.40	0.49	0.51
	2.35	0.35	0.36	0.41
	2.62	0.28	0.27	0.30
Male	1.43	0.87	0.88	0.875
	1.75	0.89	0.895	0.95
	2.05	0.83	0.83	0.89
	2.39	0.71	0.66	0.76

Table 3. 2: PESQMOS improvement (average of 8 utterances) of speech signals enhanced by the three methods

3.3.5. PMD Evaluations

Another objective measurement is the psycho-acoustically motivated distortion measure (PMD). The perceptually weighted error criterion is implemented by weighting the error spectrum with a filter, which has the shape of the inverse spectrum of the original signal [43]. Therefore, spectral peaks are not emphasized as much as spectral valleys, because noise near the formant peaks is masked and is not audible. The PMD distortion measure is defined as follows:

$$\Xi_{seg} = \frac{1}{L_s}\sum_{m=0}^{Ls-1} 10\log_{10}\left(\frac{\sum_{k=0}^{w/2+1}\left(S_k - \hat{S}_k\right)^2}{\sum_{k=0}^{w/2+1} S_k^p}\right) \tag{3.8}$$

where w denotes the window size, m denotes the frame index ranging from the first to the last (L_S) frames, and S_k and $\hat{S}_k$ denote the k-th spectral components of the clean and processed speech signals. When p=2, the PMD measure is similar to the distortion measure proposed by Itakura for comparing two autoregressive speech models. It can also be interpreted as the inverse of SNR (the expression in equation 3.8 shows "noise-to-signal ratio") except that the PMD measurement is in the frequency domain, where the weighting filter in the shape of the inverse spectrum is applied. Nine sample files are provided in the CD to give readers some idea of the subjective quality of noisy speech at different values of SNR and PMD. One of the speech files is unprocessed clean speech and the other eight are degraded by either white noise or car nosie at different values of SNR and PMD values. The SNR and PMD values of the noisy speech signals are provided in Appendix A.1.

The comparison result for the various single-channel speech enhancement methods is as shown in Figure 3.25 and Table 3.3. Lower value of PMD indicates better speech quality. The figure shows that the proposed enhancement methods of βmasking-NHR and βmasking-Corr leads to a much lower spectral distortion than β-masking method used alone. The result is as expected because the proposed methods aim to recover the over-attenuated spectral components, and each succeeds to a certain extent.

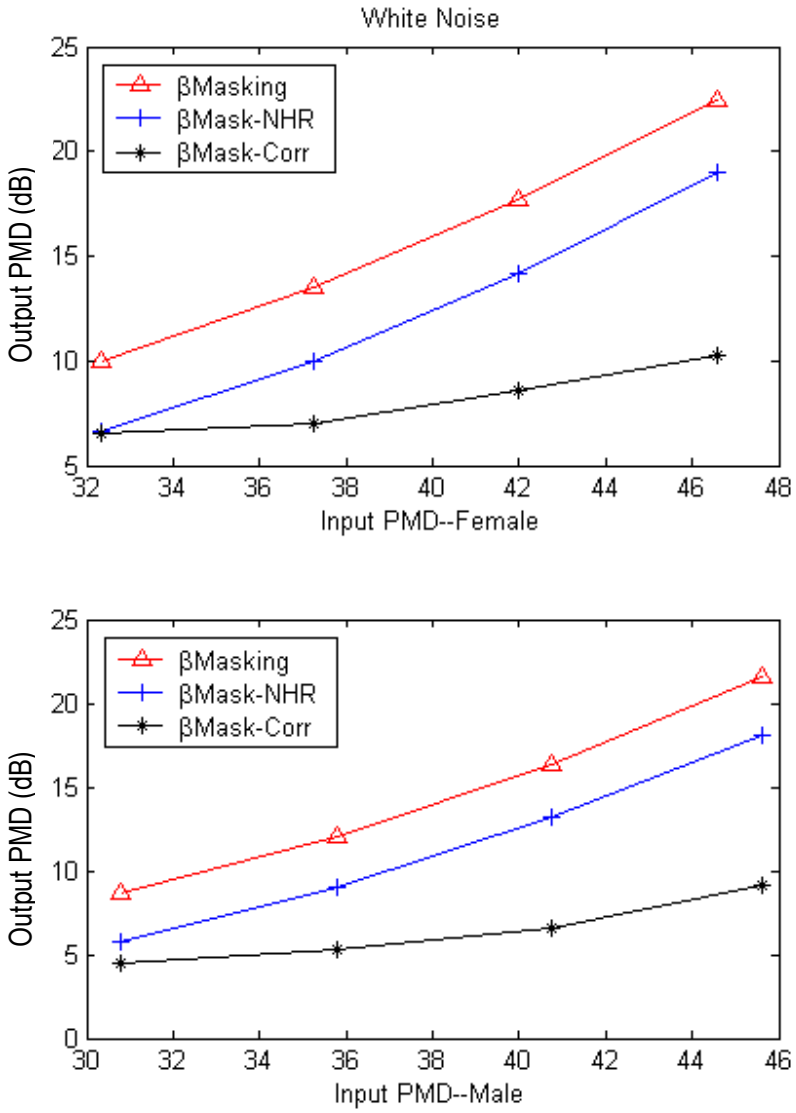

Figure 3.25: PMD measurements of the three enhancement methods (White Noise)

PMD (dB) of the unprocessed speech signals	PMD (dB) improvement of the β-masking enhanced speech signals	PMD (dB) improvement of the βmasking-NHR enhanced speech signals	PMD (dB) improvement of the βmasking-Corr enhanced speech signals

Female	46.6	24.1	27.6	36.2
	42.0	24.5	28.0	34.0
	37.2	23.7	27.55	29.6
	32.3	22.3	25.3	25.3
Male	45.7	23.7	27.7	31.7
	40.7	23.9	27.2	33.7
	35.8	23.5	27.3	28.8
	31.8	23.8	25.8	27.3

Table 3. 2: Average PMD improvement (average of 8 utterances) of speech signals enhanced by the three methods

3.3.6. Comparisons for Different noise types

Besides Gaussian white noise, different types of noises were also added to the clean speech sentences to evaluate the performance of the NHR and autocorrelation post-processing methods. Pink noise, F16 noise and car noise were used to repeat the same experiments. The PMD results are as shown in Figures 3.26 to 3.28. It is found that for F16 and pink noise corrupted noisy speech signals, the proposed NHR and autocorrelation post-processing methods do improve the speech quality by reducing the PMD measure to a low value.

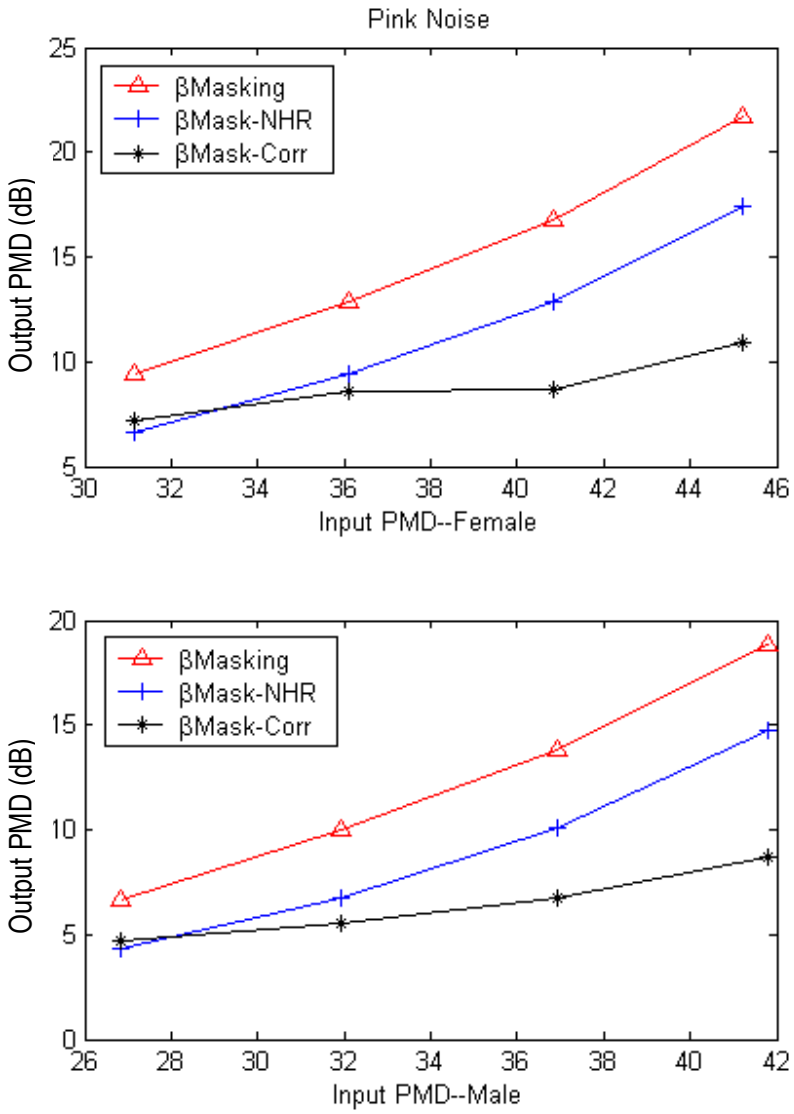

Figure 3.26: PMD measurements of the three enhancement methods (Pink Noise)

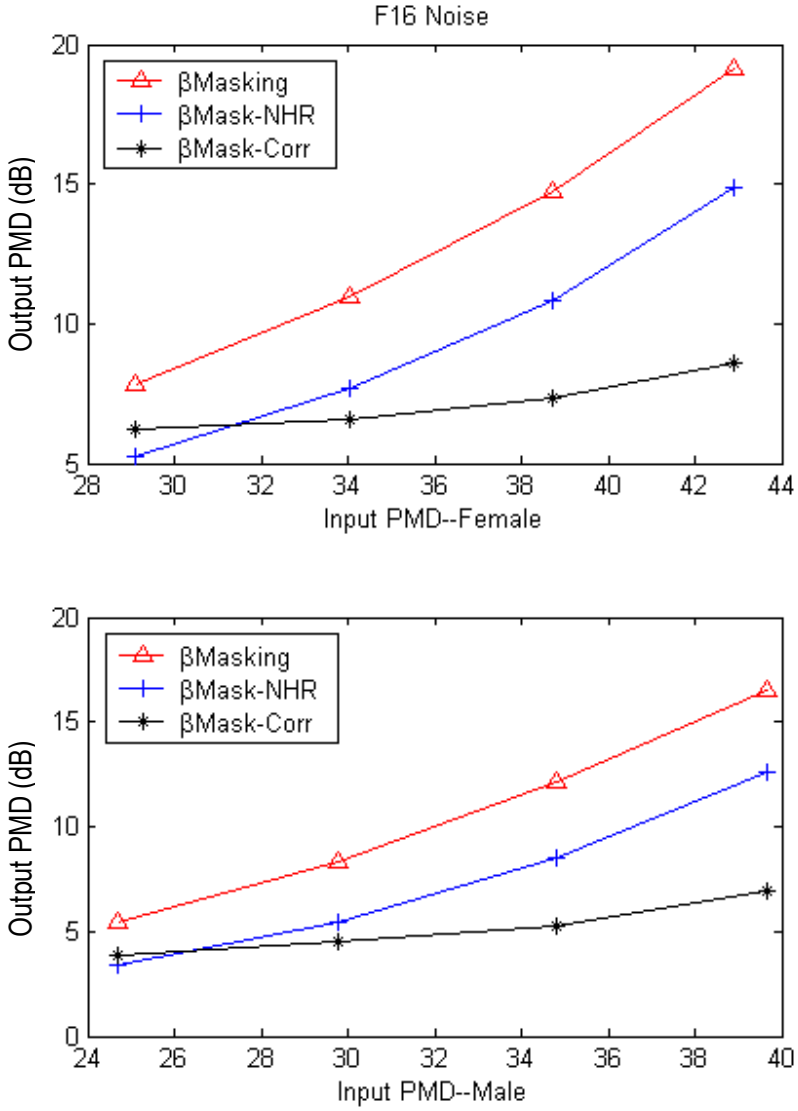

Figure 3.27: PMD measurements of the three enhancement methods (F16 Noise)

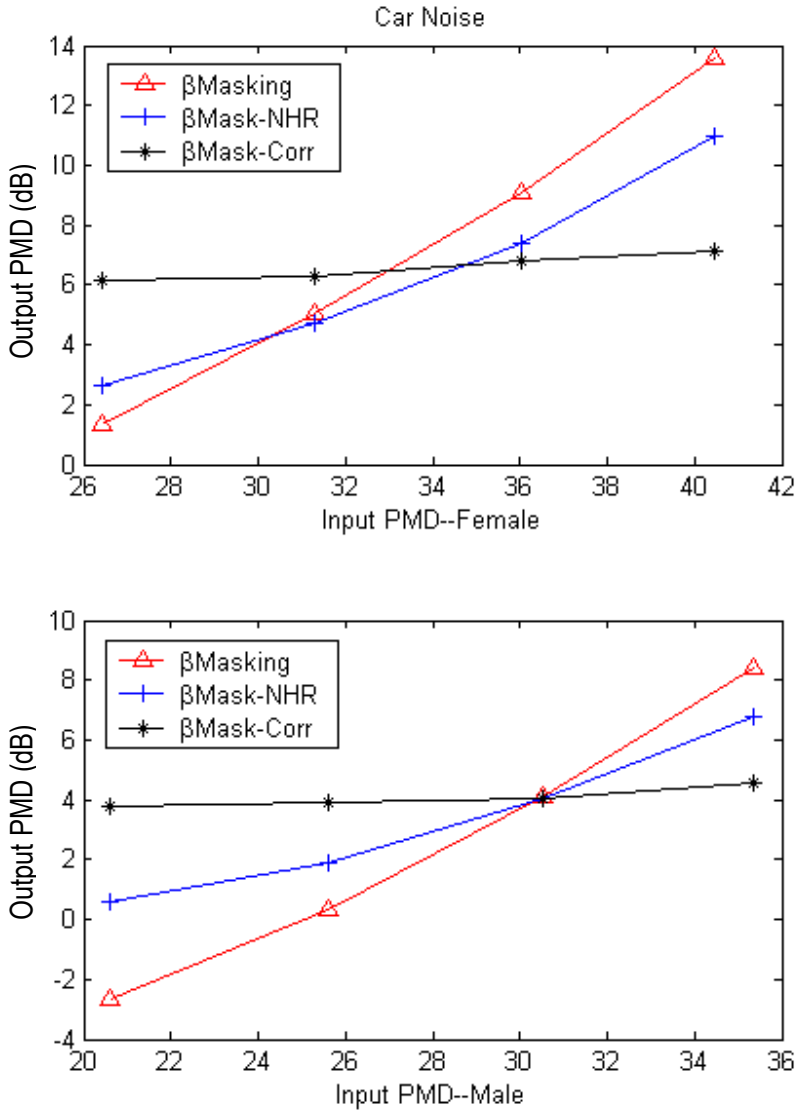

Figure 3.28: PMD measurements of the three enhancement methods (Car Noise)

For car noise corrupted signals, the two post-processing algorithms again show their advantage by reducing the PMD measurements on the processed speech signals when the noise level is high. However, when the noise level is low, their performance is worse than β-masking algorithm alone. The reason can be explained by the following spectral plots of the speech signals as shown in Figures 3.29 to 3.32. Most of the energy

of the car noise lies in the low frequency band (from Figure 3.29, the threshold is 500 Hz in this case), and thus the β-masking method alone is able to nicely reserve the spectral components above 500 Hz. Only the components below 500 Hz are difficult to be fully recovered. Since the spectral components in the β-masking-enhanced speech signals are very well matched to the clean speech, the two post-processing method becomes redundant and will introduce further distortion. E.g., in the NHR method, the re-synthesized high frequency band spectral components will be affected by the wrong information lies below 500 Hz. In the Autocorrelation method, although the harmonic information is well preserved, the magnitude becomes distorted. Therefore, it can be concluded that for car-noise-corrupted speech signals, the β-masking method alone works very well to recover the spectral information of the signal; the two post-processing methods are redundant in this case. Furthermore, the use of these two methods might cause more distortions than when the β-masking method is used alone.

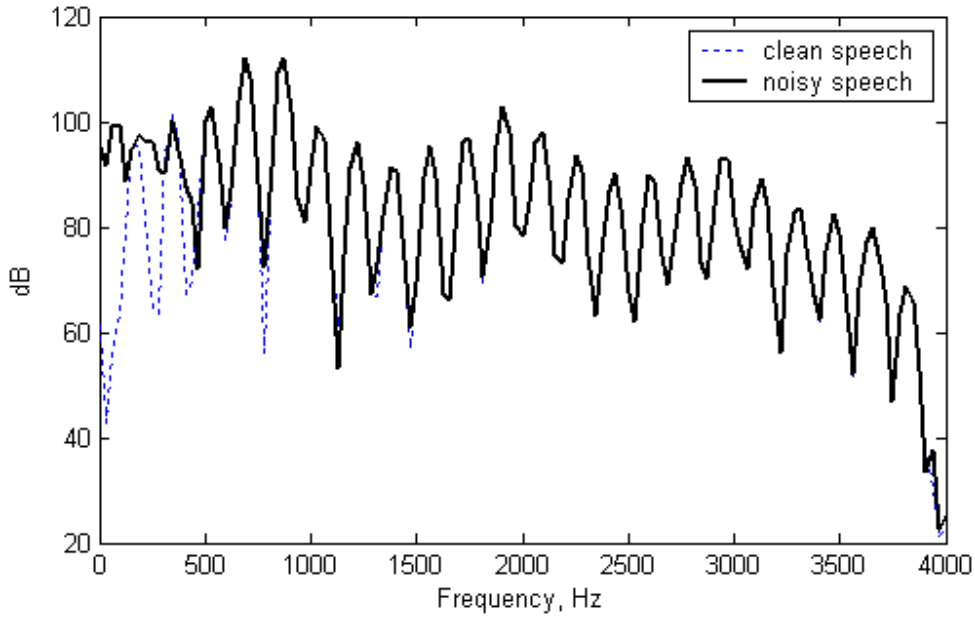

Figure 3.29: Frequency components of a car-noise-corrupted speech signal

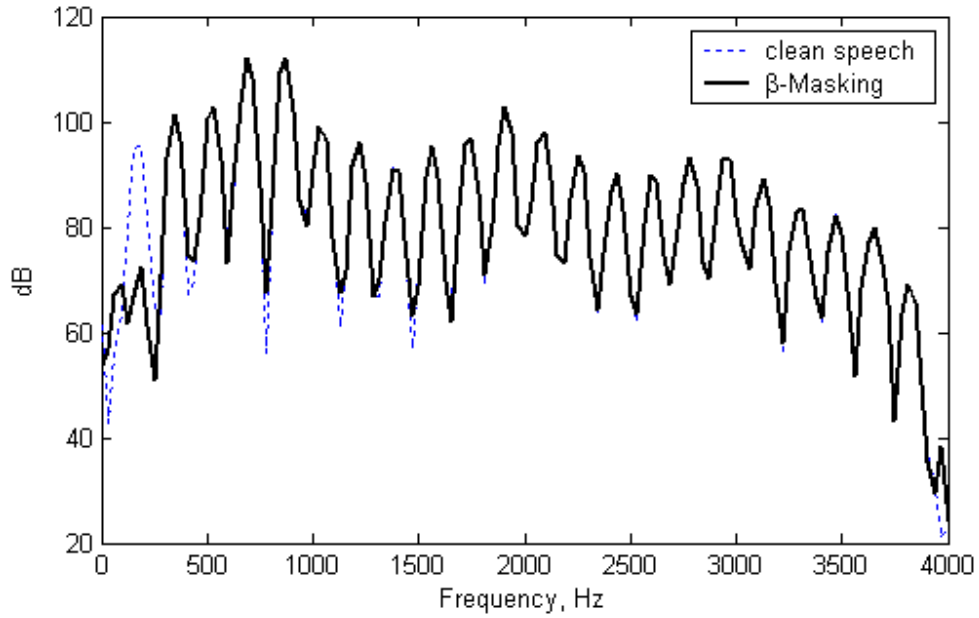

Figure 3.30: Frequency components of the β-masking enhanced speech signal

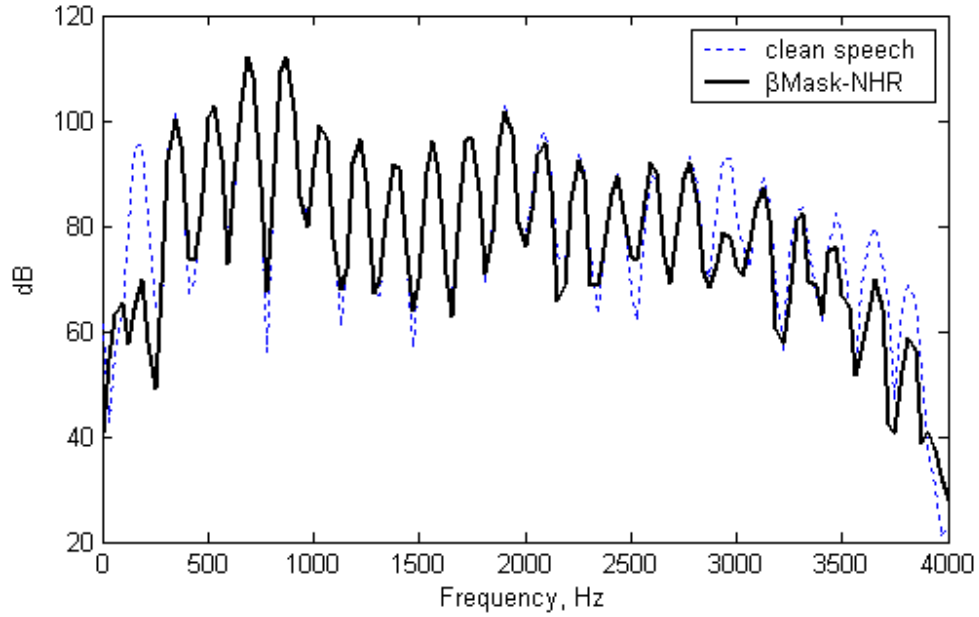

Figure 3.31: Frequency components of the βmasking-NHR enhanced speech signal

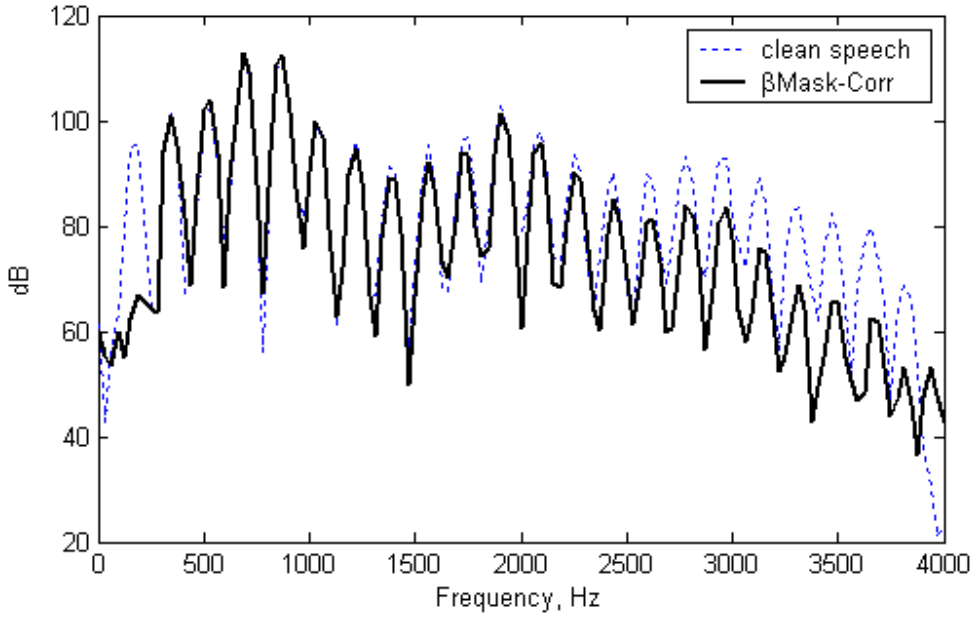

Figure 3.32: Frequency components of the βmasking-Corr enhanced speech signal

3.4 Conclusion

β-masking method is the *β*-order MMSE estimation method that incorporates the human auditory masking effects. It is good at removing the background noise while retaining most of the speech components. However, for cases where very high background noises are observed, the weak spectral components are likely to be totally masked by noise, and they are not recoverable by *β*-masking enhancement method alone. Therefore, two post-processing methods, the non-linear high-frequency regeneration and the cepstrum-aided autocorrelation methods are proposed in this chapter to reconstruct the lost voiced speech content.

The NHR algorithm tries to reconstruct the spectral details of the upper frequency band of a voiced frame by using the usually much stronger spectral energies of the lower-

band. We assume that the frequency content of the voiced speech is periodic, and by appropriately process the signal, we can spectrally "fold" the lower-band into the upper-band so as to reconstruct the periodicity in the upper-band. By repeating this periodic information, and restricting the magnitude of the reconstructed spectral components to the formant shape, we can expect to recover the distorted spectral components and at the same time to remove substantially the musical tones in the upper frequency band.

We can also regenerate the weak spectral components by using the autocorrelation of the voiced spectra. By shifting and adding the spectral magnitudes $n \times p$ points away, where n is an integer and p is the number of spectral points between any two peaks, we are able to reconstruct the periodicity of the over-attenuated weak spectral components of a voiced frame. In order to avoid local-maxima interference, cepstrum is used to help to obtain the correct periodicity. The re-synthesized components are then used to regenerate the improved β-masking gain, which has a comb-structure to be used to improve the quality of the synthesized speech.

Simulation results of the three algorithms, namely β-masking, βmasking-NHR and βmasking-Corr, show that the latter two methods are able to reconstruct the distorted spectral components that cannot be recovered by β-masking method alone. These two methods are suitable for white-noise, F16-noise, and pink-noise corrupted speech signals. However, for car-noise corrupted signals, the usage of these post-processing methods will introduce slight distortions to the spectral components recovered by the β-masking method alone.

CHAPTER 4

TWO-CHANNEL NOISE REDUCTION SYSTEM IN A CAR ENVIRONMENT

The β-masking algorithm has been shown to outperform many existing single channel speech enhancement methods. It shows its robustness by varying β under different *a-priori* and *a-posteriori* SNR conditions. However, a single-channel speech enhancement process may not give satisfactory performance under some very high background noise conditions, e.g., in a moving vehicle. One possible solution involves the idea of multi-channel speech enhancement in these cases to achieve a better performance. In this chapter, a two-channel noise reduction system to be used in a moving vehicle is studied. Cross-spectral coherence between signals of two-channel microphone array [44] will be utilized. In addition, a hybrid of the single and the two-channel speech enhancement method is then proposed. Noisy speech samples were real-time recorded in a moving car and were used in our simulation study. Subjective and objective comparison results show that the proposed hybrid speech enhancement system produces output with very low background noise and minimal speech distortion.

4.1 Introduction

The requirement for noise reduction in a car environment has increased with the development of hands-free mobile phone systems and many research works in this area have been reported. As highlighted earlier in the thesis, one of the most challenging problems of noise reduction is to attenuate the noise component of noisy speech as much as possible, while keeping the distortion to the speech components as low as possible. In a moving car environment, the noise level could be very high, and in addition, the characteristics of noise are time-varying and un-predictable. It therefore makes it difficult to reduce the noise components of a noisy speech signal recorded in a car environment as compared to normal conditions that have low and nearly stationary noise. As discussed in Chapters 2 and 3, the β-masking speech enhancement method might not be able to fully recover the spectral components that were swamped by noise. Therefore, the very robust single-channel speech enhancement method still has a less than satisfactory performance for a speech enhancement system used in a moving car environment. For some instances, high residual noise or noticeable speech distortions can be heard after the speech enhancement process for some cases of noisy speech signals. Also, some weak spectral components, especially in the high frequency bands, are sometimes over attenuated and result in speech distortions. The performance curve of the single-channel speech enhancement seems to plateau and even the best performing technique is still far from perfect. This leads to the idea of using a multi-sensor approach which is possible for certain applications. The most explored in this direction is the use of two microphones to implement a speech enhancement system.

Besides post-processing, increasing the number of microphones could possibly be used to improve the performance of the β-masking method. One advantage of using a two-channel speech enhancement system over a single-channel system is that the cross-correlation between the two microphone inputs can be incorporated into the filter gains [45]. A. Guérin *et al* proposed a two-sensor noise reduction system for a hands-free car kit. They exploit the fact that the noise inputs to the two microphones spaced at 80cm in a car are not correlated for frequencies above 210 Hz [33]. The two-sensor filter gain, $H_{css}(k)$, at frequency k is a function of the coherence of the noisy speech signals of the two channels. The two-sensor approach is good at retaining the speech components; however, the background noise of the output speech signals remains high in general. The reason is that, although noises of the two-channel input signals are not correlated above 210Hz, the speech signals are correlated. Thus the gain values at the spectral valleys are not low enough to remove the noise components in the original two-sensor approach.

In order to improve the performance of the two-sensor noise reduction system, it is proposed in this chapter that the β-masking technique be incorporated into the filter block of the two-sensor enhancement systems. Experimental results showed that the proposed approach achieves a better performance in terms of noise reduction and speech signal retention. Subjective and objective comparisons were performed for the β-masking, two-sensor and the combined βmasking-TwoSensor schemes. In addition, the noise reduction technique used in the ETSI (European Telecommunications Standard Institute) [46] Standard is also used as a performance benchmark.

4.2. Three-channel recording experiment set up

Many existing speech databases provide noisy speech signals recorded in a moving car environment. Single-channel, two-channel or even eight-channel noisy speech signals can be found in some of these databases which are mainly used for speech recognition research. Since the original clean speech is not provided in these databases, it is impossible to perform objective quality assessment for the enhanced speech. They are therefore not suitable for our use.

One possible way to obtain both clean and noisy speech signals is to playback the clean speech through a loud speaker in a moving car. At the same time, a pair of microphones are used to record the speech which is corrupted by high background noise. However, the properties of the original clean and noisy speech signals will be generally different due to the additional distortions introduced by the audio card, the loud speaker and the microphone. Therefore, it is necessary to obtain the desired speech database by recording real speech under the same condition. A three-channel recording system was set up in a moving car environment: the driver is driving while talking. In our study, we assume that only one person, the driver, is talking in a noisy car. There are three microphone inputs: x_0 is from the headset worn by the driver, which records a relatively clean speech; the other two, x_1 and x_2, are from the microphones mounted on the dashboard at 80 cm apart, and they record the noisy speech signals, as shown in Figure 4.1. The car was moving at a speed of 60 km/h with windows closed.

The speech signals recorded using the headset (x_0) are reasonably noise free so that they can be considered as the reference clean speech signals ($\tilde{s}(t)$) for performance

evaluation. Note that x_0 only plays the role of referencing, and it is not used in the speech enhancement process. The two noisy inputs, x_1 and x_2, are mainly degraded by the noises from the engine, the contact between the tires and the road, the vibration of the dashboard and the noise from the environment outside the car. The vibration noise could be minimized by setting the microphones on a piece of securely-mounted foam. About one hundred utterances from the driver were recorded, each having a duration of about five to six seconds. A sampling rate of 16 kHz was used to record the noisy utterances.

Speech signals can also be badly degraded by the noise of cars passing by, or by the wind when the windows are wound down. These noises add much difficulties to noise-level estimation. In order to examine the performance of the speech enhancement algorithm under different conditions, the recording experiment is separated into two parts. In the first part of recording, the windows were closed and no cars were passing by during speech recording. In the second part of recording, the noise level was increased by increasing the driving speed, opening windows and recording when there were cars passing by.

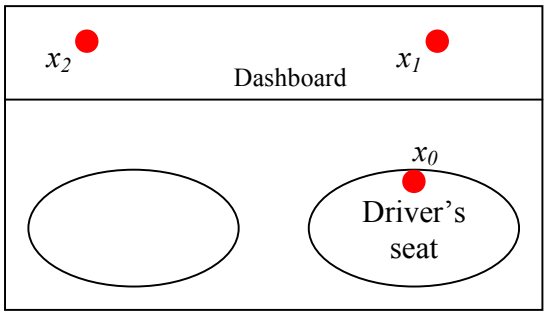

Figure 4.1: Three channel recording set up

4.3. Speech and noise characterization in a two-sensor system

For the two-sensor system, the noisy speech signals can be expressed as:

$$x_1(t) = s_1(t) + n_1(t), \quad x_2(t) = s_2(t) + n_2(t) \tag{4.1}$$

where $s_1(t)$ and $s_2(t)$ denote the clean speech signals of the two channels, $n_1(t)$ and $n_2(t)$ are the additive noise, $x_1(t)$ and $x_2(t)$ are the noisy signals recorded, respectively . $s_1(t)$ and $s_2(t)$ are highly correlated, as they contain the same speech information as the reference speech signal $\tilde{s}(t)$.

In their paper, A. Guérin *et al* use the magnitude squared coherence (MSC) to study the correlation between the two channels of speech signals:

$$MSC(k) = \left|\rho(k)\right|^2, \quad \rho(f) = \frac{\gamma_{x1x2}(k)}{\sqrt{\gamma_{x1}(k) \cdot \gamma_{x2}(k)}} \tag{4.2}$$

where $\gamma_{x1}(k)$ and $\gamma_{x2}(k)$ denote the power spectral density (PSD) of the two noisy speech signals and $\gamma_{x1x2}(k)$ is the observations' cross-PSD, at k-th spectral position.

The signal PSD and cross-PSD are defined as:

$$\gamma_{xi}(k,p) = E[X_i(k,p) \cdot X_i^*(k,p)], \quad i = 1,2 \tag{4.3}$$

$$\gamma_{x1x2}(k,p) = E[X_1(k,p) \cdot X_2^*(k,p)] \tag{4.4}$$

where $X_i(k,p)$ is the k-th spectral component at the p-th frame of a signal $x_i(t)$. The frame length used is usually 256 or 512 with an overlapping ratio of 0.5 or 0.75. For real-time applications, the PSD and cross-PSD are estimated using recursive filtering, i.e.,

$$\gamma_{xi}(k,p) = \lambda\gamma_{xi}(k,p-1) + (1-\lambda)X_i(k,p)X_i^*(k,p), \quad i = 1,2 \tag{4.5}$$

$$\gamma_{x1x2}(k,p) = \lambda\gamma_{x1x2}(k,p-1) + (1-\lambda)X_1(k,p)X_2^*(k,p) \tag{4.6}$$

where λ is a smoothing factor closed to 1. It takes small values during speech presence periods to reduce the reverberant effect, and high values during noise-only periods to minimize the fluctuating residual musical noise. It is achieved by using the following equation [33]:

$$\lambda(k,p) = 0.98 - 0.3\frac{SNR(k,p)}{1+SNR(k.p)} \tag{4.7}$$

A. Guérin *et al* proposed their two-sensor algorithm as shown in Figure 4.2. Firstly, the two channel speech signals are transformed into the frequency domain, and band pass filtered (300-3400 kHz for telephone applications). The filter should have a sharp cut-off at 300Hz so as to remove the highly correlated noise components in the low frequency band. Next, the attenuation filter gain $H_{css}(k,p)$ is calculated as:

$$H_{css}(k) = \frac{|\gamma_{x1x2}(k)| - |\gamma_{n1n2}(k)|}{\sqrt{\gamma_{x1}(k)\gamma_{x2}(k)}} \tag{4.8}$$

where $\gamma_{n1n2}(k)$ is the theoretical noise cross-PSD. This zero-phase filter $H_{css}(k,p)$ is multiplied with $X_1(k,p)$ to obtain the estimated output speech $\hat{S}_1(k,p)$. The $\gamma_{x1}(k)$, $\gamma_{x2}(k)$ and $\gamma_{x1x2}(k)$ are calculated using equation (4.4), and the noise cross-psd $|\gamma_{n1n2}(k)|$ is over-estimated by the mean noise PSD $\sqrt{\gamma_{n1}(k)\gamma_{n2}(k)}$, i.e.,

$$\sqrt{\gamma_{n1}(k,p)\gamma_{n2}(k,p)} = \alpha\left(\widetilde{SNR}_{post}(k,p)\right)\sqrt{\gamma_{n1}(k,p-1)\gamma_{n2}(k,p-1)} \tag{4.9}$$

where $$\alpha\left(\widetilde{SNR}_{post}\right) = L + (1-L)\cdot\frac{1}{1+1/\left(g\cdot\widetilde{SNR}_{post}\right)}\cdot\left(1+\frac{1}{1+g\cdot b\cdot\widetilde{SNR}_{post}}\right) \tag{4.10}$$

and $$\widetilde{SNR}_{post}(k,p) = \frac{|X_1(k,p)X_2(k,p)|}{\sqrt{\gamma_{n1}(k,p-1)\gamma_{n2}(k,p-1)}} \tag{4.11}$$

L=0.9, b=0.5 and g=1/(1-b) is used in this experiment.

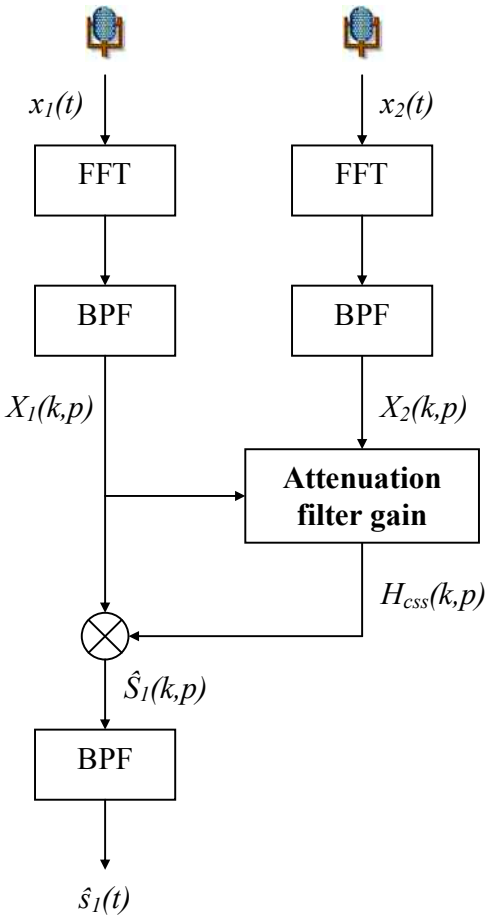

Figure 4.2: Block diagram of a two-sensor noise reduction system

The two-sensor enhanced speech signals are found to retain almost all the speech components. However, as a trade off, the residual noise is high, especially at the beginning of a speech utterance where the mean noise PSD estimation has yet to fully adapt to the correct value. The high residual noise is also due to the fact that the filter gain depends mainly on the coherence of the two-channel inputs, and thus the gain values at harmonic valleys are not low enough to suppress the noise components. In order to further suppress the residual noise, the single-channel speech enhancement method, namely β-masking proposed by C. H. You *et al*, is incorporated into the two-sensor system, as the algorithm is very effective in noise reduction.

4.4. Two-sensor noise reduction incorporating β-masking

As discussed in Chapter 3, β-masking is a single channel speech enhancement method, giving an output with very low background noise. However, some speech spectral components are over-attenuated and occasional distortions could be detected. In the case where there are two microphone inputs, we can make use of the spatial coherence of the speech signals of two channels. In their early research, R. Bouquin-Jeannès *et al.* used a cross-spectral estimator for noise reduction [44]. The noisy speech signals are passed though a filter equal to the magnitude of the coherence function:

$$H_c(k,p) = \frac{\hat{\gamma}_{ss}(k,p)}{\sqrt{\gamma_{x1}(k,p)\gamma_{x2}(k,p)}} \tag{4.12}$$

where $\hat{\gamma}_{ss}(k,p)$ is an estimate of $\gamma_{ss}(k,p)$, the PSD of a speech signal $s(t)$. The purpose is to transmit the spectral components containing speech and turn off those for which speech is

absent. Assuming speech and noise signals are totally uncorrelated, we can estimate the PSD of a speech signal by taking the magnitude:

$$\hat{\gamma}_{ss}(k,p) = |\gamma_{x1x2}(k,p)| - \bar{\gamma}_{n1n2}(k) \tag{4.13}$$

where $\bar{\gamma}_{n1n2}(k)$ represents the last estimate of the cross power spectrum $\gamma_{n1n2}(k,p-d)$ before speech activity, and the *(p-d)*-th frame is taken to be the last estimate of noise before the speech frame *p*. By approximating $\bar{\gamma}_{n1n2}(k)$ using the magnitude of the noise cross-PSD $|\gamma_{n1n2}(k)|$, the filter becomes the same as the two-sensor filter $H_{css}(k)$, i.e.,

$$H_c(k) = \frac{|\gamma_{x1x2}(k)| - |\gamma_{n1n2}(k)|}{\sqrt{\gamma_{x1}(k)\gamma_{x2}(k)}} = H_{css}(k)$$

In our study, we weight the β-masking enhanced speech by the filter $H_c(k)$, which is proportional to the coherence of the two-channel inputs. Therefore, passing the β-masking enhanced speech signals through filter $H_c(k)$ is functionally equivalent to cascading the β-masking gain filter $G_{\beta p}(k)$ with the two-sensor filter $H_{css}(k)$. Hence the whole process is as shown in Figure 4.3.

Cascading of two filters is the same as multiplying the two gains in the frequency domain. Both filters are of zero-phase and the output (the estimation of first channel speech signal) becomes

$$\hat{S}_1(k) = H_{css}(k)G_{\beta p}(k)X_1(k) = G_{new}(k)X_1(k) \tag{4.14}$$

Figure 4.4 shows the spectral plots of a frame of noisy and clean speech, and the corresponding two-sensor gain ($H_{css}(k)$), the β-masking gain ($G_{\beta p}(k)$), and the new gain (multiplication of the two gains, $G_{new}(k)$). This figure shows how $G_{\beta p}(k)$ can be constrained and improved by $H_{css}(k)$ to obtain a new gain $G_{new}(k)$.

One common feature of $H_{css}(k)$ and $G_{\beta p}(k)$ is the harmonic-structure, which has high values at the harmonic peaks, and low values at the valleys. The value of $H_{css}(k)$ seldom falls below 0.5. This causes the ineffectiveness in noise removal in the two-sensor approach. In contrast, the comb-structure sometimes does not appear (as indicated by the circles in Figure 4.4) in $G_{\beta p}(k)$, and the value of the gain might also be greater than unity. This causes the over-amplification effect, which is especially noticeable in the form of musical tones when it happens at the high frequency region. The multiplication of the two gains, $G_{new}(k)$, can combine the characteristics of both of them. The resulting gain curve retains the comb-structure which is not apparent in $H_{css}(k)$. In addition, almost all the gain values are kept under unity, which reduces tonal distortions in the output signals as compared to the signals processed by the β-masking method. Most importantly, for spectral peaks, the gain values are kept close to unity and for spectral valleys, the gain values are smaller than $G_{\beta p}(k)$ and $H_{css}(k)$. In a noisy speech signal, the speech components in the spectral valleys are usually swamped by noise. The values of $G_{new}(k)$ around the spectral valleys are low enough to effectively attenuate the noise components, resulting an output signal with little background noise.

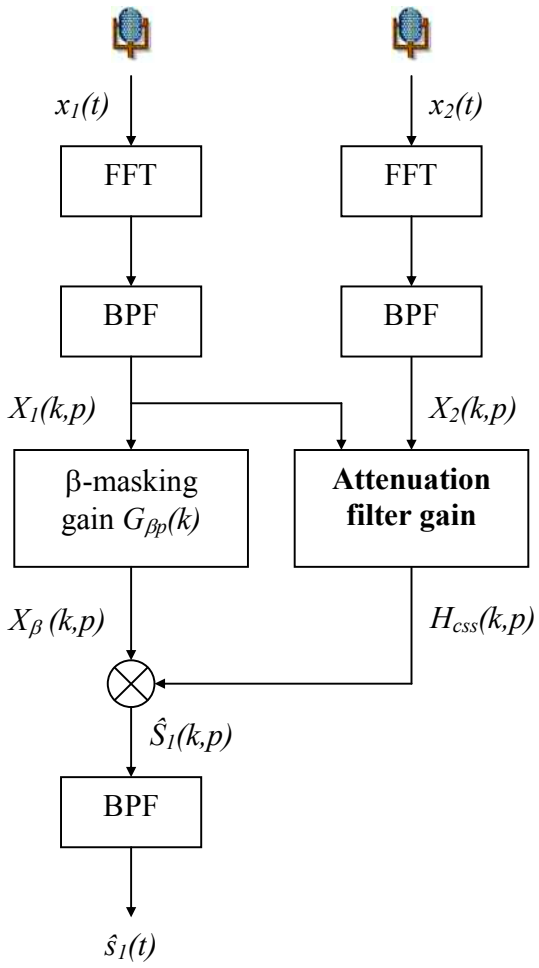

Figure 4.3: Two-sensor noise reduction system incorporating β-masking

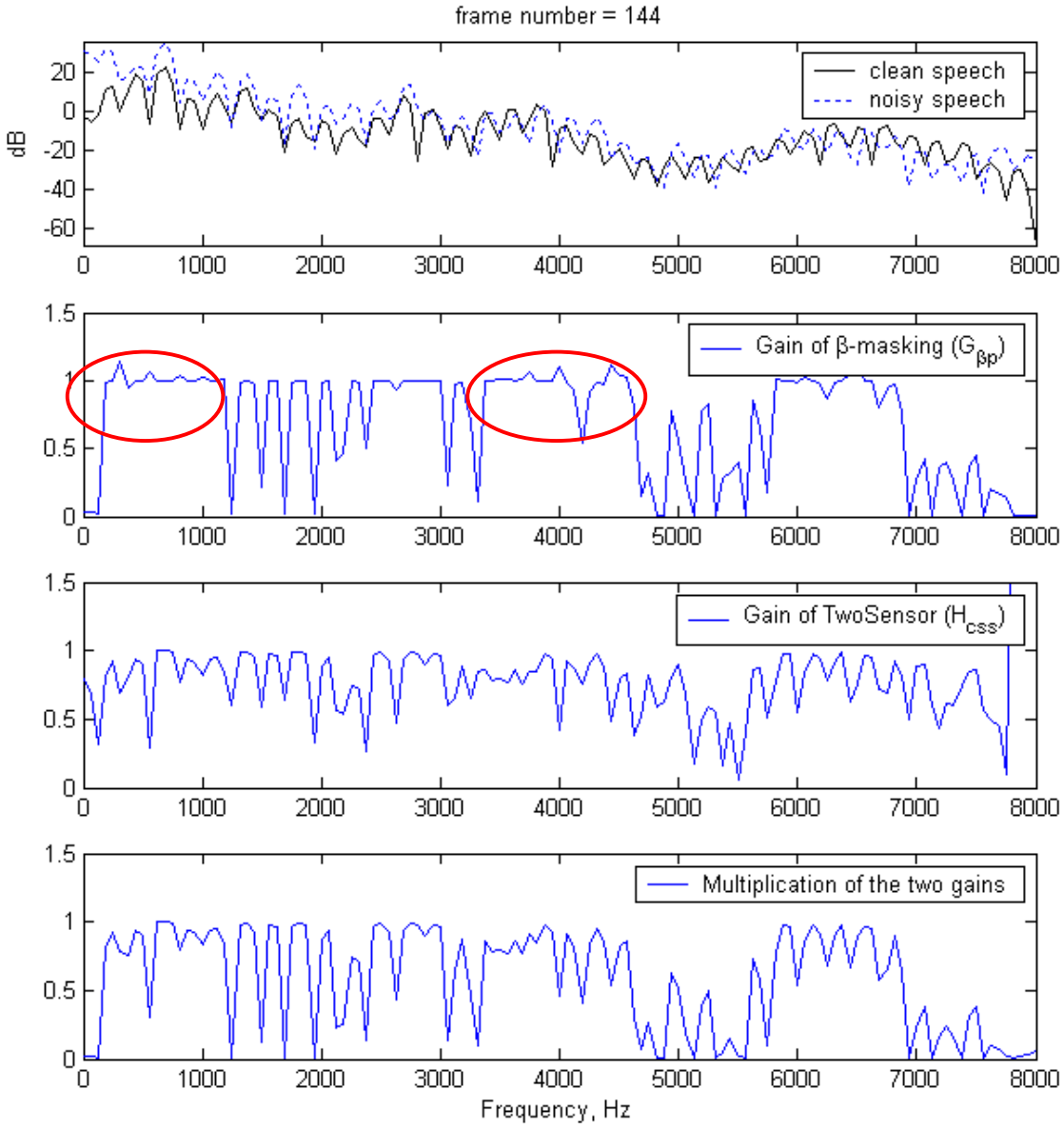

Figure 4.4: (1) The spectral plot of a frame of noisy and clean speech and the corresponding frequency domain filter Gains for (2) β-masking approach, (3) Two Sensor approach and (4) Multiplication of the two gains

4.5. The noise reduction technique used in ETSI standard

In our study, the noise reduction technique used in the ETSI [46] standard is also taken for performance comparison. The complete ETSI standard is developed mainly as a speech recognition front end. The part used here is their noise reduction approach before speech feature extraction. Here "ETSI" is used to denote the noise reduction technique. It is a two-stage Mel-wrapped Wiener filter approach, which is depicted in Figure 4.5.

In the first stage, the spectrum is estimated and the PSD mean is calculated to derive the frequency domain Wiener Filter (WF) coefficients. The calculation of the WF coefficients also involves noise estimation, which comes from a voice activity detector (VADNest stands for Voice Activity Detection used for Noise estimation). Next a Mel filter bank is used to smooth the linear WF coefficients. The impulse response of this filter is obtained by applying a Mel Inverse Discrete Cosine Transform (Mel-Wrapped IDCT).

The output signal from the first stage is the input signal to the second stage and the second stage is basically the same as the first stage except with the addition of gain factorization to control the effect of noise reduction. At the end of the second stage (or the end of the whole noise reduction process), the DC offset is removed. The detailed functionality and formulation of each block is given in [46].

Simulation results show that the ETSI standard removes most of the undesirable noises in a noisy car environment. However, some low level "trembling noise" is introduced in the processed signals. Detailed comparisons will be given in the following section.

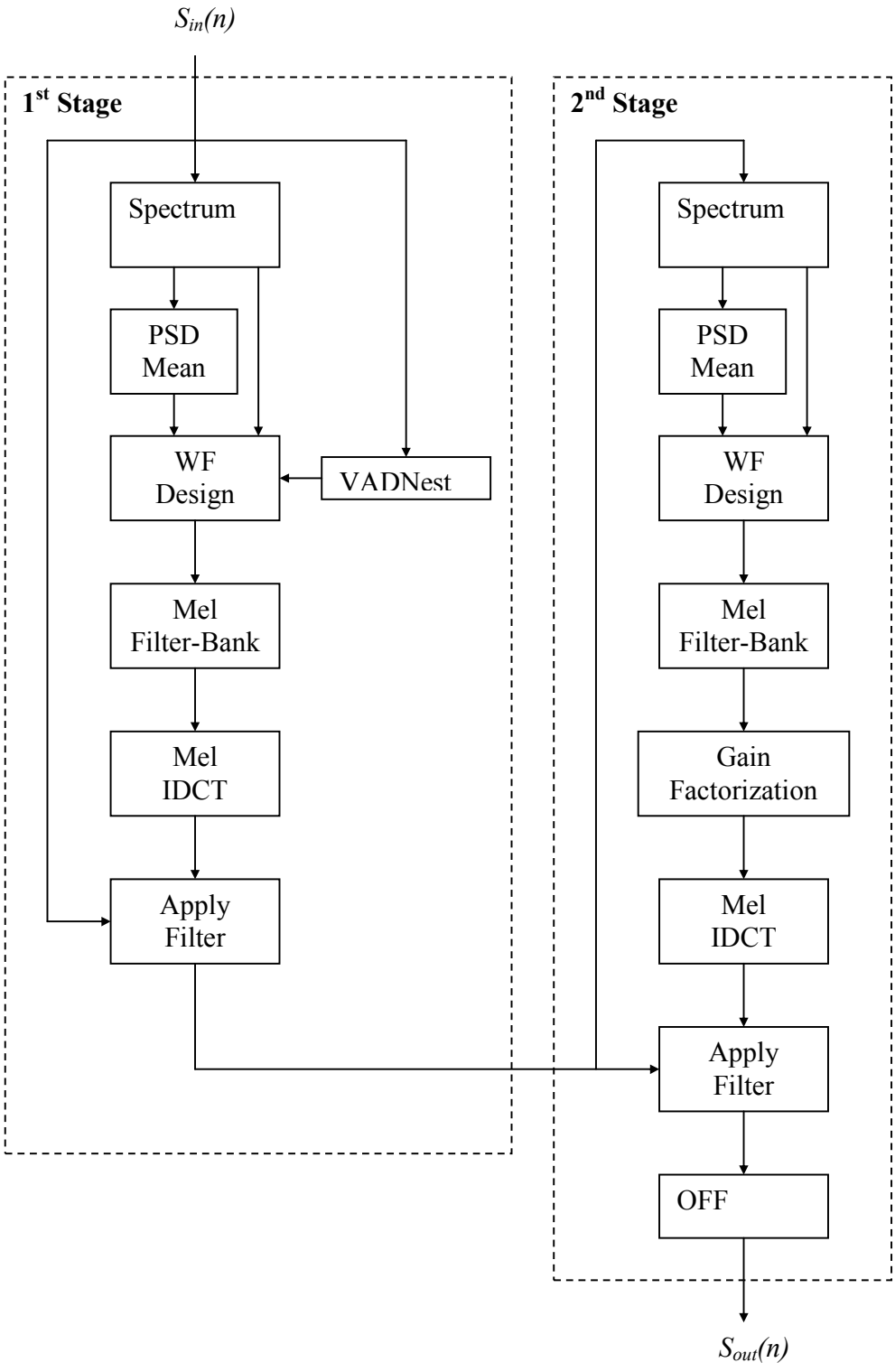

Figure 4.5: Block diagram of noise reduction in the ETSI standard

4.6. Performance Comparisons

As stated in section 4.2, the recording experiment has two parts. The comparisons of the different speech enhancement algorithms are separated into two sections so that we can examine the performance of the algorithms under different circumstances.

4.6.1 Data Recorded with Low Level of Noise

In the first part of the recording experiment, the speech signals were recorded without cars passing by, and the windows were closed. The noises that corrupt the speech signal are kept to the lowest possible level. Sixteen sets of clean and noisy speech signals were used for simulation and performance comparison.

Figure 4.6 shows the waveforms of one set of clean speech, noisy speech and the enhanced speech (two-sensor, β-masking, ETSI and βmasking-TwoSensor) signals. The first waveform is the "quasi-clean" speech signal recorded by the headset microphone and some very slight background noise can be heard in this signal. However, the noise level is very low which is quite acceptable subjectively. Therefore, this waveform is used as the reference for taking objective measurements such as SNR. The average segmental SNR of the unprocessed two-channel inputs is -12 dB.

The ETSI-enhanced speech signals have a relatively high SNR of 0.26dB. However, some low level "trembling noise" can be heard. The two-sensor enhanced speech signals have a low level of speech distortion, but with high background noise retained, especially for the first few frames where the noise level and noise cross-PSD

have not yet been fully adapted. In contrast, the β-masking enhanced speech has an overall lower level of background noise, but with some tonal distortions observed in the processed speech. The combination of the two algorithms gives a low background noise and low speech distortion, and the reason is that β-masking gain always takes low values for noise frames. Experimentally, the βmasking-TwoSensor enhanced speech signals sound very close to the clean speech.

The improvement in noise reduction resulting from the use of the discussed speech enhancement algorithm can be observed by examining the segmental SNR measurements shown in Figure 4.7 and Table 4.1. The proposed βmasking-TwoSensor method has an SNR improvement of more than 0.8dB as compared to ETSI, more than 2dB as compared to the β-masking method and more than 6dB compared to the original two-sensor method.

The general psycho-acoustically motivated distortion measure (PMD) is repeated here for readers' convenience:

$$\Xi_{seg} = \frac{1}{L_s}\sum_{m=0}^{Ls-1} 10\log_{10}\left(\frac{\sum_{k=0}^{w/2+1}\left(S_k - \widehat{S}_k\right)^2}{\sum_{k=0}^{w/2+1} S_k^p}\right) \tag{4.16}$$

where w denotes the window size, m denotes the frame index ranging from the first to the last (L_S) frames; S_k and $\widehat{S}_k$ are the k-th spectral components of the clean and processed speech signals, respectively. For p=2, the distortion measure is similar to the distortion measure proposed by Itakura for comparing two autoregressive speech models. The comparison result is as given in Figure 4.8 and Table 4.2. It shows that the proposed βmasking-TwoSensor enhanced speech signals have much less spectral distortion than those enhanced by β-masking alone.

It is observed from Figures 4.7 and 4.8 that, although the segmental SNR evaluation shows that the ETSI algorithm performs better than β-masking, the PMD evaluation shows that the ETSI performance is poorer than β-masking. The way of calculating PMD is very similar to the inverse of SNR measured in the frequency domain. By listening to the ETSI-enhanced and the β-masking enhanced speech signals, it is found that the PMD measurement results are closer, than segmental SNR, to the subjective perceptual quality of the enhanced speech.

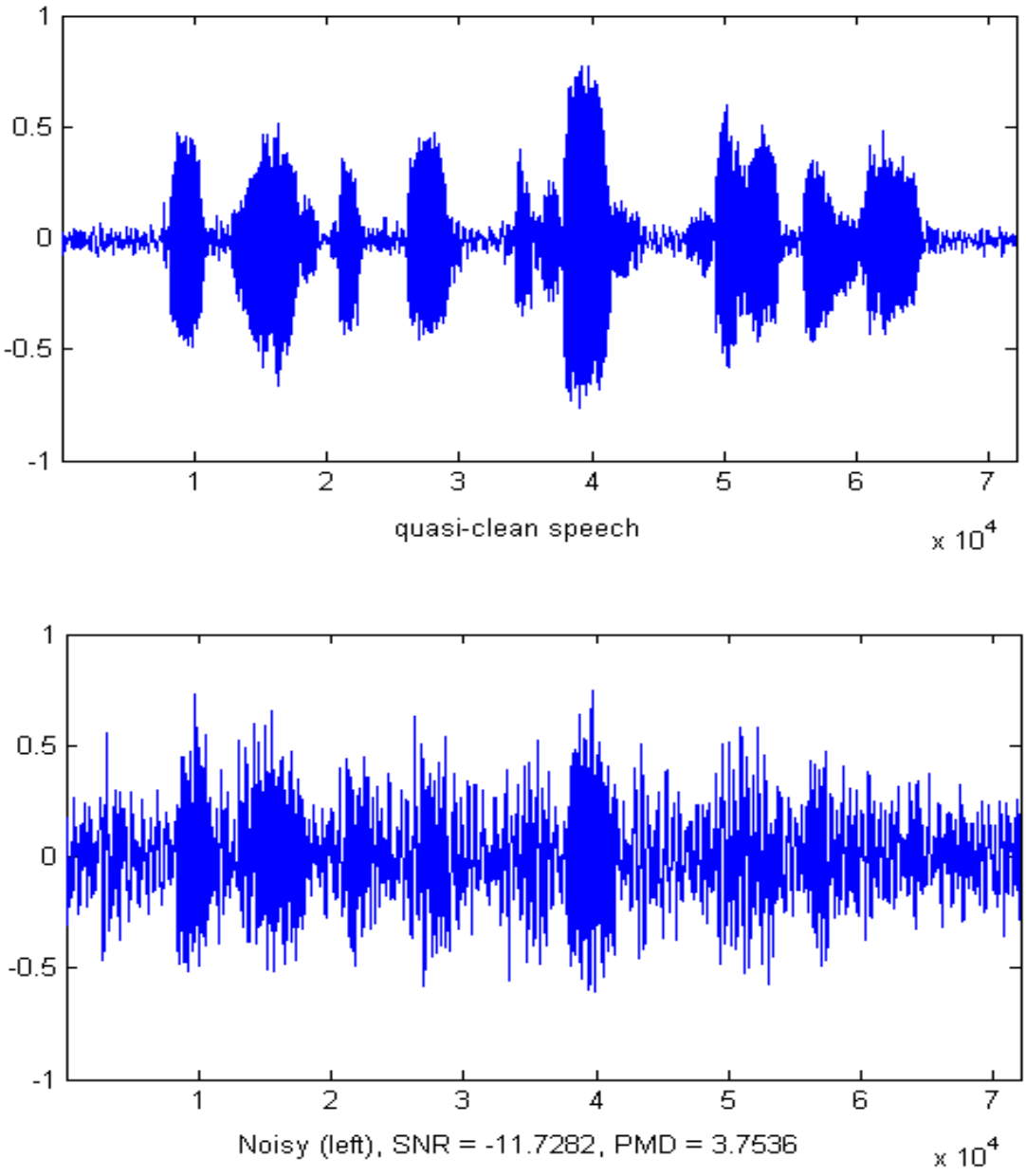

Figure 4.6 (1): Waveform of the quasi-clean speech

Figure 4.6 (2): Waveform of Noisy speech (left channel)

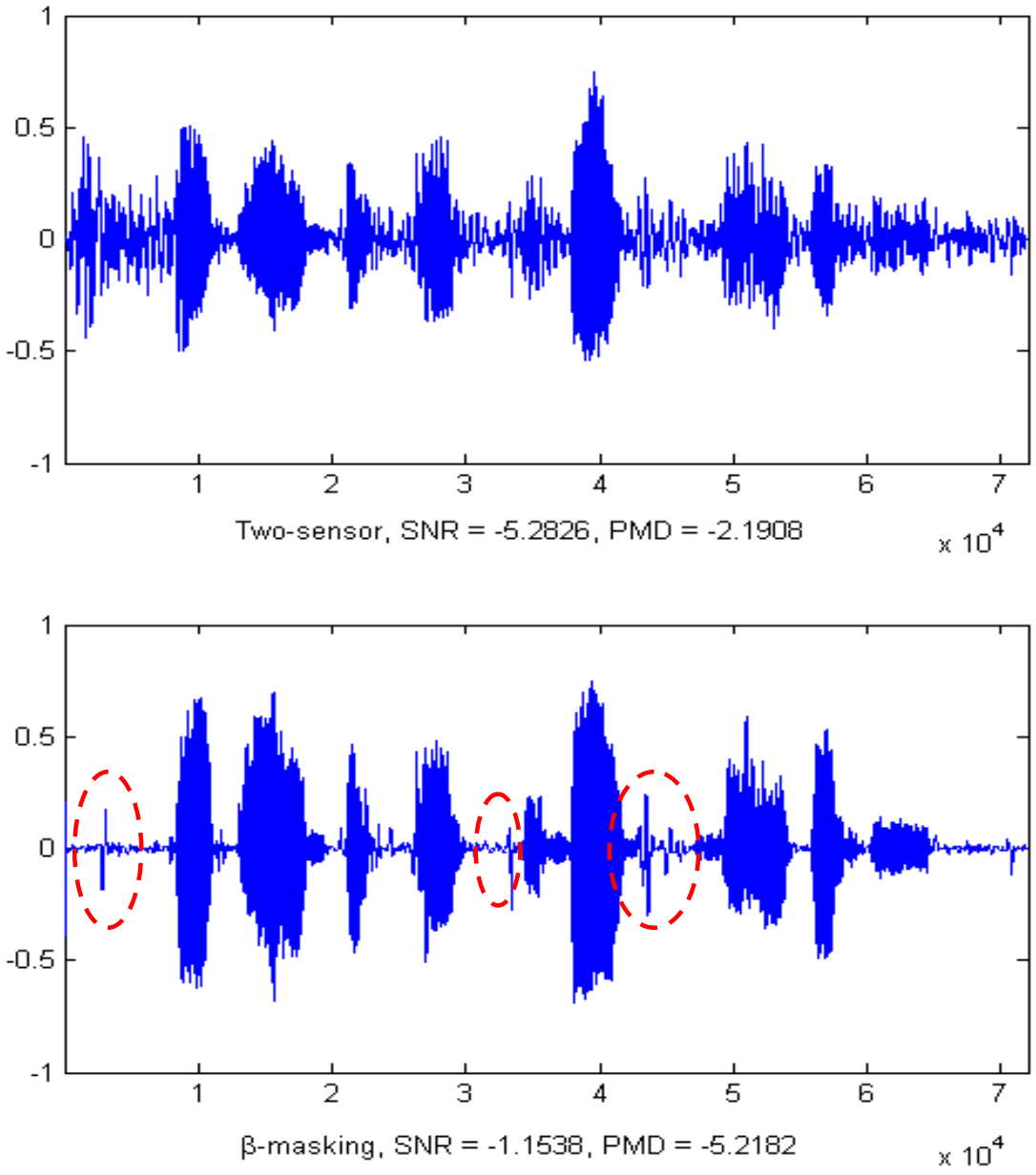

Figure 4.6 (3): Waveform of the speech enhanced by two-sensor approach

Figure 4.6 (4): Waveform of the speech enhanced by β-masking approach

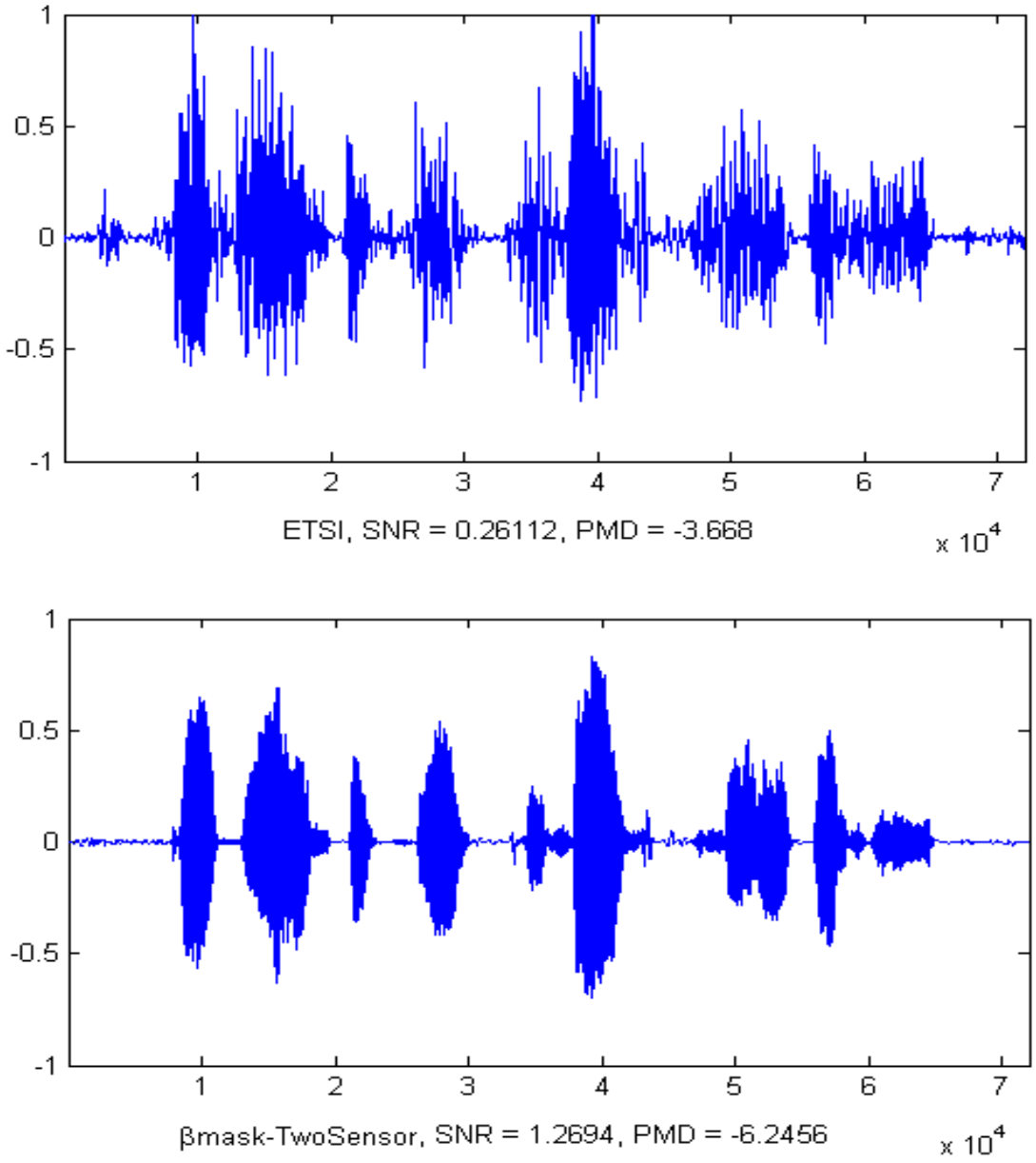

Figure 4.6 (5): Waveform of the speech enhanced using the ETSI Standard

Figure 4.6 (6): Waveform of the speech enhanced by βmask-TwoSensor approach

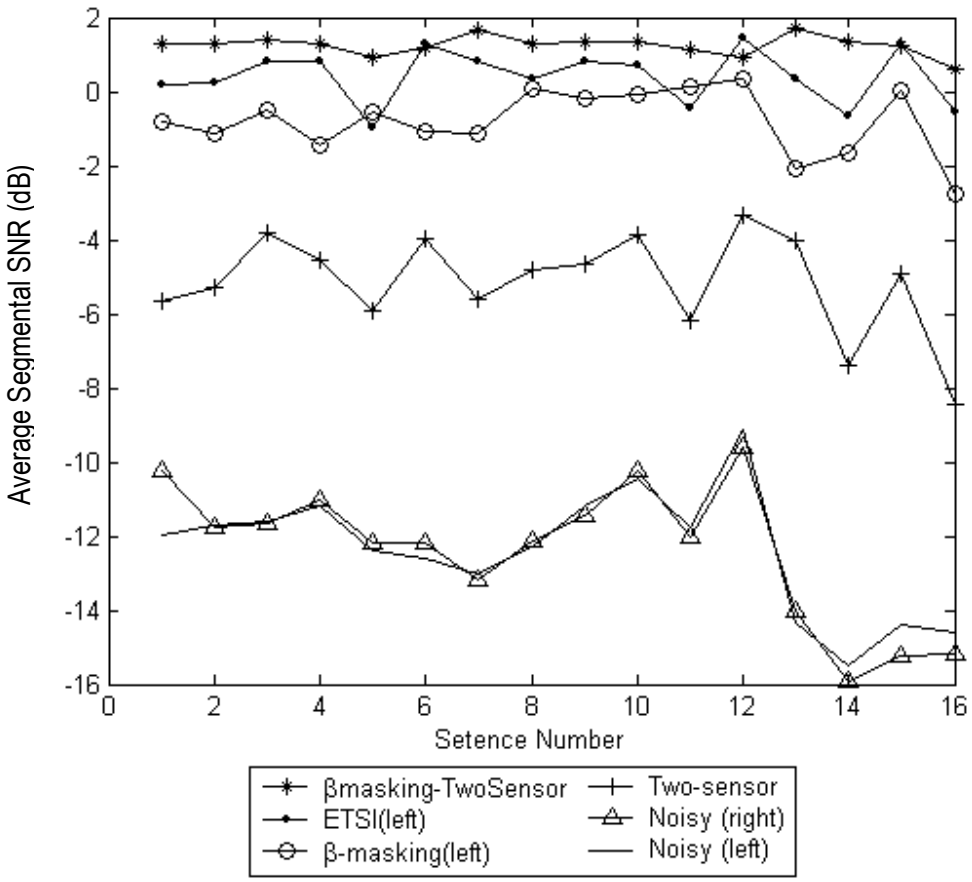

Figure 4.7: Segmental SNR measurement for speech utterances with low level of noise

Speech enhancement method	**Average SNR Value**
*β*masking-TwoSensor	1.25
ETSI (left channel)	0.41
β-masking (left channel)	-0.81
Two-sensor	-5.16
Noisy (right channel)	-12.38
Noisy (left channel)	-12.39

Table 4.1: Average Segmental SNR measurement (low level of noise)

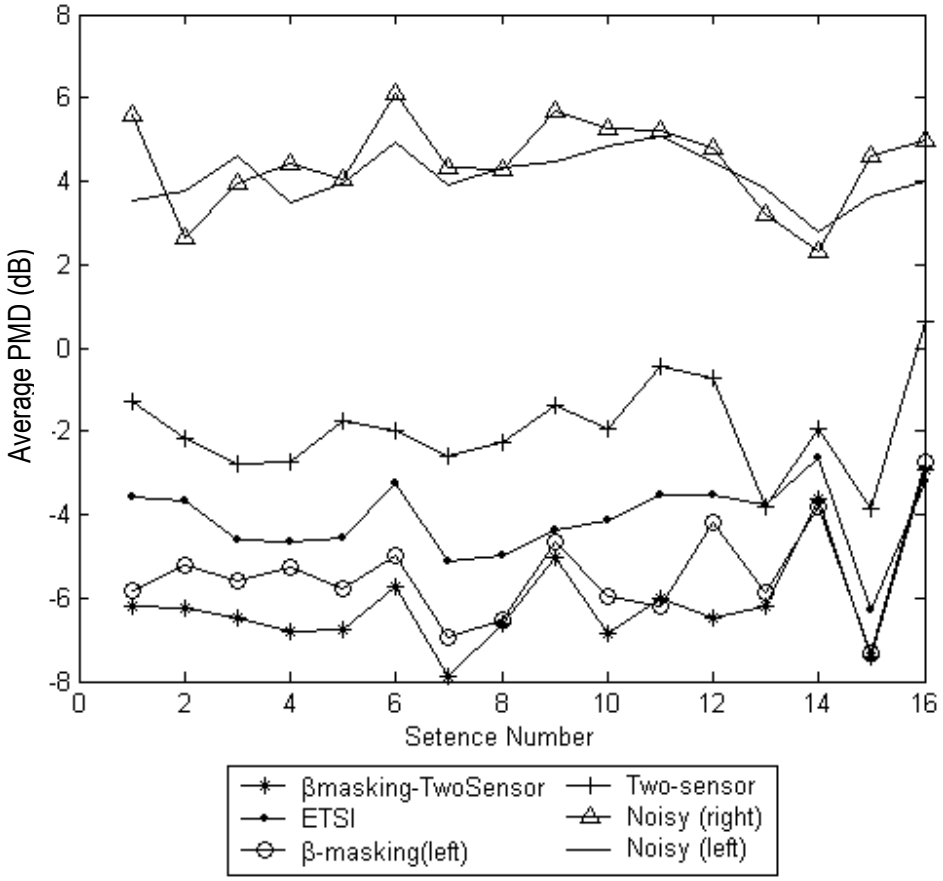

Figure 4.8: PMD measurement for speech utterances with low level of noise

Speech enhancement method	**Average PMD Value**
*β*masking-TwoSensor	-6.1
ETSI (left channel)	-4.13
β-masking (left channel)	-5.43
Two-sensor	-1.95
Noisy (right channel)	4.47
Noisy (left channel)	4.09

Table 4.2: Average PMD measurement (low level of noise)

4.6.2 Data Recorded with Higher Level of Noise

In the second part of the recording experiment, thirty sets of speech utterances were used. This time, the noise source is more complex than that of the first part. The driving speed was increased, car window (right side) was wound down, and the speech signals were recorded when there were cars passing by. In the first part, the noise levels of the two channel inputs are about the same as they have very close segmental SNR values of -12dB. However, in the second part, the noise levels between the two channels are greatly different. The right channel speech signals have a much lower SNR than the left channel speech signals which have segmental SNRs of between -14dB and -10dB. This is because other cars were passing by on the right hand side and it badly degraded the right channel input signals. Note that the left channel speech signals of the second part have an average segmental SNR of -10dB while those of the first part have an average segmental SNR of -12dB. A higher SNR in the second part does not mean that the noise level in the second part is lower. In fact the noise level is much higher and the noise characteristics are much more complex. The higher SNR value was due to the fact that the driver talked much louder in the second part of the experiment and thus increasing the numerical SNR value of the noisy speech signals. This is normal because people tend to talk louder when the environment suddenly becomes noisy.

As the noise becomes more complex with short transient noise than the first part, the noise level becomes more difficult to estimate accurately. Nevertheless, the proposed hybrid algorithm still shows its advantage in improving the segmental SNR and decreasing PMD, as shown in Figures 4.9, 4.10 and Tables 4.3, 4.4. The amount of improvement is still quite significant.

Again, comparisons of simulation results using the set of recorded noisy utterances show that the proposed hybrid algorithm, i.e., βmasking-TwoSensor, is suitable to be a noise reduction scheme under different conditions for a hands-free communication car kit.

4.7. Conclusion

In this chapter, a two-sensor noise reduction technique is studied, and it is used to enhance speech signals recorded in a moving vehicle where the noise level is generally very high. The attenuation filter used in the two-sensor enhancement method is a function of the spectral coherence between the two-channel noisy speech signals. The output speech signal enhanced by the two-sensor method has very little speech distortion. However, the filter gain value is not low enough to suppress the noise to a reasonably low level.

Compared to the two-sensor method, the performance curve of the single-channel β-masking speech enhancement method seems to be the opposite. The noise level in the processed speech signals is low, but the speech signals are more distorted. As the background noise level is high in a moving vehicle, the weak spectral components are swamped by the noise, and they are likely to be over-attenuated. Also, for some instances, over-amplification of some spectral components occurs during speech period, which leads to slight tonal distortions.

After studying the characterizations of the speech signals enhanced by the β-masking method and the two-sensor method, the idea of combining these two methods was considered. Passing the β-masking enhanced speech signals through a coherence-

weighting filter is equivalent to multiplying the two gains in the frequency domain. Multiplying the two gains can improve and constrain each other. It can recover the harmonic-structure to a certain extent and reduce the over-amplification of some of the spectral components due to the β-masking gain.

The experimental data used in our simulations were real-time recorded in a moving car. The recording experiment has two parts. In the first part, speech signals were recorded under the condition that the noise level was low; the windows were closed and no car passed by during recording. In the second part, speech signals were recorded in a much noisier environment; the windows were wound down and the recording was done when there were cars passing by. Subjective and objective comparisons show that the proposed hybrid method can remarkably improve the output speech quality for both conditions. Objectively, the speech signals processed by the βmasking-TwoSensor method have a significant SNR improvement as compared to the individual methods. Subjectively, the output speech signals processed by the βmasking-TwoSensor method have very low background noise and very little tonal distortions. It is therefore concluded that the proposed βmasking-TwoSensor enhancement method is a very effective method for enhancing speech signals corrupted by noise in a moving car environment.

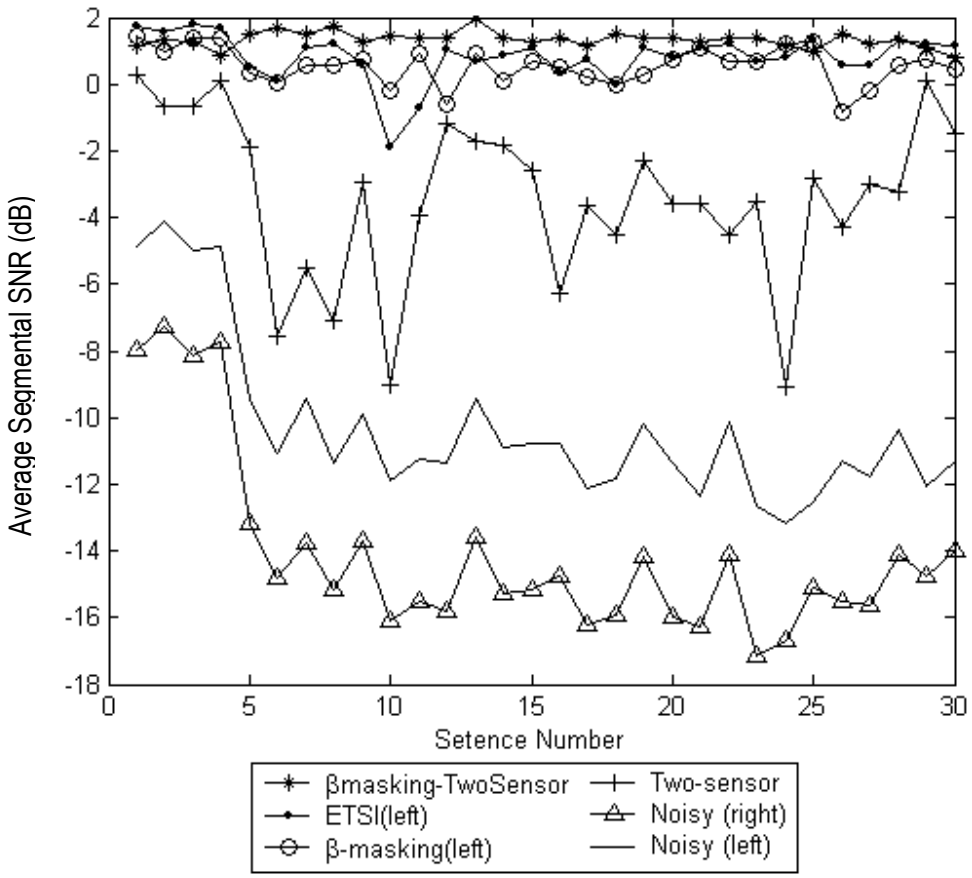

Figure 4.9 Segmental SNR measurements for speech utterances with higher level of noise

Speech enhancement method	**Average SNR Value**
*β*masking-TwoSensor	1.34
ETSI (left channel)	0.81
β-masking (left channel)	0.55
Two-sensor	-3.40
Noisy (right channel)	-14.12
Noisy (left channel)	-10.32

Table 4.3: Average Segmental SNR measurement (higher level of noise)

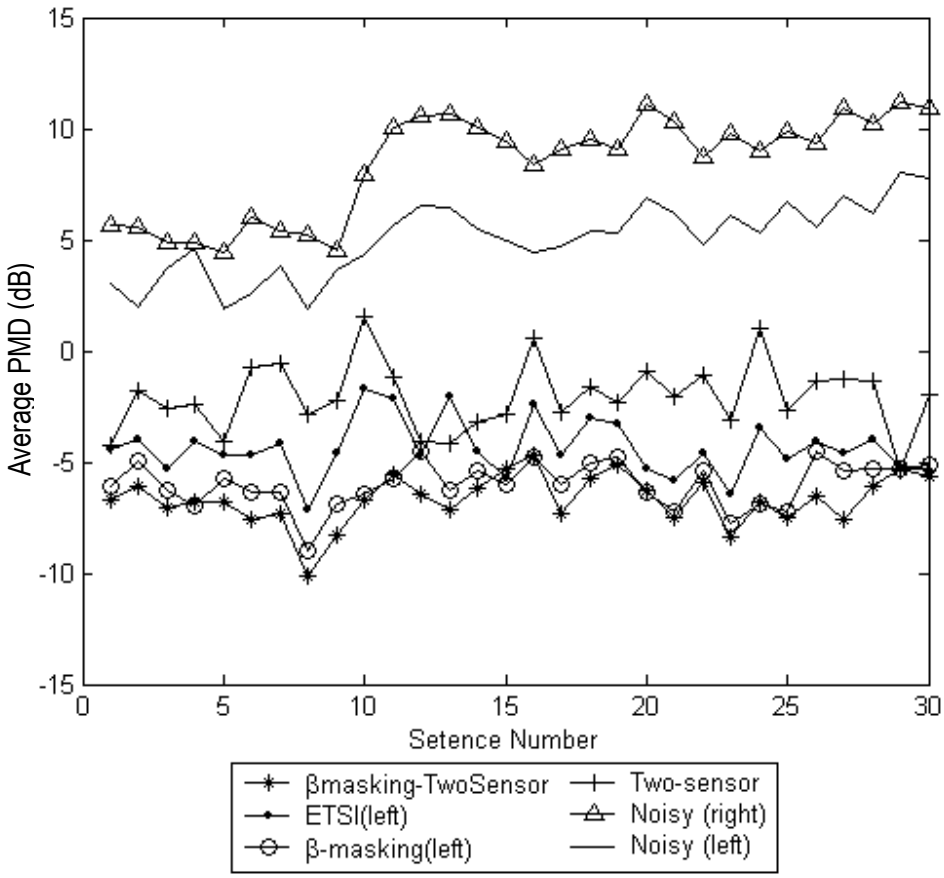

Figure 4.10: PMD measurements for speech utterances with higher level of noise

Speech enhancement method	**Average PMD Value**
*β*masking-TwoSensor	-6.68
ETSI (left channel)	-4.38
β-masking (left channel)	-6.00
Two-sensor	-2.06
Noisy (right channel)	8.41
Noisy (left channel)	5.03

Table 4.4: Average PMD measurement (higher level of noise)

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Conclusion

In this thesis, we have examined a very robust single channel speech enhancement algorithm, namely β-masking. This algorithm derives the MMSE estimator using an adaptive β-order cost function, which incorporates the auditory masking effects. It can effectively attenuate noise to a very low level. However, for cases where a very high level of background noise is observed, the weak spectral components are likely to be totally swamped by the noise, and they are not recoverable by the β-masking enhancement method alone. For some instances, over-attenuation of most spectral components and over-amplification of a few spectral components occur during speech periods, causing perceptible tonal speech distortions. Therefore, two post-processing methods, namely the non-linear high-frequency regeneration and the cepstrum-aided autocorrelation methods, are proposed to compensate for the over-attenuation of some of the voiced speech components.

The NHR algorithm is used to reconstruct the upper frequency band information of a voiced frame by means of the lower-band information. We assume that the frequency content of a voiced speech signal is perfectly periodic and we can fold the lower-band so as to regenerate the upper-band spectral structure. By reconstructing this periodic

information, and restricting the magnitude of the reconstructed spectrum to the general formant shape, we can remove the undesirable tonal distortions that occur in the high frequency band.

The weak spectral components predominantly in the high frequency region can also be regenerated by using the autocorrelation of the strong voiced spectral components in the lower frequency band. By shifting and adding the spectral magnitudes n points away, where n is the number of spectral points between any two peaks, we are able to reconstruct the periodicity of the over-attenuated weak spectral components of a voiced frame. In order to avoid local-maxima interference, cepstrum is used to help to obtain the correct periodicity.

Simulation results of the three algorithms, namely β-masking, βmasking-NHR and βmasking-Corr, show that the latter two methods are able to re-synthesize spectral components that cannot be recovered by the β-masking method alone. This might help to increase speech intelligibility, and the performance of an automatic speech recognizer. Based on various objective performance measures used in this thesis, these two post-processing methods work very well for white-noise, F16-noise and pink-noise corrupted speech signals. However, their performance is not better for car-noise corrupted speech signals when the noise level is low because the β-masking method alone works sufficiently well and further processing by NHR or the Autocorrelation method actually causes more distortions.

This thesis also examines a well-known two-sensor noise reduction method and shows how the single-channel β-masking method could be used to improve its overall noise reduction performance. The two-sensor car noise reduction system makes use of the

spatial coherence of two speech signals. This algorithm is very good at retaining speech components, but it appears not so efficient in noise attenuation. Passing the β-masking enhanced speech signal through a coherence-weighting filter is equivalent to cascading the β-masking gain and the two-sensor filter gain (or multiplication of the two gains in the frequency domain). Multiplying the two gains can improve and constrain each other and we have demonstrated through experiments that it can recover the comb-structure to a certain extent and reduce the over-amplification effect of the β-masking gain. The proposed scheme was tested using real noisy speech signals recorded in a car under different conditions. Objective measurements such as SNR and PMD show that it is effective in improving the quality of the processed speech. Subjectively, the output speech signals processed by the proposed method have very low background noise and very low speech distortion. As compared to the noise reduction technique used in the ETSI standard, the proposed βmasking-TwoSensor enhancement method also has a far superior performance.

5.2 Recommendations for future research

Although the two post-processing methods, namely the non-linear high-frequency regeneration and the cepstrum-aided autocorrelation methods, can improve the performance of the β-masking method alone by re-synthesizing the upper-band weak spectral components, they are just an "add-on" to the β-masking method and therefore the overall enhancement scheme is not computationally efficient. It would be a neater solution if the idea of the periodicity regeneration to enhance the weak spectral components could be incorporated into the β-masking enhancement algorithm directly.

Though the two-sensor scheme that incorporates the β-masking enhancement method leads to the best performance for the enhanced speech signals, the combined complexity is relatively high. Further investigations could be focused on the use of less complicated MMSE methods in place of β-masking to make the overall scheme more acceptable in terms of implementation complexity. The existing two-sensor scheme exploits the fact that the noise inputs to the two microphones spaced at 80cm in a car are uncorrelated for frequencies above 210 Hz. The spacing issue could be re-examined and new enhancement schemes could perhaps be developed to achieve a better performance.

REFERENCES

[1] J. S. Lim and A.V. Oppenheim, "Enhancement and bandwidth compression of noisy speech", *Proc. IEEE*, Vol. 67, No. 12, pp. 1586-1604, Dec. 1979.

[2] M. R. Weiss, E. Aschkenasy and T. W. Parsons, "Study and development of the INTEL technique for improving speech intelligibility", *Report NSC-FR/4023,* Nicolet Scientific Corporation, Dec. 1974.

[3] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. Acoust., Speech, Signal Processing,* Vol. ASSP-27, No. 2, pp. 113-120, Apr. 1979.

[4] J. S. Lim, "Enhancement and bandwidth compression of noisy speech by estimation of speech and its model parameters", *Sc. D. dissertation, Dept. Elec. Eng. And Comput. Sci., Massachusetts Inst. Technol. Cambridge*, Aug 1978.

[5] P. Scalart and J. V. Filho, "Speech Enhancement based on a A Priori Signal to Noise Estimation", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing,* Vol. 2, pp. 629-632, 1996.

[6] I. B. Thomas and R. J. Niederjohn, "Enhancement of speech intelligibility at high noise levels by filtering and clipping", *J. Audio Eng. Soc.,* Vol. 16, pp. 412-415, Oct 1968.

[7] M. Berouti, R. Schwartz and J. Makoul, "Enhancement of speech corrupted by acoustic noise", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing,* Vol. 4, pp. 208-211, Apr. 1979.

[8] R. J. McAulay and M. L. Malpass, "Speech Enhancement using a soft-decision noise suppression filter", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-28, No. 2, pp. 137-145, Apr. 1980.

[9] R. A. Curtis and R. J. Niederjohn, "An investigation of several frequency domain processing methods for enhancing the intelligibility of speech in wideband random noise", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 3, pp. 602-605, Apr. 1978.

[10] M. W. Callahan, "Acoustic signal processing based on the short-time spectrum", *Ph. D dissertation, Dep. Comput. Sci., Univ. Utah, Salt Lake City*, Mar 1976.

[11] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part 1*. New York,: Wiley, 1968.

[12] W. K. Pratt, *Digital Image Processing*. New York: Wiley, 1978.

[13] H. C. Andrews and B. R. Hunt, *Digital Image Restoration*. Englewood Cliffs, N. J.: Prentice-Hall, 1977.

[14] H. Liang, J. Rosca and R. Balan, "Independent component analysis based single channel speech enhancement using Wiener filter", *Siemens Corporate Research, 755 College Road East, Princeton*.
http://www.scr.siemens.com/en/pdf/mt_pdf/ISSPIT2003.pdf

[15] R. H. Frazier, S. Samsam, L. D. Braida and A. V. Oppenheim, "Enhancement of speech by adaptive filtering", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 251-253, Apr. 1976.

[16] T. W. Parsons, "Separation of speech from interfering speech by means of harmonic selection", *J. Acoust. Soc. Amer*., Vol. 60, pp. 911-918, Oct. 1976.

[17] R. D. Preuss, "A frequency domain noise cancelling preprocessor for narrowband speech communications systems", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 4, pp. 212-215, Apr. 1979.

[18] B. Widrow, J. R. Glover, Jr., J. M. McCool, J. Kaunitz, C. S., Williams, R. H. Hearn, J. S. Zeidler, E. Dong, Jr., and R. C. Goodlin, "Adaptive noise canceling; Principles and applications", *Proc. IEEE,* Vol. 63, pp. 1692-1716, Dec. 1975.

[19] J. S. Lim and A.V. Oppenheim, "All-pole modeling of degraded speech", *IEEE Trans. Acoustics, Speech, Signal Processing*, Vol. ASSP-26, pp. 197-210, Jun. 1978.

[20] J. S. Lim, "Enhancement and bandwidth compression of noisy speech by estimation of speech and its model parameters", *Sc.D. dissertation, Dept. Elec. Eng. And Comput. Sci., Massachusetts Inst. Technol. Cambridge*, Aug. 1978.

[21] M. Pagano, "Estimation of models of autpregressive signal plus white noise", *Ann. Statist.,* Vol. 2, pp. 99-108, 1974.

[22] K. Steiglitz, "On the simultaneous estimation of poles and zeros in speech analysis", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 25, pp. 229-234, Jun. 1977.

[23] W. J. Done and C. K. Rushforth, "Estimating the parameters of a noisy all-pole process using pole-zero modeling", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, pp. 228-231, Apr. 1979.

[24] B. R. Musicus and J. S. Lim, "Maximum likelihood parameter estimation of noisy data", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 4, pp. 224-227, Apr. 1979.

[25] A.V. Oppenheim and R. W. Schafer, "Homomorphic analysis of speech", *IEEE Trans. Audio Electroacoust.,* Vol. 16, pp. 221-226, Jun. 1968

[26] J. S. Lim, "Spectral root homomorphic deconvolution system", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 27, pp. 223-233, Jun. 1979

[27] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 32, No. 6, pp. 1109-1121, Dec. 1984.

[28] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log Spectral Amplitude Estimator", *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. 33, No. 2, pp. 443-445, Apr. 1985.

[29] C. H. You, S. N. Koh and S. Rahardja: "Adaptive β-order MMSE Estimation for Speech Enhancement", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 1, pp. 900-903, Apr. 2003.

[30] C. H. You, S. N. Koh and S. Rahardja: **"**β-order MMSE spectral amplitude estimation for speech enhancement**"** *IEEE Trans. Speech and Audio Processing,* Vol. 13, No. 4, pp. 475-486, Jul. 2005.

[31] C. H. You, S. N. Koh and S. Rahardja: "Adaptive β-order MMSE speech enhancement application for mobile communication in a car environment", *Proc.Int. Conf. Information, Communication and Signal Processing, ICICS03 Singapore*, Dec. 2003.

[32] C. H. You, S. N. Koh and S. Rahardja: "An MMSE speech enhancement approach incorporating masking properties", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, 2004. Vol. 1, pp. 725-728 , May 2004.

[33] A. Guérin, R. L. Bouquin-Jeannès and G. Faucon, "A two-sensor noise reduction system: applications for hands-free car kit", *EURASIP Journal on Applied Signal Processing,* Vol. 2003, No. 11, pp. 1125-1134, 2003.

[34] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products.* New York: Academic, 1980.

[35] N. Virag, "Single Channel Speech Enhancement Based on Masking Properties of the human Auditory system*", IEEE Trans. Speech and Audio Processing*, Vol. 7, No. 2, pp. 126-137, Mar. 1999.

[36] J. H. L. Hansen, S. Nandkumar, "Robust Estimation of Speech in Noisy Background Based on Aspects of the Auditory Process", *J. Acoust. Soc. Amer*., Vol. 97, No. 7, pp. 3833-3849, Jun. 1995.

[37] C. Plapous, C. Marro and P. Scalart, "Speech enhancement using harmonic regeneration", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 1, pp. 157-160, Mar. 2005.

[38] I. Y. Soon, S. N. Koh and W. H. Ngo: "Transformation of narrowband speech into wideband speech with aid of zero crossings rate", *IEE electronics letter,* Vol. 38, No.24, pp. 1607-1608, Nov 2002.

[39] S. A. Samad, A. Hussain and L. K. Fah, "Pitch detection of speech signals using the cross-correlation technique", *Proc. TENCON*, Vol. 1, pp. 283-286, Sep. 2000.

[40] D. Nelson, "Correlation based speech formant recovery", *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, Vol. 3, pp. 1643-1646, Apr. 1997.

[41] L.R. Rabiner and R.W. Schafer, "Homomorphic Speech Processing", *Digital Processing of Speech Signals*, Chapter 7.

[42] "Perceptual evaluation of speech quality (PESQ), an objective method for end-toend speech quality assessment of narrow-band telephone networks and speech codecs", *ITU-T Recommdation* P.862, 2001.

[43] P. C. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum", *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 5, Part 2, pp. 857-869, Sep. 2005.

[44] R. L. Bouquin-Jeannès, A. A. Azirani and G. Faucon, "Enhancement of speech degraded by coherent and incoherent noise using a cross-spectral estimator", *IEEE Trans. Speech and Audio Processing,* Vol. 5, No. 5, pp. 484-487, Sep. 1997.

[45] S. A. Samad, A. Hussain and K. F. Low, "Pitch Detection of Speech Signals using the Cross-Correlation Technique", *Proc. TENCON 2000*, Vol. 1, pp. 283-286, Sep. 2000.

[46] ETSI ES 202 050 v1.1.4 (2005-11), "Speech Processing, Transmission and Quality Aspect (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms".

APPENDIX A

LIST OF SPEECH FILES ATTACHED IN THE CD

Some sample speech files are included in the CD in order to subjectively demonstrate the quality of speech signals before and after processing. This appendix gives the list and information of the speech files to guide users on browsing the CD.

A.1 Single-channel noisy speech with different noise levels

There are nine speech files under the folder "A1_Single-channel noisy speech with different noise levels". The noisy speech signals are degraded by either white noise or car noise. These files help readers understand subjectively on noise level changes with 5dB variation. From Table A.1 it is noted that for different noise types, same SNR value (-10 dB) does not correspond to same PMD value (37.14 for car noise and 43.11dB for white noise). The reason is that SNR and PMD values are calculated in different domains. Perceptually, the -10dB noisy speech signal degraded by white noise sounds worse than the -10dB noisy speech signal degraded by car noise. From this point of view, PMD measurement result is closer to the perceptual result. For same noise type, a larger value of SNR indicates better subjective quality and a lower value of PMD indicates better subjective quality.

File Name	SNR (dB)	PMD (dB)	Note
FA.pcm	-	-	clean
fa_010_car.pcm	-10	37.14	noisy (car)
fa_n05_car.pcm	-5	33.92	noisy (car)
fa_p00_car.pcm	0	30.10	noisy (car)
fa_p05_car.pcm	5	25.60	noisy (car)
fa_010_white.pcm	-10	43.11	noisy (white)
fa_n05_white.pcm	-5	40.06	noisy (white)
fa_p00_white.pcm	0	36.37	noisy (white)
fa_p05_white.pcm	5	32.16	noisy (white)
Data format: 16-bit Intel PCM (LSB, MSB) Sampling rate: 8000 Hz Channel: Mono Resolution: 16 bit			

Table A.1: File list under the folder

"A1_Single-channel noisy speech with different noise levels".

A.2 Single-channel speech before and after processing

Under the folder "A2_Single-channel noisy speech before and after processing", there are two subfolders specified by the noise level. Each has one set of female and one set of male speech signals (randomly selected). Each set has a clean speech, a noisy speech signal degraded by additive white noise, a β-masking enhanced speech signal, a β-masking-NHR enhanced speech signal and a β-masking-Corr enhanced speech signal.

In the "moderate noise" subfolder, the clean speech signals FC.pcm (female) and MH.pcm (male) are degraded by white noise to obtain the noisy speech with SNR of –5dB. The residual musical tones found in the β-masking enhanced speech signal is suppressed in the β-masking-NHR enhanced speech signal and the β-masking-Corr enhanced speech signal.

File Name	Note
FC.pcm	Clean speech (female)
fc_n05_white.pcm	Noisy speech degraded by additive white noise
fc_n05_white_bMask.pcm	β-masking enhanced speech
fc_n05_white_bMask_NHR.pcm	β-masking-NHR enhanced speech
fc_n05_white_bMask_Corr.pcm	β-masking-Corr enhanced speech
MH.pcm	Clean speech (male)
mh_n05_white.pcm	Noisy speech degraded by additive white noise
mh_n05_white_bMask.pcm	β-masking enhanced speech
mh_n05_white_bMask_NHR.pcm	β-masking-NHR enhanced speech
mh_n05_white_bMask_Corr.pcm	β-masking-Corr enhanced speech
Data format: 16-bit Intel PCM (LSB, MSB) Sampling rate: 8000 Hz Channel: Mono Resolution: 16 bit	

Table A.2: File list under the folder

"A2_Single-channel noisy speech before and after processing / moderate noise".

In the "high noise" subfolder, the clean speech signals FC.pcm (female) and MH.pcm (male) are degraded by white noise to obtain the noisy speech with SNR of –

10dB. The residual musical tones found in the β-masking enhanced speech signal is suppressed in the β-masking-NHR enhanced speech signal and the β-masking-Corr enhanced speech signal.

File Name	Note
FC.pcm	Clean speech (female)
fc_n10_white.pcm	Noisy speech degraded by additive white noise
fc_n10_white_bMask.pcm	β-masking enhanced speech
fc_n10_white_bMask_NHR.pcm	β-masking-NHR enhanced speech
fc_n10_white_bMask_Corr.pcm	β-masking-Corr enhanced speech
MH.pcm	Clean speech (male)
mh_n10_white.pcm	Noisy speech degraded by additive white noise
mh_n10_white_bMask.pcm	β-masking enhanced speech
mh_n10_white_bMask_NHR.pcm	β-masking-NHR enhanced speech
mh_n10_white_bMask_Corr.pcm	β-masking-Corr enhanced speech
Data format: 16-bit Intel PCM (LSB, MSB) Sampling rate: 8000 Hz Channel: Mono Resolution: 16 bit	

Table A.3: File list under the folder

"A2_Single-channel noisy speech before and after processing / high noise".

A.3 Two-channel speech before and after processing

Under the folder "A3_Two-channel noisy speech before and after processing", there are two subfolders. One is for low level of noise and the other is for high level of noise. Each has two sets of speech signals (randomly selected). Each set has a quasi-clean speech signal, two real-time recorded noisy speech signals (left channel and right channel), an ETSI enhanced speech signal, a β-masking enhanced speech signal, a Two-Sensor enhanced speech signal and a βmasking-TwoSensor enhanced speech signal.

In the "low noise" subfolder, the clean speech signals m04.wav and m09.wav were recorded with windows closed and no cars passing by.

File Name	Note
m04.wav	A Quasi-clean speech recorded using a very closed headset
m04_nsL.wav	Noisy speech (Left-channel) recorded with low noise
m04_nsR.wav	Noisy speech (Right-channel) recorded with low noise
m04_ETSI.wav	ETSI enhanced speech
m04_2sensor.wav	Two-Sensor enhanced speech (Left-channel)
m04_bMask.wav	β-masking enhanced speech (Left-channel)
m04_bMask2sensor.wav	βmasking-TwoSensor enhanced speech (Left-channel)
m09.wav to m09_bMask2sensor.wav	Another set of speech signals using the same naming convention.
Data format: wave file Sampling rate: 16000 Hz Channel: Stereo Resolution: 16 bit	

Table A.4: File list under the folder

"A3_Two-channel noisy speech before and after processing / low noise".

In the "high noise" subfolder, the clean speech signals m05.wav and m09.wav were recorded with a higher driving speed, window (right side) wound down, and cars passing by.

File Name	Note
m05.wav	A Quasi-clean speech recorded using a very closed headset
m05_nsL.wav	Noisy speech (Left-channel) recorded with high noise
m05_nsR.wav	Noisy speech (Right-channel) recorded with high noise
m05_ETSI.wav	ETSI enhanced speech
m05_2sensor.wav	Two-Sensor enhanced speech (Left-channel)
m05_bMask.wav	β-masking enhanced speech (Left-channel)
m05_bMask2sensor.wav	βmasking-TwoSensor enhanced speech (Left-channel)
m09.wav to m09_bMask2sensor.wav	Another set of speech signals using the same naming convention.
Data format: wave file Sampling rate: 16000 Hz Channel: Stereo Resolution: 16 bit	

Table A.5: File list under the folder

"A3_Two-channel noisy speech before and after processing / high noise".