

**STFT BASED ENVELOPE TRACKING HARMONIC
GENERATOR DESIGN WITH APPLICATION
TO ENHANCING BAND LIMITED
AUDIO SIGNALS**

By

JAE YONG LEE

**Bachelor of Science in Electrical Engineering
Republic of Korea Air Force Academy
Chong Ju, Korea
1984**

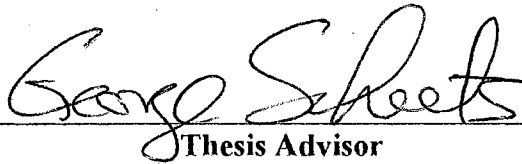
**Master of Science
Oklahoma State University
Stillwater, Oklahoma
1991**

**Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
May, 1998**

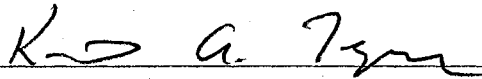
Thesis
1998D
L478s

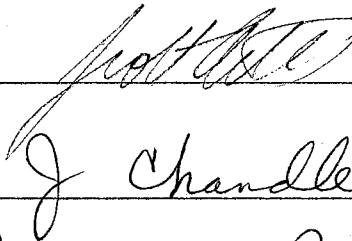
**STFT BASED ENVELOPE TRACKING HARMONIC
GENERATOR DESIGN WITH APPLICATION
TO ENHANCING BAND LIMITED
AUDIO SIGNALS**

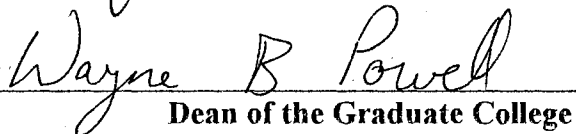
Thesis Approved:



Thesis Advisor







Dean of the Graduate College

ACKNOWLEDGEMENT

This work would not have been possible without the generous support and guidance from my major advisor Dr. George M. Scheets. Great appreciation must be given to his limitless patience, encouragement, and understanding, which were always available to me despite his many responsibilities. These ‘ergodic’ parameters made every step of the way in last four years a pleasure .

Thanks must be extended to other committee members, Dr. Keith Teague, Dr. Scott Acton, and Dr. John Chandler for their support and understanding.

I also want to thank my wife Yong Joo, sons Kee Wan and Kee Hyun, my mother and father, and my wife’s family for their constant love, prayer, and encouragement throughout the years.

Finally, I want to thank for the Republic of Korean Air Force for the financial support.

TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION	1
1.1 Statement of Problem	1
1.2 Overview of Spectrum Analysis Techniques	3
1.2.1 Classical Techniques	3
1.2.2 Parametric Spectrum Estimation	7
1.3 Overview of Chapters	9
1.4 Summary	10
II. ENVELOPE TRACKING HARMONIC GENERATOR	11
2.1 Introduction	11
2.2 Other Higher Harmonic Generating Methods	11
2.2.1 Frequency Doubler	11
2.2.2 High Frequency Regeneration	12
2.2.3 Summary	14
2.3 ETHG Algorithm	15
2.3.1 Introduction	15
2.3.2 Waveform Model	18
2.3.3 ETHG Algorithm :	
Single Frequency, Rectangular Window Case	32
2.3.4 Error Analysis	39
2.3.4.1 Introduction	39
2.3.4.2 Definition of Error	40
2.3.4.3 Analysis of the Effects of the Window Function	41
2.3.4.4 Error Pattern	53
2.3.4.5 Error Analysis : Theoretical Derivation	66

Chapter	Page
2.3.5 An Analysis of the Single Frequency Case	74
2.3.5.1 Result 1: Constant Envelope, Single Sinusoid	77
2.3.5.2 Result 2: Time Varying Envelope, Single Sinusoid	87
2.3.6 The Multiple Frequency Case	100
2.3.7 Summary	108
III. DISTORTION MEASURES.	109
3.1 Introduction.	109
3.2 Time Domain Distortion Measure.	110
3.2.1 Signal to Noise Ratio	110
3.2.1.1 Frequency Weighted Segmental SNR	111
3.2.2 Mean Square Error	112
3.2.3 Normalized Squared Error	113
3.3 Frequency Domain Distortion Measure	114
3.3.1 Distance Measure	115
3.3.2 Other Spectral Distance Measures	116
3.4 Cepstral Distance Measure	117
3.5 Summary.	117
IV. SIMULATION WITH AUDIO SIGNAL	119
4.1 Introduction.	119
4.2 Sound Generating Mechanism and Synthesis Techniques	119
4.2.1 Definitions	120
4.2.2 Speech Model and Synthesis	120
4.2.3 Analysis and Synthesis of Musical Signals	121
4.3 Psychoacoustics	123
4.4 Audio Characteristics.	125
4.4.1 Statistical Analysis of Audio Signals	126
4.5 Subjective Quality Analysis:	
Generation of Harmonics for 5 ~10 KHz Region	130
4.6 Subjective Quality Analysis:	
Generation of Harmonics for 10 ~ 15 KHz region	133
4.7 Summary.	135

Chapter	Page
V. CONCLUSION	136
5.1 Summary	136
5.2 Future Research Considerations	138
REFERENCES	140
APPENDIX I	145

LIST OF TABLES

Table	Page
2.1 Examples of sum of absolute error of middle $L/2$ points between the window and the convolution result	52
2.2 Obtainable K for constant envelope, single sinusoid	77
2.3 Obtainable K for time varying envelope, single sinusoid	88
2.4 Comparison of MNSE (FFT length = 1024)	89
2.5 Window Selection Criterion (WSC)	98
2.6 Comparison of average MNSE of 1024 FFT and 2048 FFT	98
2.7 Obtainable K for time varying envelope, double sinusoid	103
4.1 Mean spectral magnitude ratios	129
A.1 Result of MLE program	156
A.2 Parameters of AR(100) model in Fig. 2.4 in section 2.3.2	159

LIST OF FIGURES

Figure	Page
1.1 Role of the ETHG algorithm	2
2.1 High frequency regeneration technique	13
2.2 Waveform Envelope Generation Model for a typical window	20
2.3 Envelope Generation Process by AR modeling	21
2.4 Comparison of power spectrum	24
2.5 Examples of waveforms (23 msec)	25
2.6 Statistical nature of real audio signal envelope Histogram and Autocorrelation from a popular music.	27
2.7 Statistical nature of real audio signal envelope Histogram and Autocorrelation from a classical music.	28
2.8 Histogram and Autocorrelation function of model envelope.	29
2.9 Histogram and Autocorrelation function of model envelope.	30
2.10 Block diagram of the Envelope Tracking Harmonic Generator	33
2.11 Signals during the higher harmonic generation process	34
2.12 Signals during the higher harmonic generation process	35

Figure	Page
2.13 Spectrum translation	45
2.14 Convolution result of 64 and 256 length rectangular window	47
2.15 Convolution result of 64 and 256 length Hamming window	49
2.16 Convolution result of 64 and 256 length Hanning window	51
2.17 Minimum squared error position (middle)	54
2.18 Minimum squared error position (middle)	55
2.19 Minimum squared error position (right of middle)	56
2.20 Minimum squared error position (left of middle)	57
2.21 (a) MSE and average MSE, Rectangular window	59
2.21 (b) MSE and average MSE, Rectangular window, 512 zeros padded	60
2.22 (a) MSE and average MSE, Hamming window	61
2.22 (b) MSE and average MSE, Hamming window, 512 zeros padded	62
2.23 (a) MSE and average MSE, Hanning window	63
2.23 (b) MSE and average MSE, Hanning window, 512 zeros padded	64
2.24 Comparison of Average MNSE	79
2.25 Average MNSE for $L = 64, 128, 256, 512, 1024, K = 4$	80
2.26 Average MNSE for $L = 64, 128, 256, 512, 1024, K = 8$	81
2.27 Average MNSE for $L = 128, 256, 512, 1024, K = 16$	82
2.28 Average MNSE for $L = 256, 512, 1024, K = 32$	83
2.29 Average MNSE for $L = 512, 1024, K = 64$	84

Figure	Page
2.30 Average MNSE for $L = 1024$, $K = 128$	85
2.31 MNSE for FFT length = 1024, $K = 128$	90
2.32 Average MNSE for $L = 64, 128, 256, 512, 1024$, $K = 4$ Time varying envelope case.	92
2.33 Average MNSE for $L = 128, 256, 512, 1024$, $K = 8$ Time varying envelope case.	93
2.34 Average MNSE for $L = 256, 512, 1024$, $K = 16$ Time varying envelope case.	94
2.35 Average MNSE for $L = 512, 1024$, $K = 32$ Time varying envelope case.	95
2.36 Average MNSE for $L = 1024$, $K = 128$ Time varying envelope case.	96
2.37 Waveform with double frequencies generation for typical window.	101
2.38 Mean squared error, 512 length Hanning window	102
2.39 Average MNSE for $L = 256, 512, 1024$, $K = 32$ Time varying envelope, 2 sinusoid case	104
2.40 Average MNSE for $L = 512, 1024$, $K = 64$ Time varying envelope, 2 sinusoid case	106
2.41 Average MNSE for $L = 1024$, $K = 128$ Time varying envelope, 2 sinusoid case	107
4.1 The human auditory organ	124
4.2 Mean magnitude spectrum	128
4.3 Example of harmonic generation to 5 ~ 15 KHz region	134

Figure	Page
A.1 Box-Jenkins method of estimation sequence.	146
A.2 GPAC	147
A.3 An envelope to be modeled	152
A.4 GPAC of the envelope signal in Fig. A.3	153
A.5 Comparison of AR model's normalized frequency response	154
A.6 Plot of ACF and PACF of residual	157

CHAPTER I

INTRODUCTION

1.1 Statement of Problem

This dissertation addresses the problem of developing a new envelope tracking harmonic generator (ETHG) algorithm using the Short Time Fourier Transform (STFT) technique. This envelope tracking harmonic generator (hereafter ETHG) uses the existing bandlimited harmonic structure of a signal to generate estimates of higher harmonics by a modified spectral translation technique, while minimizing the mean normalized squared estimation error. The estimation error is minimized by noting that the harmonic generation process has the least amount of error in the middle of the Short Time Fourier Transform processed window. The ETHG has the potential to improve the sound quality of current analog band limited systems such as AM radio, the telephone, or audio-over-the-Internet. The human ear is capable of hearing up to 20 KHz, and many musical instruments including human voice are rich in harmonics, however many systems in use today are designed with base band bandwidths less than 5 KHz. As a result, they do not yield the sound quality that we desire.

The goal of this dissertation is to theoretically analyze the ETHG algorithm's behavior, and find its optimal performance in terms of minimizing the mean normalized

squared error between a processed signal and its full fidelity original, by investigating the performance of the algorithm when various parameters, such as window type, length, and zero padding, are adjusted.

Fig. 1.1 shows a simplified block diagram of the ETHG algorithm. The high fidelity original $X(n)$ with full harmonic structure is an input signal to the baseband equivalent narrow band system, and the distorted output $Y(n)$ is the band limited audio with a missing higher harmonic structure. In this research, we assume that the degree of band limiting is so severe that no information exists as to the missing higher frequency energy, meaning that *traditional linear filtering techniques will not work*.

The ETHG algorithm nonlinearly processes this distorted output, and generates the processed output $\hat{X}(n)$ which has an estimate of the missing higher harmonic structure. This algorithm has the potential capability of improving the sound quality of severely band limited systems without the need of changing the transmitter, channel, or receiver configurations. In other words, instead of modifying the large installed base of transmitters and receivers, instead of increasing the bandwidth available in the channel (which for AM broadcasting would require FCC approval as well as bandwidth which just isn't there), the ETHG might be usable to process the band limited signal at the receiver output, giving the user the subjective impression of extended high frequency response. To do this well, the behavior of the ETHG needs to be well understood.

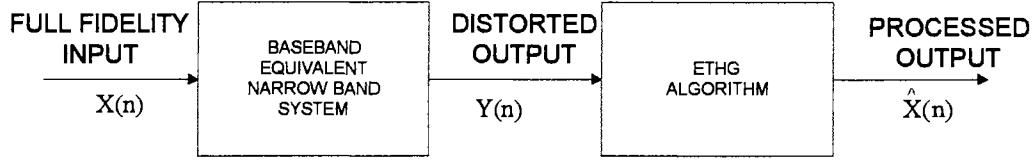


Fig.1.1 Role of the ETHG algorithm

1.2 Overview of Spectrum Analysis Technique

A summary on the spectrum estimation techniques and the comparison of their performance is discussed in this section. Understanding the current available techniques could provide diverse options which may lead to achieving better performance in ETHG.

1.2.1 Classical Techniques

The periodogram and Blackman-Tukey spectral estimators are classified as the classical spectral estimation techniques, which are based on the Fourier analysis. The periodogram is the most popular spectral estimator that is in use today. The periodogram estimator is defined as [Haye 96, Kay 88],

$$\hat{S}_{xx_{per}}(e^{j\omega}) = \sum_{k=-N+1}^{N-1} \hat{R}_{xx}(k) e^{-j\omega k}, \quad (1.1)$$

where the hat (Λ) indicates an estimated approximation. It is based on the Wiener-Khinchin theorem which is defined as

$$S_{xx}(e^{j\omega}) = \sum_{k=-\infty}^{\infty} R_{xx}(k) e^{-j\omega k}, \quad (1.2)$$

where $S_{xx}(e^{j\omega})$ is the power spectrum, and $R_{xx}(k)$ is an autocorrelation function of input $x(n)$ described as,

$$R_{xx}(k) = E[X(n)X(n+k)], \quad (1.3)$$

and E is the expected value operator. For ergodic finite data the autocorrelation estimate is

$$\hat{R}_{xx}(k) = \frac{1}{N} \sum_{n=0}^{N-1} x(n+k)x^*(n), \quad (1.4)$$

where $*$ denotes complex conjugate. We can rewrite Eq (1.4) using window function as

$$\hat{R}_{xx}(k) = \frac{1}{N} \sum_{n=0}^{N-1} wx(n+k)wx^*(n), \quad (1.5)$$

where $wx(n)$ is a windowed $x(n)$. Taking the Fourier Transform of Eq (1.5) yields

$$\hat{S}_{xx_{per}}(e^{j\omega}) = \frac{1}{N} WX(e^{j\omega})WX^*(e^{j\omega}) \quad (1.6)$$

$$= \frac{1}{N} |WX(e^{j\omega})|^2, \quad (1.7)$$

where

$$WX(e^{j\omega}) = \sum_{n=0}^{N-1} wx(n)e^{-j\omega n}. \quad (1.8)$$

Therefore, the periodogram can be directly obtained from the Fourier Transform of the input data as

$$\hat{S}_{xx_{per}}(e^{j\omega}) = \frac{1}{N} \left| \sum_{n=0}^{N-1} wx(n) \exp(-j\omega n) \right|^2. \quad (1.9)$$

Using the DFT (Discrete Fourier Transform) notation, Eq (1.9) can be written as,

$$\hat{S}_{xx_{per}}^{\Lambda}(e^{j\frac{2\pi k}{N}}) = \frac{1}{N} \left| \sum_{n=0}^{N-1} wx(n) \exp(-j2\pi kn / N) \right|^2 \quad (1.10)$$

where $k = 0, 1, \dots, N-1$ [Kay 88, Haye 96].

The periodogram estimate of Eq (1.10) is asymptotically unbiased, meaning the estimate approaches the true power spectral density as $N \rightarrow \infty$, and it can be computed via the Fast Fourier Transform (FFT). The FFT, credited to Cooley and Tukey, is a fast and efficient method to compute the discrete Fourier Transform (DFT).

The commonly used FFT algorithms are decimation-in-time (DIT) and decimation-in-frequency (DIF) algorithms for the sequences with length of integer power of 2. Although the FFT algorithm enjoys the efficient computation order of $N \log_2 N$ complex additions and multiplications when N is a power of 2 [Oppe 89], it has a frequency resolution problem, especially with short sequences which commonly occur in radar, sonar and seismic signal processing, and a leakage problem due to windowing of the data sequence [Kay 81, Kay 88, Soli 90, Oppe 89, Haye 96].

The Blackman-Tukey method is one of the modified periodogram techniques designed to reduce the variance of the periodogram's spectral estimation. The estimation variance is reduced by discarding the autocorrelation estimate with largest variance. Then take the Fourier Transform with the remaining autocorrelation estimate. The Blackman-Tukey periodogram is defined as

$$\hat{S}_{xx_{BT}}^{\Lambda}(e^{j\omega}) = \sum_{k=-M}^M \hat{R}_{xx}(k) w(k) e^{-jk\omega}, \quad (1.11)$$

where $w(k)$ is a lag window, and M is the window length [Kay 88, Haye 96].

The Blackman-Tukey method was the most popular method until the introduction

of the FFT algorithm in 1965. It introduces less variance into the spectral estimate, however for signals with narrowband power spectral densities, it introduces a significant amount of bias, because using reduced number of the autocorrelation estimate results in a broader spectrum. The periodogram results in less bias [Kay 88, Haye 96, Oppe 89].

For a time varying signal which is stationary only for a short time, we should use the Short Time Fourier Transform (STFT). The ETHG algorithm uses this STFT because it provides highly efficient measurements of time varying harmonic structures of the audio signals [RMoo 90, Port 80, Port 81, Kay 88]. The STFT is

$$X_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m}, \quad (1.12)$$

where $x(m)$ is the input sequence, and $w(n-m)$ is the window which is positioned at time n along the input sequence [Kay 88]. The window function will be discussed in Chapter 2.

Several techniques to improve the accuracy of the STFT's frequency estimate have been proposed [Brow 93, Brow 96, Ando 93, Tabe 88, Park 86]. Among them, the Quadratic Fit Method and the frequency estimation by phase vocoder using the Single Frame Approximation (SFA) technique produce a frequency estimate with a higher accuracy. However, the computation load increases significantly with these techniques, and the increased accuracy is confined to the frequency estimate only, whereas the ETHG needs both frequency and phase estimates.

In order to overcome problems related with the FFT, many algorithms have been introduced, and the advantages and disadvantages of some selected spectral estimators will be discussed in following sections.

1.2.2 Parametric Spectrum Estimation

Model based spectrum estimation techniques have been developed to achieve better spectral estimation which is not possible, under certain conditions, with the conventional spectrum analysis methods such as the periodogram or Blackman-Tukey method. If one has an idea of how the signal was formed by an arbitrary transfer function, then one can generate better spectral estimates by using a model which adequately describes the transfer function. The model selecting and evaluating procedure is described in Box-Jenkins [BoxJ 94] and Pankratz [Pank 83]. The general model is an ARMA (autoregressive moving average) model, which is a pole-zero model, described as,

$$y(n) = \sum_{l=0}^q b_l x_{n-l} - \sum_{k=1}^p a_k y_{n-k} . \quad (1.13)$$

The transfer function via z- transform is,

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^q b_k z^{-k}}{1 + \sum_{k=1}^p a_k z^{-k}} . \quad (1.14)$$

If $b_0 = 1$ and all the remaining b_i are zero then it becomes an AR (autoregressive) model, which is an all pole model, as,

$$y(n) = - \sum_{k=1}^p a_k y_{n-k} + x_n . \quad (1.15)$$

The MA (moving average) model, an all zero model, is,

$$y(n) = \sum_{l=0}^q b_l x_{n-l} . \quad (1.16)$$

There are many algorithms to calculate the model parameters such as

autocorrelation method, covariance method, adaptive linear prediction methods using various algorithms like LMS (least mean square), RLS (recursive least squares), and the MLE (maximum likelihood estimator). The estimator's properties such as unbiasedness, efficiency and consistency are described in Mendel [Mend 95], and a performance comparison is found in Hayes [Haye96], Kay [Kay 88], Haykin [Hayk 96], and Widrow [Widr 85].

Even though these techniques yield better estimates than classical methods under certain circumstances, it is not always easy to find the parsimony model for the given realization, especially for a time varying transfer function structure. If the model does not adequately describe the transfer function, it will yield a poor estimation. Another major drawback is that the computational complexity of these algorithms is much more intensive than classical methods [Kay 81, Kay 88, Haye 96].

In addition to the parametric estimation methods, we have other spectral estimation techniques based on eigendecomposition of the autocorrelation matrix such as the Pisarenko Harmonic decomposition method and MUSIC (multiple signal classification method). These eigen-analysis based techniques generally require a very intensive computation load, and may yield a biased estimate because of the estimated autocorrelation matrix [Haye 96, Kay 88].

The ETHG should be able to follow the time varying structure of diverse audio signals with computational efficiency, so the short time FFT based method seems most appropriate for this purpose when it is carefully implemented, with a certain window type. In later sections, an analysis on the effects of window type, size and zeropadding on the estimation error, and the error effects on perceptive quality of generated audio signal with

higher harmonics will be discussed. Following is an overview of this paper.

1.3 Overview of Chapters

A literature review of other harmonic generation technique is presented in section 2.2, followed by an introduction of the ETHG in section 2.3. In section 2.3.2, a model waveform generation scheme is introduced. Theoretical discussions about the optimal window type along with other window types and their role in the convolution with the frequency truncation impulse response is in section 2.3.4.3. Also, the error pattern of the ETHG will be shown, by computer simulations in section 2.3.4.4 and by mathematical derivation in section 2.3.4.5, to be minimum in the middle of the STFT processed window. Extensive computer simulation results using the window types described above and zero padding to enhance the frequency resolution are given in section 2.3.5 and section 2.3.6.

A review of time domain and frequency domain distortion measures, which includes discussions of why the normalized squared error was chosen for error analysis of the ETHG, is presented in Chapter 3.

A statistical analysis of real audio signals and simulation results using real audio signals are included in Chapter 4. The general relationship between harmonic structures, and the error effects on perceptive quality will be discussed.

In Chapter 5, a summary of the key works done so far to achieve the goal of this paper, and the future research considerations are presented.

1.4 Summary

We have discussed the goal of this dissertation in section 1.1, which is to optimize performance by analyzing the ETHG algorithm's behavior. The motivation to improve the audio quality of a bandlimited audio source using the ETHG, while not modifying the configuration of current bandlimited systems such as AM broadcasting, the telephone system, and audio-over-the-Internet was mentioned as a possible use for this algorithm. A comparison of spectrum analysis techniques was made in section 1.2. The Short Time Fourier Transform was selected for the ETHG algorithm to efficiently follow the time varying nature of audio signals. The computation load of Short Time Fourier Transform appears to be much less than other complicated spectrum analysis techniques.

CHAPTER II

ENVELOPE TRACKING HARMONIC GENERATOR

2.1 Introduction

In this chapter, a discussion of other higher harmonic generating methods and problems is presented. Then a theoretical explanation of the ETHG algorithm and the computer simulation results which support the theory will be discussed.

2.2 Other Higher Harmonic Generating Methods

There are a limited amount of related works on higher harmonic generation from a band limited signal. In the following sections, some of the works worthwhile to note are presented.

2.2.1 Frequency Doubler

A simple squaring device called a frequency doubler can generate the higher harmonic terms from a given sinusoidal input [Stre 90]. Define an input sinusoid as,

$$x(t) = A \cos(w_0 t) . \quad (2.1)$$

If we suppose the input-output function is,

$$y(t) = a_1 x(t) + a_2 x^2(t) , \quad (2.2)$$

then the output can be written as,

$$y(t) = a_1 A \cos(w_0 t) + a_2 A^2 \cos^2(w_0 t) \quad (2.3)$$

$$= \frac{1}{2} a_2 A^2 + a_1 A \cos(w_0 t) + \frac{1}{2} a_2 A^2 \cos(2w_0 t). \quad (2.4)$$

The last term in Eq (2.4) is the generated second harmonic. Likewise, the third harmonics and so on can also be generated. This is a simple illustration of the concept, and cannot be applied directly to the band limited audio signal, because of following reason. Suppose $x(t)$ is composed of two sinusoids as

$$x(t) = A \cos(w_0 t) + B \cos(w_1 t). \quad (2.5)$$

Squaring of $x(t)$ will generate two cross terms. Since an audio signal can be considered to be made up of possibly thousands of sinusoids, this squaring device will have a problem with many unwanted cross terms.

2.2.2 High Frequency Regeneration

This technique was proposed by Makhoul for the purpose of regenerating the high frequency portion of the excitation in a baseband coder for speech signal processing [Makh 79]. The high frequency regeneration technique is simply duplicating the baseband spectrum at higher frequencies in two possible ways: spectral folding, and spectral translation. Fig. 2.1 (a) is the baseband spectrum. Fig. 2.1 (b) shows the result of spectral folding, and Fig. 2.1 (c) the result of spectral translation. It is assumed that the signal bandwidth W is an integer multiple of the baseband width B , for simplicity. The spectral folding in Fig. 2.1 (b) shows that the spectrum in the second band (between B and $2B$) is the mirror image of the baseband, and the spectrum in the third band is the mirror image of

the second band. The spectral translation, Fig. 2.1 (c), shows the spectrum of second and third band as identical to the baseband. Details of spectral folding and translation are described in [Makh 79].

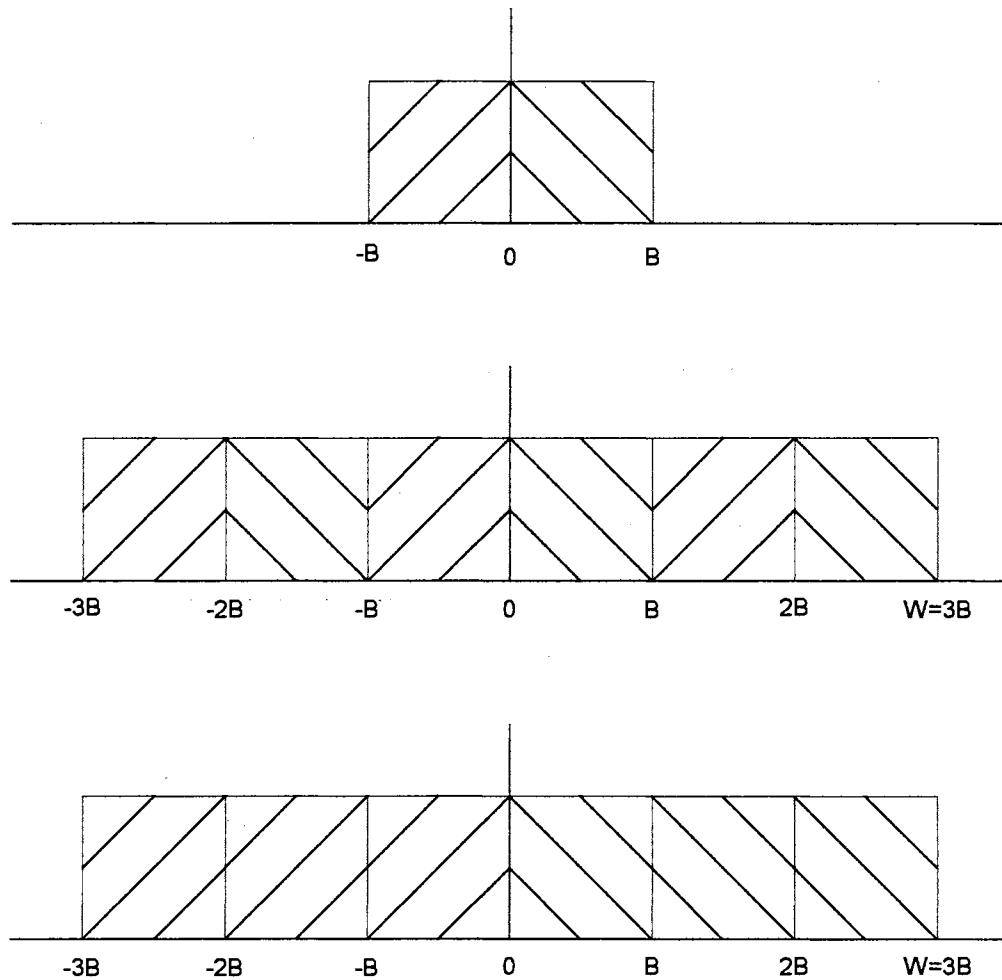


Fig. 2.1 High frequency regeneration technique.

(a) Baseband spectrum

(b) 3-band spectral folding

(c) 3-band spectral translation [Makh 79]

This spectral duplication causes a breaking of the harmonic structure, thus creating distortions in the output speech. Makhoul stated that there were a number of distortions in the form of added tones, and these tones were generally more audible with a large number of bands and for higher-pitched voices. Some of the distortion tones could be reduced, but there were other audible tones which were difficult to trace.

The basic idea of frequency duplication is used in the ETHG algorithm in that lower frequency energy is shifted to higher frequency regions, however, there is a significant difference. In the ETHG algorithm, the shifted energy is split up into pieces to preserve the harmonic structure, while the spectral duplication technique shifts whole sections of energy into higher frequency regions which results in a breaking of the harmonic structure.

2.2.3 Summary

It appears that there is presently no algorithm specifically designed for regenerating missing harmonically related high frequency energy, given knowledge of energy at lower frequencies, other than the ETHG algorithm. The closest reference would be the high frequency regeneration technique noted above, and no additional works regarding higher frequency regeneration have been noticed in the literature since Makhoul's work in 1979. It would appear that the technique had been abandoned as unworkable.

2.3 ETHG Algorithm

2.3.1 Introduction

The purpose of the ETHG algorithm is to take a band limited audio signal, and generate desired harmonics without the cross frequencies and distortion of the alternate techniques noted in section 2.2. The long term goal is to possibly use this technique to enhance the audio quality of band limited systems such as AM radio, telephone systems, and Internet audio without changing the current system configurations except at the receiver location. Many of these systems will be existing for years to come because of the economic unfeasibility of changing the large installed base. Therefore, an inexpensive device that could artificially insert higher frequency harmonics at the receiver, and make audio sounds more pleasing to our ear seems to be in order.

The original idea of ETHG algorithm was developed by Scheets [Sche 93, Sche 94]. In [Sche 94], the general idea of ETHG algorithm, and preliminary simulation results with problems are described. The mathematical derivation of an error function using a constant amplitude sinusoid input with a rectangular window is found in [Sche 93]. These references form the foundation of this paper, however, additional work is needed to better understand and better characterize the behavior of the ETHG algorithm before it can be successfully applied to real audio.

Following is a brief explanation of areas where the ETHG might be implemented.

AM Radio

AM radio is very popular because of its long range and inexpensive receivers. There are huge numbers of existing AM radio stations throughout the world. Its biggest

problem is that the baseband frequency response is generally limited to 5KHz [Enne 74, Stre 90]. Within recent years, stereo AM was introduced but flopped because of an inability to arrive at a single standard and because of the relatively poor audio quality compared to FM. DSP chips running the ETHG could be installed in AM receivers to improve the subjective quality, making the music and voice sound more pleasing. It would not be necessary to change any of the broadcasting or transmission equipment of existing installations.

Telephone

The analog telephone's bandwidth is limited to roughly 3.5 KHz [Rey 84]. This is a result of filtering and the sampling rate of 8000 samples per second. There are many existing advances currently ongoing in terms of fiber to the home, interactive cable TV, etc. The fact is, however, that this band limited analog system will be with us for many years. A possible use for this algorithm would be as an addition to the telephone receiver. In turn, the audio quality from the telephone line would sound crisper and more realistic.

Internet Phone, Radio

The use of personal computers on the Internet is rapidly increasing. Two of the promising features of the Internet are the Internet phone and Internet radio, however, the current quality of these services is rather poor, largely because of bandwidth limitations. This is due to the fact that the bandwidth comes at a premium. The ETHG might be included in computer systems, and would provide improved Internet phone and radio quality using the currently available telephone lines.

Restoration of old recordings

Some of the best musicians made their recordings when the equipment was inferior

to what is available today. Both noise and frequency response is a problem on these recordings. Vaseghi reported a technique for restoration of old recordings, but his trial was limited to remove the noise disturbances such as scratches, impulse noise, and white noise [Vase 92]. The ETHG algorithm has the potential to enhance the audio quality of old recordings.

Compression

There are numerous compression techniques for digital links. This technique may make it possible to compress analog links or additionally compress digital links. By filtering the source audio to a smaller bandwidth before transmitting through the channel, less channel bandwidth would be necessary. The received audio signal may then be enhanced by the device attached at the receiver side.

The layout of remainder of section 2.3 is as follows. In section 2.3.2, the input waveform model to be used in the ETHG will be defined, and the generating scheme of this input waveform will be discussed. In section 2.3.3, a description of the ETHG and sources of errors are presented. In section 2.3.4.2, the definition of error for the ETHG is defined. An analysis of effects window function will be presented in section 2.3.4.3. The mean error between harmonic generator output and the desired signal will be shown to be minimum in the middle of the STFT processed window, by computer simulations in section 2.3.4.4 and by mathematical derivation in section 2.3.4.5, for the time varying amplitude, single input frequency case. In section 2.3.5, the results of computer simulations for single input frequencies will be discussed, and important findings regarding window type and size will be addressed. Computer simulation results from multiple input frequencies will be discussed in section 2.3.6. Summary of key findings is in section 2.3.7.

2.3.2 Waveform Model

In this section, a possible input model for use in the ETHG algorithm will be defined. Selection of an adequate input model for a real audio signal is an important step in order to get satisfactory computer simulation results. The input model should be able to represent real audio properly, and render mathematical tractability as well. For example, in the audio synthesis arena, simple sinusoids are used to both synthesize audio and analyze the performance mathematically [Moor 77, RMoo 90, Hans 83]. In the additive synthesis audio technique, which is one of the popular methods to synthesize an audio signal and will be described in Chapter 4, it is assumed that an audio signal is comprised of multiple sinusoidal components with different amplitude and harmonic frequencies [Moor 77]. A cosine function, described below in Eq (2.6), is a good model to illustrate the harmonic structure of audio signals, and allows mathematical analysis of the ETHG algorithm [Sche 93, Sche 94, Hans 84, Serr 90]. In [Sche 93, Sche 94], the input signal is assumed as,

$$x(i) = v(i) \cos(2\pi \frac{fin}{fs} i + \theta), \quad (2.6)$$

and the desired signal with N th harmonic is

$$d(i) = v(i) \cos(N2\pi \frac{fin}{fs} i + N\theta), \quad (2.7)$$

where $v(i)$ is an arbitrary time varying envelope, fin is the input frequency in Hertz which is assumed to be stationary (or nearly so) during the L point window interval, fs is the sampling frequency in Hertz, and θ is an arbitrary phase angle, which is also assumed to be stationary (or nearly so) during the L point window interval. Eq (2.7) is assumed to

missing due to severe band limiting. The ETHG algorithm processes Eq (2.6) to generate an estimate of Eq (2.7) which may be written as

$$\hat{d}(i) = \hat{v}(i) \cos(N2\pi \frac{\hat{f}n}{fs} i + N\hat{\theta}), \quad (2.8)$$

where $\hat{v}(i)$, $\hat{f}n$, and $\hat{\theta}$ are all estimates of the input parameters [Sche 93, Sche 94].

For the cosine function above, a Rayleigh envelope, and a pulse envelope were used for performance analysis in [Sche 93].

We formulate the input model by modifying Eq (2.6) to accommodate multiple sinusoids, as

$$x(i) = \sum_{h=1}^M v_h(i) \cos(2\pi \frac{f n_h}{fs} i + \theta_h), \quad (2.9)$$

where M is the number of sinusoids, $v_h(i)$ is the envelope amplitude of the h th sinusoid at time i , $f n_h$ is the input frequency of h th sinusoid in hertz which is assumed to be stationary (or nearly so) during the L point window interval, fs is the sampling frequency in hertz, and θ_h is the phase angle of h th sinusoid which is also assumed to be stationary (or nearly so) during the window interval with length L . These parameters are defined similar to [Sche 93].

The desired signal $d(i)$ is defined similarly as,

$$d(i) = \sum_{h=1}^M v_h(i) \cos(N2\pi \frac{f n_h}{fs} i + N\theta_h), \quad (2.10)$$

where N is the desired harmonic number, i.e., $N = 2$ means the second harmonic, and so on. Eq (2.9) and Eq (2.10) describe the harmonic structure of audio signals that the ETHG is dealing with. The desired signal $d(i)$ is a model of the missing higher harmonic structure

with an arbitrary number of sinusoids in it.

The envelope functions described in [Sche 93] may be seen in parts of real audio signals, however they do not fully represent the various randomly varying envelopes of audio signals. Therefore, it is needed to develop a real generic audio-like envelope function, which may be altered for more specific application such as speech or music if needed, for computer simulation of the ETHG's performance. This envelope generation scheme should be a good approximation of the envelopes seen on real audio signals. Fig. 2.2 shows the generation scheme of the input waveform, which uses autoregressive (AR) modeling.

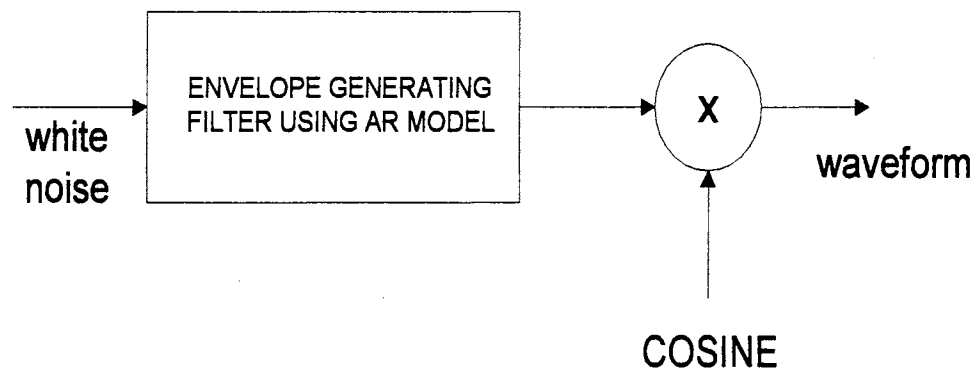


Fig. 2.2 Waveform Envelope Generation Model for a typical window.

The envelope function is generated by filtering a white noise sequence with the envelope generating filter using AR (p), where p is the order of AR model parameters. The role of the envelope generating filter is to generate an envelope shape from white noise which statistically resembles the shape of audio signal envelopes. The AR model has been widely used in many signal processing areas, including speech and musical signal processing, because it provides a sufficiently accurate representation of given realization [Haye 96, RMoo 90, Kay 88]. For example, AR models with 10 to 20 orders are used in speech signal processing, and orders of 50 to 100 are used for musical sound representation [RMoo 90].

The AR (100) model was chosen as accurate model, as discussed in Appendix I. Parameters of the AR model were obtained by a Box-Jenkins modeling procedure [BoxJ 94], which is also discussed in Appendix I. Fig. 2.3 shows the block diagram of the envelope generation process. An overview of the modeling process, and an example using an envelope generated by the low pass filter in Fig. 2.3, is presented in Appendix I.

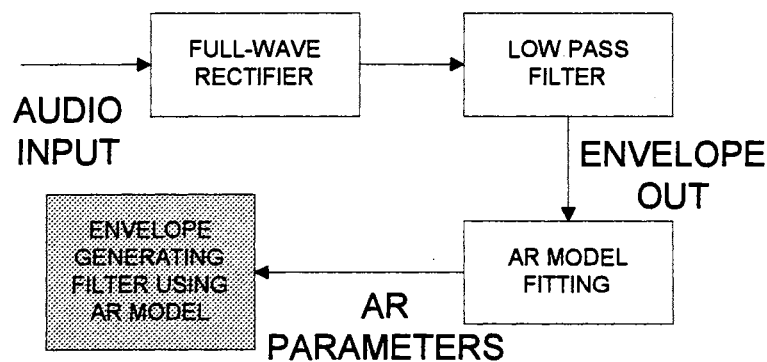


Fig. 2.3 Envelope Generation Process by AR modeling.

Note the purpose of this process is to generate an envelope which is slowly varying within a window interval, and independent of any sinusoidal frequencies modulated by the envelope. In a real audio signal, the envelope amplitude is relatively slowly varying in a short window interval, less than 25 msec, and is independent of frequency [Smit 87, Moor 90, Hall 91]. For example, an instrument can be played loudly or softly without changing the frequency [Hall 91]. This phenomenon may be explained by using the concept of simple harmonic motion (SHM), which is the same as a sinusoidal wave or pure tone. SHM occurs whenever the restoring force of vibration is linear. For example, the restoring force of a string or spring is directly proportional to the degree of displacement from the equilibrium status. Thus restoring forces of musical instruments such as violin and air-filled pipes can be described as one of SHM. The oscillation frequency is

$$f = \frac{\sqrt{K/M}}{2\pi}, \quad (2.11)$$

where K is the stiffness and M is the mass of the medium [Hall 91]. We note that the oscillation frequency is not dependent on the displacement amplitude in Eq (2.11). Therefore, we may assume that the envelope function is independent of the sinusoidal frequency.

Following is a brief discussion of the envelope generation process depicted in Fig. 2.3. The purpose of this process is to generate an average power spectrum from various audio signals to obtain the estimated autocorrelation function (ACF). Non-overlapping one thousand point sections from a 3 second long audio signal, with 44.1 KHz sampling frequency, were an input to the full-wave rectifier, and followed by a fifth order

Butterworth low pass filter with 3 dB cut off frequency of 177 Hertz, which was obtained from experimentation. Thus an average power spectrum of 132 power spectrums is obtained from this 3 second audio signal. Repeat the process more with 23 three second audio signals, and we obtain the average power spectrum of 24 three second audio envelopes. The autocorrelation function is then obtained from the inverse Fourier Transform of the average spectrum. After this, the AR (p), where p is an order of AR model, model's coefficients are obtained by solving the Yule-Walker equation which is described in the Appendix I. We observed that AR (100) model, in general, accurately represents the average power spectrum. Fig. 2.4 (a) shows the normalized mean power spectrum, and Fig. 2.4 (b) shows the normalized power spectrum of AR (100) coefficients. Both figures show frequency range of 0 to .001 Nyquist frequency. The estimated AR (100) parameters are shown in Appendix I.

A comparison of statistical characteristics of envelopes from selected audio signals, and generated envelopes from AR filtering of white noise will be presented in the following discussion.

Note here that the purpose of using the AR (100) model is to generate a waveform which has an arbitrary frequency with a real-audio-like envelope shape. An example of a generated waveform is shown in Fig. 2.5 (b). Fig. 2.5 (a) shows a section of classical music which is a relatively rapidly changing passage with many instruments. We see that the generated waveform shows considerable similarities. Note Fig. 2.5 (b) has one sinusoid, and Fig. 2.5 (a) has many sinusoids. Also note in Fig. 2.5 (a) that the input frequencies appear to be relatively stationary, and the envelope is relatively slowly varying during the window interval, which is generally true for real audio signals, thus generally

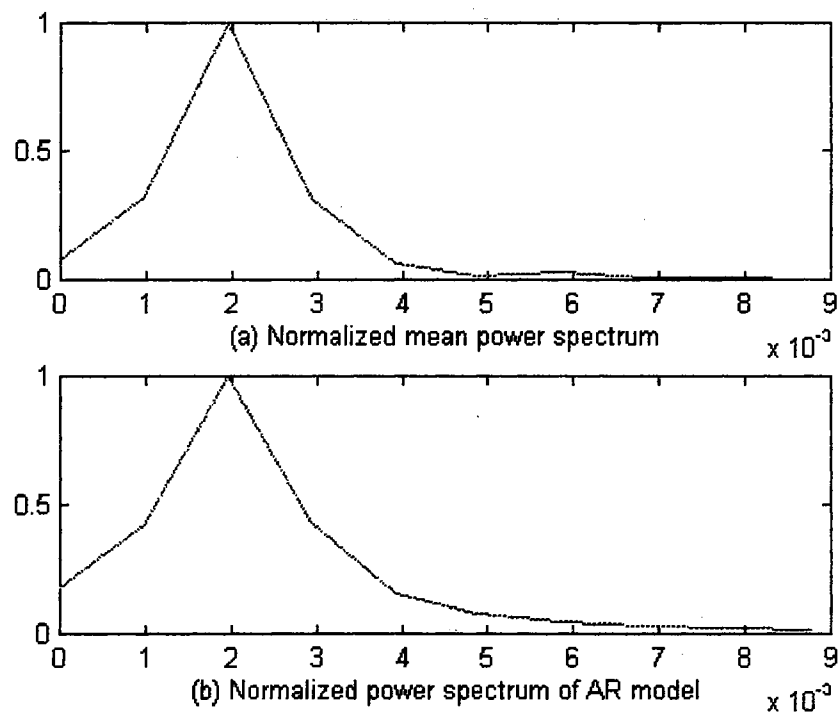


Fig. 2.4 Comparison of power spectrum.

- (a) Normalized mean power spectrum
(0 ~ .001 Nyquist frequency).
- (b) Normalized power spectrum of AR model
(0 ~ .001 Nyquist frequency).

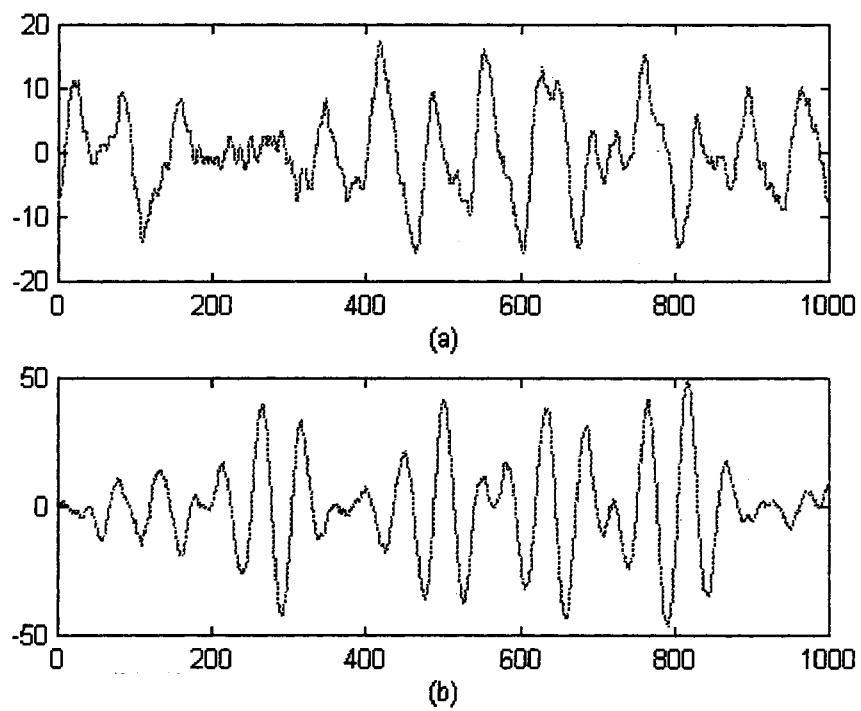


Fig. 2.5 Example of waveforms (23 msec).
(a) Real audio (Classical music).
(b) Generated waveform (single sinusoid).

assumed in many references [Smit 87, Moor 90, Ando 93].

The histogram, and the autocorrelation function of the model envelope agree with those of real audio signals. Fig. 2.6 shows a histogram and autocorrelation function of a 44100 data points envelope which is obtained from popular music (1 second long), which is composed of dominant voice and background music. The autocorrelation function shows up to .1 second. Likewise, the envelope histogram and autocorrelation plot of Fig. 2.7 is obtained from classical music. These plots have similar histogram shapes, and their autocorrelation functions show periodicities up to higher lags. These characteristics are typical of other real audio samples, and are also found in the histograms and autocorrelation functions of the model envelopes. Fig. 2.8 and Fig. 2.9 show examples of histograms, and autocorrelation functions of model generated envelopes. They appear to have similar histogram pattern, and periodicity in the autocorrelation plots, which implies the first and second order statistics of the model envelopes are similar to those of the real audio signals. Therefore the model waveforms shown here, which consist of a single sinusoid modulated by the generated envelope, are seen to have similar characteristics of real audio.

A word about the envelope amplitudes before we proceed further. The envelope amplitudes of Eq (2.9) and Eq (2.10) are the same, i.e. $v_h(n)$ is same for both Eq (2.9) and Eq (2.10). This assumption is not valid for real audio signals. In a real audio signal, the amplitude of each harmonic frequency is different. Every instrument, or every human has their own harmonic structure with different amplitudes at each harmonic frequency. It is the harmonic structure that determines a timbre of its own [Earg 95, Hall 91, RMoo 90].

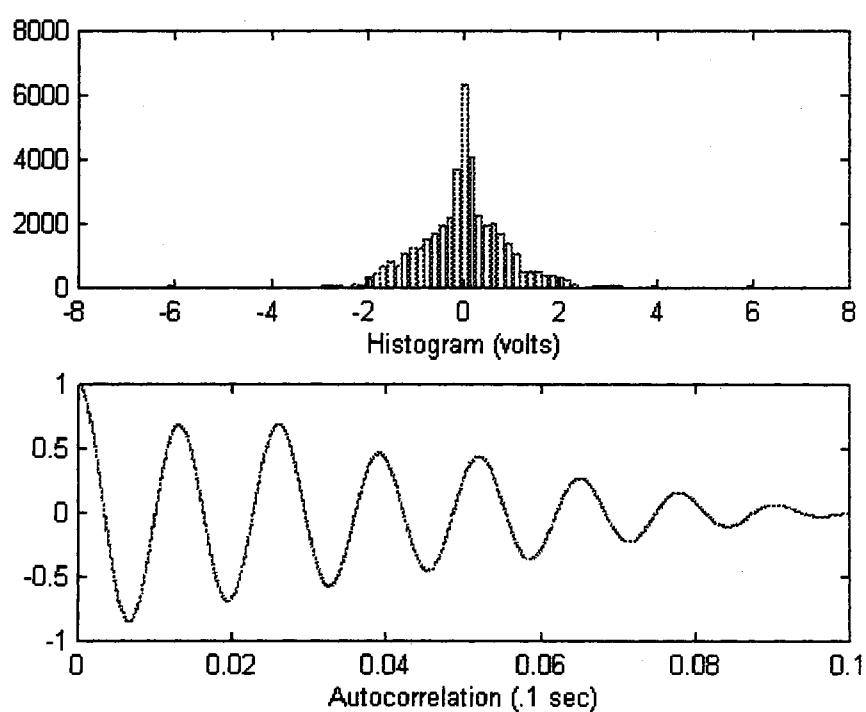


Fig. 2.6 Statistical nature of real audio signal envelope. Histogram and Autocorrelation from a popular music (voice plus background music). Mean zeroed out.

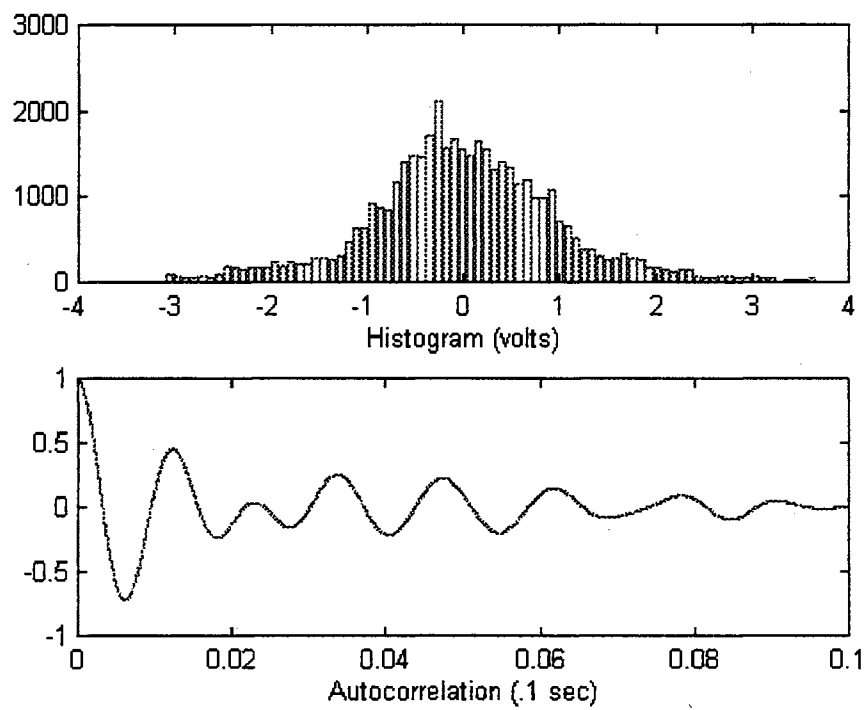


Fig. 2.7 Statistical nature of real audio signal envelope. Histogram and Autocorrelation from a classical music. Mean zeroed out.

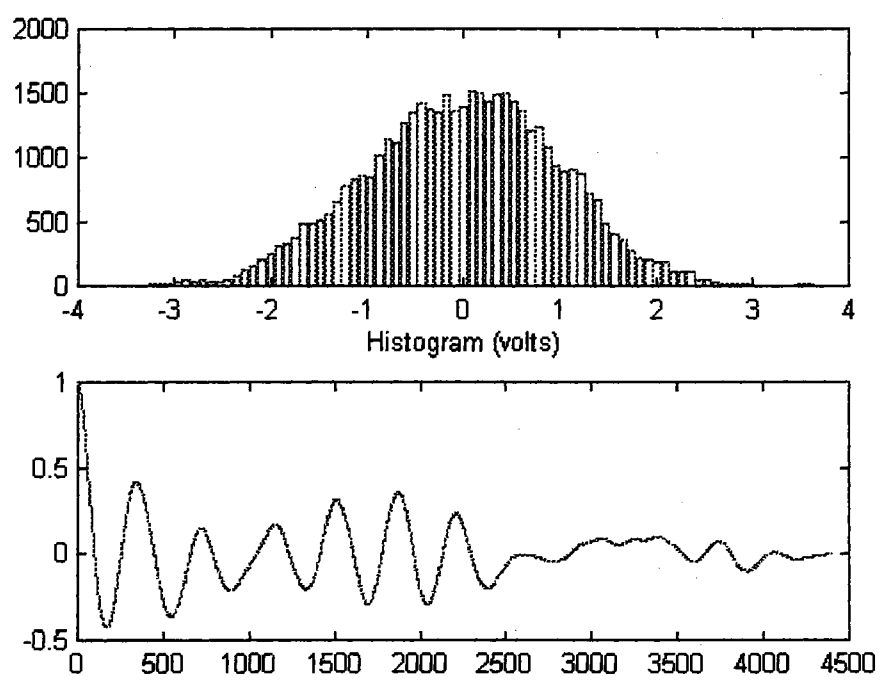


Fig. 2.8 Histogram and Autocorrelation function of model envelope.
Mean zeroed out.

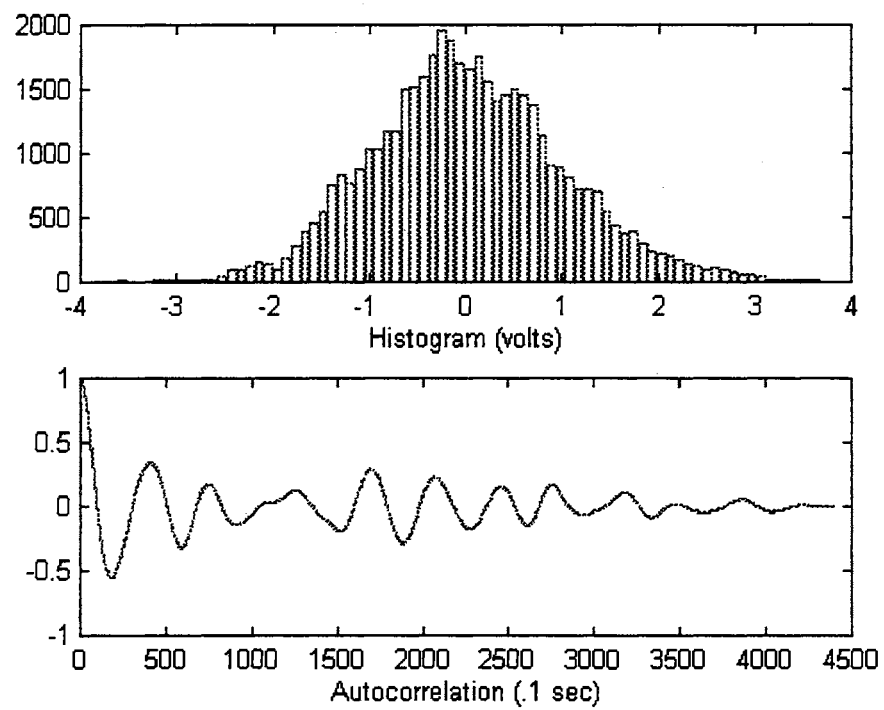


Fig. 2.9 Histogram and Autocorrelation function of model envelope.
Mean zeroed out.

In the ETHG algorithm, however, the generated output higher harmonic's amplitude will be approximately the same as that of the lower harmonic, because it was generated by using the lower harmonic structure, if no adjustment is included. To more perfectly match reality, we will have to estimate the unknown amplitude of the higher harmonics, and analyze the subjective quality of the estimated higher harmonic amplitudes. Later in Chapter 4, the generated higher harmonic's amplitude will be adjusted in a statistical sense, and the subjective quality will be analyzed.

Therefore, even though the assumption of equal amplitude is not valid for real audio signals, Eq (2.9) and Eq (2.10) are adequate models to describe the behavior of the ETHG algorithm's performance, which also renders good mathematical tractability. Details of ETHG algorithm with the model input waveform with an arbitrary envelope amplitude and single frequency will be introduced in section 2.3.3, followed by a discussion of the error characteristics of the ETHG algorithm with single frequency sinusoid using several windows in section 2.3.4.

2.3.3 ETHG Algorithm : Single Frequency, Rectangular Window Case

In this section, the ETHG algorithm will be explained in detail with a single frequency case using a rectangular window. This entire section is largely based on Scheets' work [Sche93, Sche 94].

Fig. 2.10 illustrates the concept of missing higher harmonic generation [Sche 93], and Fig. 2.11 and Fig. 2.12 depict example signals during the higher harmonic generation process. It is assumed that a signal's higher harmonic is missing due to a severely narrow transmission bandwidth. In other words, only a signal with lower harmonic structure is available at the input side of the ETHG. This input signal is multiplied by a window with length L in step 1, and becomes $w(i)$, as shown in Fig. 2.11 (a), where a 256 length rectangular window was used. Fig. 2.11 (b) is a portion of the desired signal with same length as Fig. 2.11 (a). This desired signal is the second harmonic of Fig. 2.11 (a), with the same envelope. A 256 point Fast Fourier Transform yields $W(f)$ in step 2. Fig. 2.11 (c) shows the magnitude spectrum of $W(f)$. Note the magnitude spectrum of the desired signal depicted in Fig. 2.11 (d), which is a shifted version of $W(f)$ with different height and some difference at the lower and higher frequency sections. In step 3, $W(f)$ is examined to determine the local peak frequency bin (B_{peak}), and the phase angle of this bin (B_{phase}). The voltage spectra around this bin (B_{peak}) is shifted to a bin associated with the proper harmonic. Frequency bins shifted to the right of $L/2$ are discarded. Also the phase angle of the relocated voltage spectra is rotated according to the local peak (B_{phase}) of the spectrum in step 4 to synchronize the generated output signal segments. The voltage spectra shift, and phase angle shift formulas are,

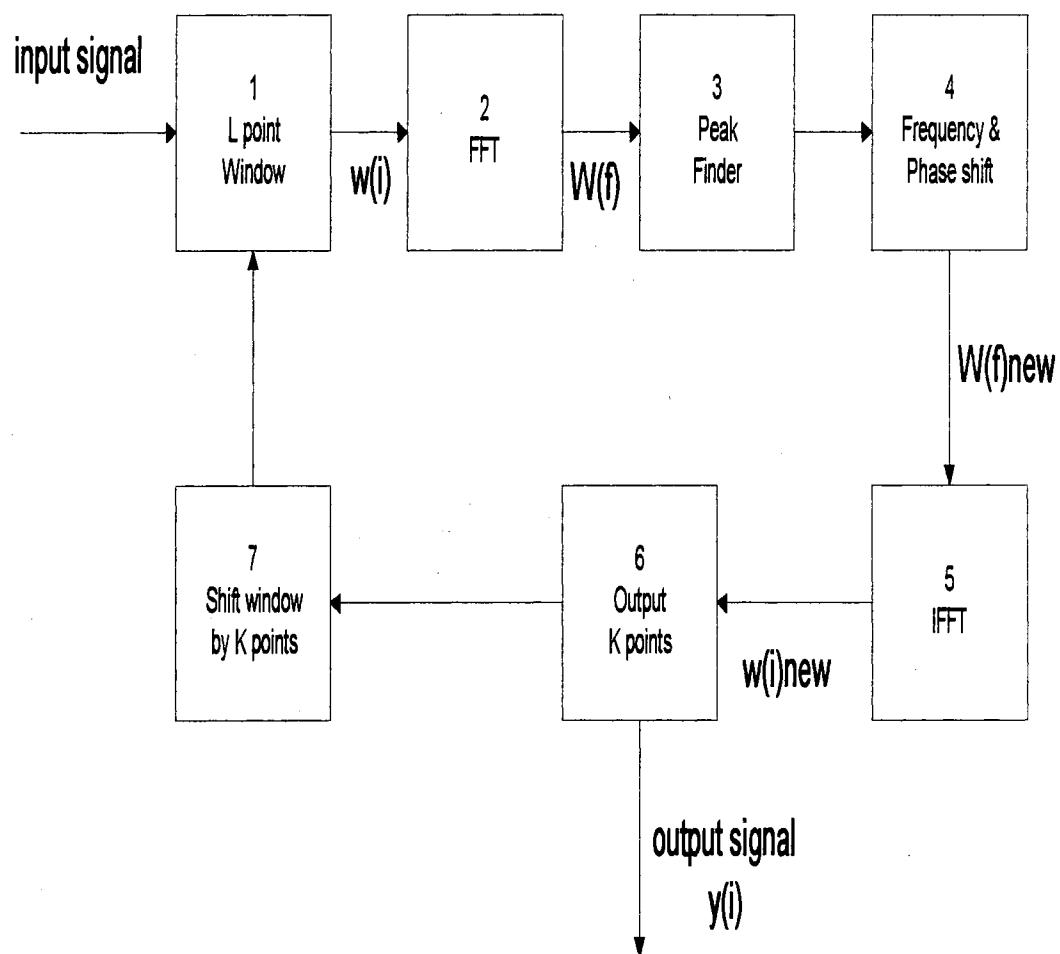


Fig. 2.10 Block diagram of the Envelope Tracking Harmonic Generator.
[Sche 93].

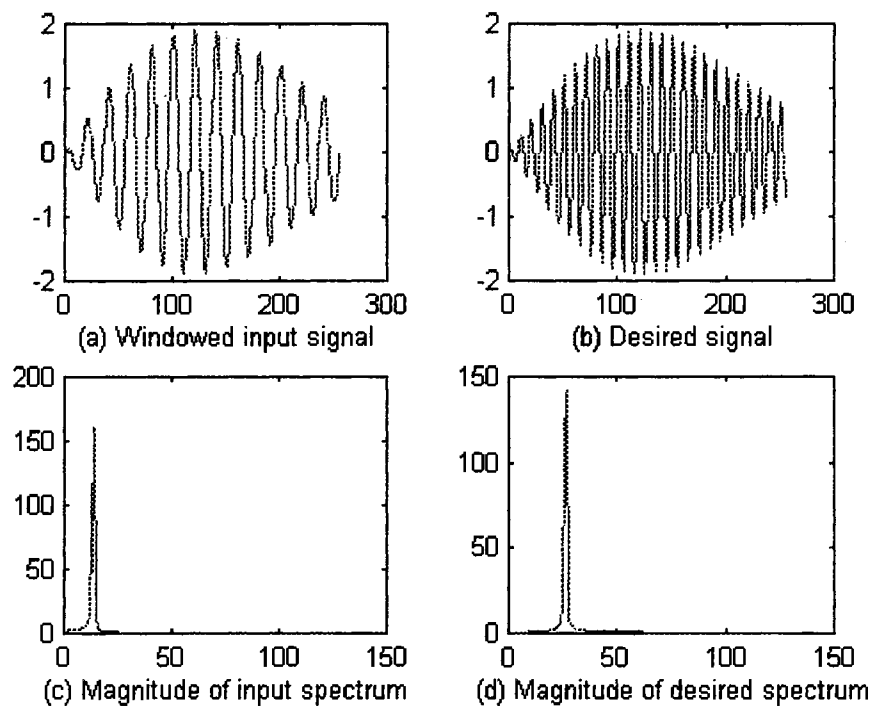


Fig. 2.11 Signals during the higher harmonic generation process.

- (a) Windowed input signal
- (b) Desired signal
- (c) Magnitude spectrum of input signal
- (d) Magnitude spectrum of desired signal

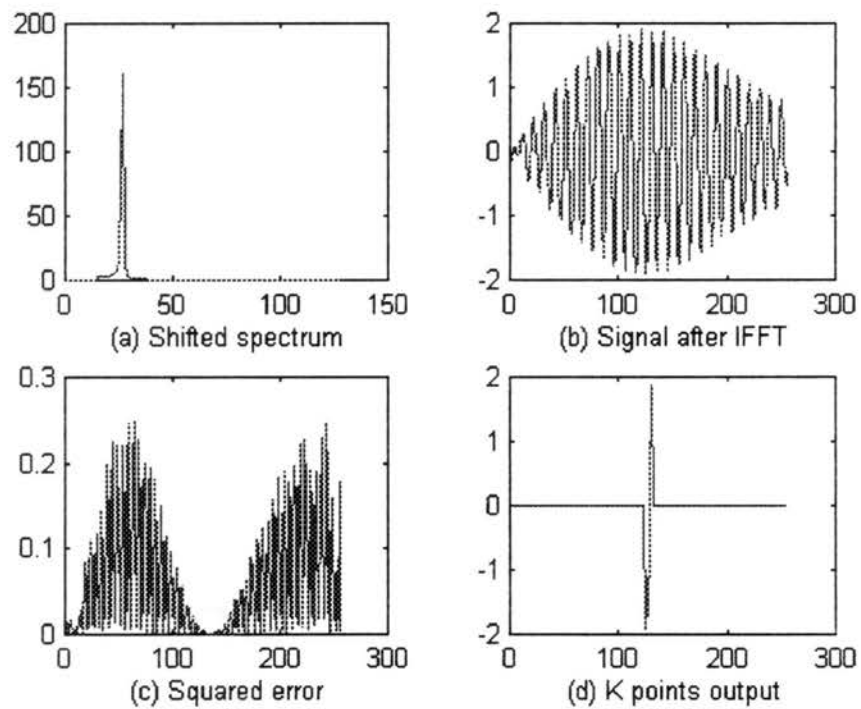


Fig. 2.12 Signals during the higher harmonic generation process.

- (a) Shifted magnitude spectrum of Fig. 2.11 (c)
- (b) Signal after IFFT
- (c) Squared error between input and the desired signal
- (d) K points output in the middle of window

$$\text{Voltage shift} = (N-1) \times B_{\text{peak}} \quad (\text{bins}), \quad (2.12)$$

and

$$\text{Phase shift} = (N-1) \times B_{\text{phase}} \quad (\text{radians}), \quad (2.13)$$

where N is the desired harmonic.

The Output signal from the step 4 is $W(f)_{\text{new}}$, which is shown in Fig. 2.12 (a). An L point IFFT is performed in step 5, which yields $w(i)_{\text{new}}$. Fig. 2.12 (b) shows $w(i)_{\text{new}}$. Then K ($K \leq L$) points from the middle section are extracted in step 6, as shown in Fig. 2.12 (d). This middle window section is the area where the estimation error tends to be at the minimum. Fig. 2.12 (c) shows the squared error between the generated output signal and the desired signal. Note the amount of squared error in the middle of the window. Then the window is shifted to the right by K points in step 7, and the process is repeated until the shifting window spans all data points of the input. Since $L \gg K$, the data points are overlapped by the window. Therefore, larger K allows less overlap in the window, which results in faster computation speed, while the amount of error increases. Finally all K outputs at each iteration are concatenated to form the final output $y(i)$, which is the estimate of the desired signal which has higher harmonic structures [Sche 93, Sche 94].

The author noticed that the shifting of the spectrum introduces several main sources of error as follows [Sche 93, Sche 94]:

- Bin resolution error is caused by the peak FFT bin and the actual unknown input frequency peak most likely not coinciding. As the FFT bins lie $fs / (L)$ Hertz apart, where fs is the sampling frequency, a shift of the spectrum based on the peak bin will usually result in the estimated harmonic frequency being incorrect. For example, with a constant amplitude sinusoid signal with a normalized frequency of .0878, and 32 length rectangular

window, the peak spectral magnitude lies in bin 3, where the frequency estimate is $3/32 = .09375$ Hertz. A spectral shift of three bins results in an estimated second harmonic at $6/32 = .1875$ Hertz, whereas the actual second harmonic lies at $.1756$ Hertz.

- A gap in the low frequency section appears, where some energy should be in that section.

- An improper mix of aliased energy occurs in the higher frequency section.

- An improper mix of negative frequency energy occurs in the shifted lower frequency section.

Despite these errors, the author showed that for a constant amplitude, single frequency signal, the minimum amount of error exists in the middle of the processed window. The author also commented that additional work is needed to better understand the performance of the ETHG algorithm under various conditions before this algorithm can be successfully applied to real audio. Areas that need further research include :

- Determining the optimum window size in the sense of minimizing the error while maximizing the processing speed.

- Determining the optimum window type to minimize error.

- Determining the applicability of this technique to audio signals.

- Determining the best manner to split the energy with a multiple frequency input signal [Sche 93, Sche 94].

- Proving that, on average, the minimum error occurs in the middle of the processed window with a time varying envelope input signal.

Following is a brief discussion of possible techniques which might be able to

reduce Scheets' initial problems with errors described above.

The error problems mentioned above are related to the FFT length, therefore their effects on the accuracy of estimation are decreased as the FFT length goes up, while the computation load increases. The bin resolution error could be reduced by using a longer window, or using zero padding when it is appropriate to do so. Using a window other than the rectangular window may also result in reduced error. Discussions on the use of other window types, and implementing zero padding to improve the performance of the ETHG algorithm are found in section 2.3.5.

The problem of dividing multiple peaks of the spectrum when the input signal is composed of multiple sinusoids may be related to the window's sidelobe magnitude. Windows with different sidelobe magnitude should be compared, with a notion that the human auditory system is more sensitive to the peaks of the spectrum than the spectral troughs [Flan 72, Rabi 78, Dell 93].

The constant amplitude sinusoid signal used in the derivation for the estimation error is not a practical model for real audio signal. Envelopes of real audio signals are dynamically varying with time. Also, it was noticed that the error is not always minimum in the middle of the output window with time varying envelopes when the rectangular window is used. Therefore, it is necessary to understand the performance of the ETHG using input signals with arbitrary time varying envelopes, and to determine where the error is consistently minimum when other types and sizes of windows are implemented. This will be discussed in section 2.3.4.

2.3.4 Error Analysis

2.3.4.1 Introduction In this section the error criteria, and the error pattern of the processed output window of the ETHG are described for the single input frequency case. In [Sche 93], the author derived the error function for a constant envelope, single frequency case, and noted that the error was minimum in the middle of the window.

However, experiments have shown that some processed time varying envelopes do not have a minimum error in the middle of the window. A question that needs to be addressed is whether or not the error in the middle of the processed window is minimum on average, i.e., is the middle the best place to extract output points? This question is discussed in the following sequence. In section 2.3.4.2 the error in the window interval is defined, followed by a discussion of the effects of window function in section 2.3.4.3. It will be noted that the truncation in frequency domain due to non-linear frequency translation in ETHG algorithm corresponds to a convolution between windowed waveform sequence and the impulse response of truncation function in time domain. This convolution, which depends on the choice of window type, results in distortions to the waveform. A prediction of the ETHG algorithm's performance, based on analysis of the amount of distortion of each window type, will be presented. After this discussion, the mean error pattern will be examined by computer simulations in section 2.3.4.4. Then it will be mathematically shown, in the mean sense, that the error approaches zero in the middle of the processed window in section 2.3.4.5.

2.3.4.2 Definition of Error The desired signal, from [Sche 93], given an L point window, is the N th harmonic of Eq (2.7) with the exact same envelope $v(i)$,

$$d(i) = v(i) \cos(N w_d i + N \theta), \quad i = 0, \dots, L-1 \quad (2.14)$$

$$\text{where } w_d = 2\pi \frac{fin}{fs} \quad (2.15)$$

We note here that even though we use the desired signal for analysis purpose, it is not available in reality. The desired signal is used to analyze the general error pattern of the processed output window of the ETHG algorithm.

The windowed input signal with length L, which is assumed to have the form of Eq (2.6), will become an estimate of Eq (2.14) after processing by the first five blocks of the ETHG in Fig. 2.4, as

$$w(i)_{new} = \hat{d}(i) = \frac{1}{w(i)} [w(i)v(i) \cos(N \hat{w}_d i + N \hat{\theta}) * h_t(i)], \quad i = 0, \dots, L-1, \quad (2.16)$$

$$\text{where } \hat{w}_d = 2\pi \frac{\hat{f} in}{fs} = 2\pi \frac{(fin + \Delta)}{fs}, \quad (2.17)$$

$$\Delta \text{ is the frequency error equal to } \hat{fin} - fin, \quad (2.18)$$

$w(i)$ is the window function, $*$ denotes convolution operator, and $h_t(i)$ is the impulse response of the truncation function which will be discussed in next section.

The error signal is obtained by subtracting Eq (2.16) from Eq (2.14), which can be written as

$$e(i) = d(i) - w(i)_{new} \quad (2.19)$$

$$= v(i) \cos(N w_d i + N \theta) - \frac{1}{w(i)} [w(i) v(i) \cos(N \hat{w}_d i + N \hat{\theta}) * h_i(i)]$$

$$i = 0, \dots, L - 1. \quad (2.20)$$

The squared error of Eq (2.20) is going to be examined in section 2.3.4.4 by computer simulations using envelopes generated by the waveform generation model depicted in Fig. 2.2, following a discussion of convolution result between the impulse response of truncation function and the window function in section 2.3.4.3. We assume that the envelope amplitude is slowly changing within a window interval, as discussed in section 2.3.2. The equations from Eq (2.14) through Eq (2.20) assumes that we are using an arbitrary L length window. The choice of window function affects the error in Eq (2.19), because it affects the spectral shape in frequency domain, which in turn affects the amount of truncation of side lobes due to the non-linear frequency translation used in the ETHG. Truncation of frequency spectrum is equivalent to the convolution between the truncation impulse response and the windowed time domain waveform, which results in distortions to the time domain waveform. The amount of distortion depends on the selection of window function. Details of the window function, the reason why the selected windows were chosen, and the distortion after the convolution will be discussed in the following section.

2.3.4.3 Analysis of the Effects of the Window Function A discussion of window functions on their roles in both time and frequency domain before the presentation of the simulation results seems to be necessary in order to provide a better understanding of the ETHG algorithm. The convolution relationship with the impulse response that truncates

the frequency content in frequency domain will be emphasized. Following is a discussion of window functions.

It is well known that the short time Fourier Transform is useful for signals which are stationary for a short period. It has been widely used in audio signal synthesis [Smit 87, Smit 91, Port 80, Moor 90], and audio signal analysis [Ando 93, Brow 93, Brow 96, Moor 90]. The short time Fourier Transform is defined as [Kay 88]

$$Y_n(e^{j\omega}) = \sum_{m=-\infty}^{\infty} x(m)w(n-m)e^{-j\omega m}, \quad (2.21)$$

where $x(m)$ is the input signal and $w(n-m)$ is the window sequence positioned at time n along the input signal. Eq (2.21) results in a convolution in the frequency domain, which affects the spectral distribution.

Since we have to use a short data sequence, the problem of resolution occurs. It is well known that time resolution and frequency resolution of the short time Fourier Transform are dependent on the window size [Oppe 89, Harr 78]. A shorter window yields poorer frequency resolution, but the time resolution will be better since the input properties are averaged over short time intervals. The larger window, on the other hand, gives poorer time resolution and better frequency resolution. Therefore, the short time Fourier Transform cannot have arbitrarily good time and frequency resolution. The time-frequency resolution is limited by the Uncertainty Principle, which is also known as the Time Bandwidth Product [Coh 95]. The Uncertainty Principle describes that the broadness of time and frequency resolution is defined as σ_n and σ_ω , respectively, where

$$\sigma_n^2 = \int (n - n_0)^2 |x(n)|^2 dn \quad (2.22)$$

and

$$\sigma_w^2 = \int (w - w_0)^2 |X(w)|^2 dw. \quad (2.23)$$

The mean time n_0 is defined as,

$$n_0 = \int \tau x(\tau) w(\tau - n) d\tau, \quad (2.24)$$

where $w(n)$ is a window function. The mean frequency w_0 is

$$w_0 = \int w |F(w)|^2 dw, \quad (2.25)$$

where $F(w)$ is the Fourier Transform of windowed signal.

The Uncertainty Principle states

$$\sigma_n \sigma_w \geq \frac{1}{2}. \quad (2.26)$$

This uncertainty principle places limits on the resolution properties of the short time Fourier Transform procedure.

Three different window types are used for comparison. The rectangular window provides the best frequency resolution, while the high proportion of energy leakage outside the main lobe is a problem. It is defined as [Oppe 89]

$$\begin{aligned} w(i) &= 1 \quad i = 0, \dots, L-1 \\ &= 0 \quad \text{otherwise.} \end{aligned} \quad (2.27)$$

The peak side lobe is down 13 dB from the main lobe level, and the rest of side lobes show fall off rate of 6 dB per octave [Jack 96]. In the ETHG algorithm, while the frequency content of a lower harmonic is translated into a higher position, the spectral

energy outside of the main lobe is truncated. Therefore, the energy outside of the main lobe, a.k.a., leakage problem, should be minimized in order to minimize the loss of frequency content during the translation.

Truncation in the frequency domain can be explained by using an ideal low pass filter with a frequency response as

$$H_t(f) = \begin{cases} 1 & \text{for } -f_c < f < f_c \\ 0 & \text{otherwise,} \end{cases} \quad (2.28)$$

where f_c is an arbitrary cut off frequency. The time domain impulse response of this filter by using the inverse Fourier Transform for finite-energy digital signals [Kunt 86] is,

$$h_t(i) = \int_{-f_c}^{f_c} e^{j2\pi fi} df \quad (2.29)$$

$$= \frac{\sin(2\pi f_c i)}{\pi i} \quad (2.30)$$

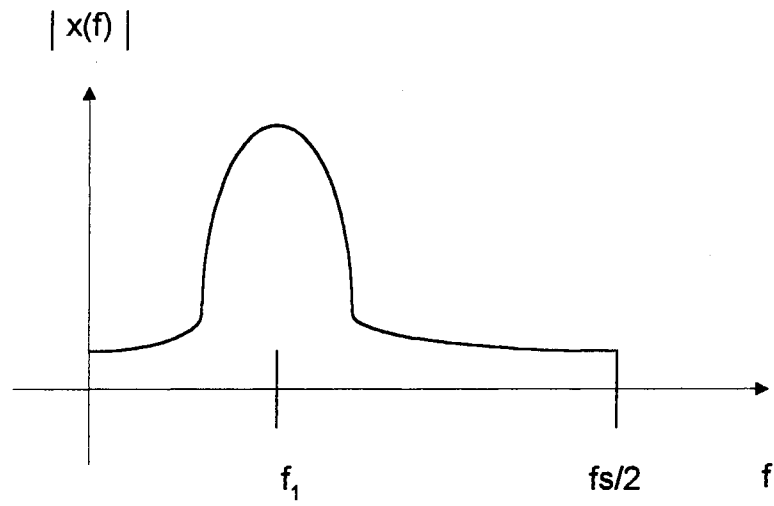
We may express Eq (2.21) as

$$Y(f) = X(f) * W(f) \quad (2.31)$$

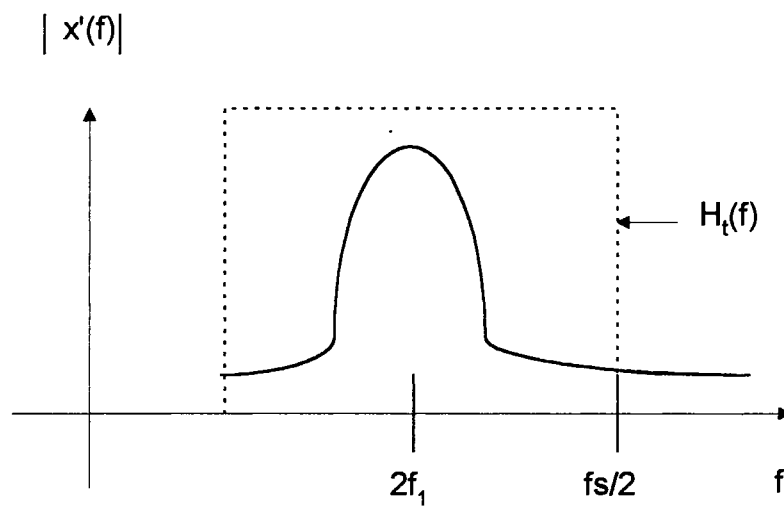
where $*$ denotes convolution notation. Since the truncation of $Y(f)$, after frequency translation, by an ideal LPF corresponds to a multiplication in frequency domain, Eq (2.31) becomes

$$Y'(f) = [X'(f) * W'(f)] H_t(f), \quad (2.32)$$

where $X'(f)$ is the translated spectrum to an arbitrary harmonic frequency N , and the $W'(f)$ is the translated window function. This concept is illustrated in Fig. 2.13. Fig. 2.13 (a) and (b) shows an example of original spectrum and the truncated spectrum after frequency translation, respectively. Eq (2.32) corresponds to a convolution of the window



(a)



(b)

Fig.2.13 Spectrum translation. Note that this can be represented as the original spectrum $|X(f)|$ shifted and multiplied by a rectangular window $H_t(f)$.

function with the ideal low pass filter in time domain, and may be written as,

$$y'(i) = x'(i) w(i) * h_t(i), \quad (2.33)$$

where $x'(i)$ is the estimated signal with N th harmonic frequency and the same envelope shape. We see in Eq (2.33) that the result $y'(i)$ depends on the choice of window function. In the following discussion, we will see the convolution results between the selected window functions and the truncation impulse response.

When the impulse response in Eq (2.29) is convolved with the rectangular window, we would see ripples in the window interval, which is the Gibbs phenomenon [Jack 96, Oppe 89]. It was noticed that the amount of the ripples in both corners of the rectangle stays the same regardless of the window length. However, the amount of the ripples in the middle of the window is decreased with a longer window length. For example, Fig. 2.14 (a) shows the result of convolution when the length of the rectangular window is 64, and Fig. 2.14 (b) is when the length is 256. The solid line is the rectangular window, and the circles represent the convolution result. We note here that as the window length increases, the amount of ripples in the middle of the window is decreased, however this still causes small distortions in the time domain waveform.

This ripple problem is reduced when the window is tapered on both sides. The tapering of the window in time results in smaller frequency domain side lobes, but the width of the main lobe gets wider as well. When the spectrum with a tapered window is truncated outside of the main lobe, it will lose less energy compared to the rectangular window case. The tapered window also generates a smoothed convolution result in the

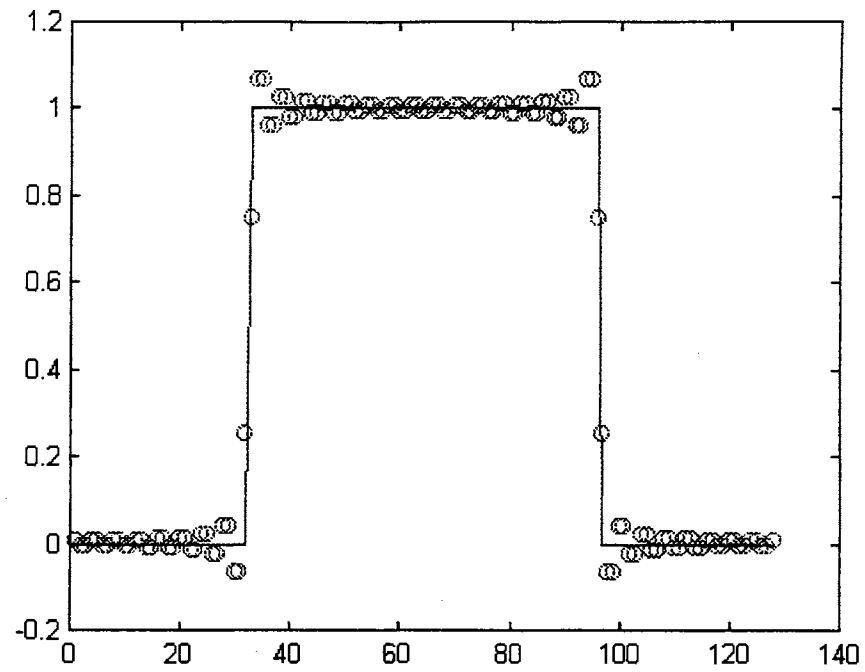


Fig. 2.14 (a) Convolution result with 64 length rectangular window.

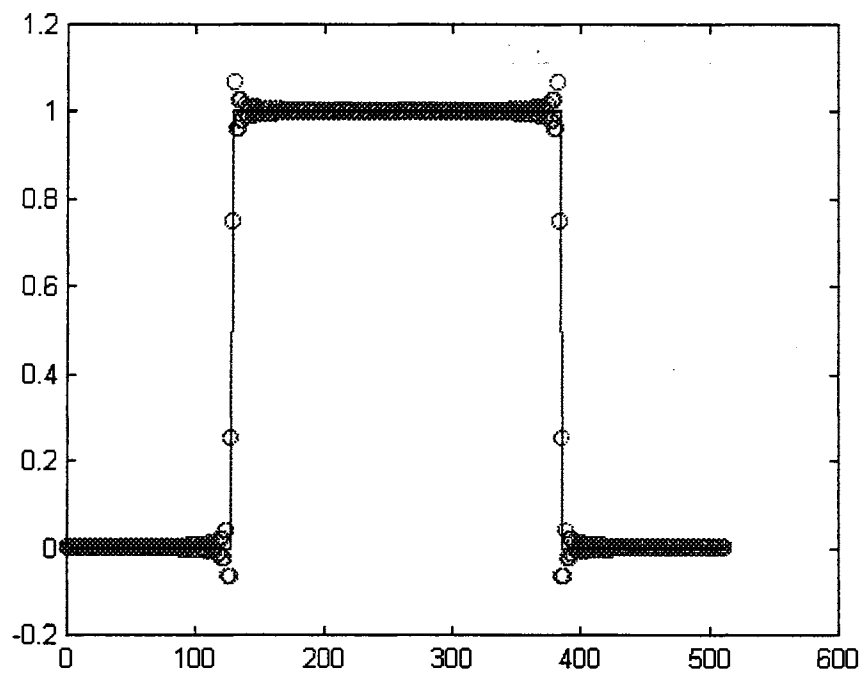


Fig. 2.14 (b) Convolution result with 256 length rectangular window.

time domain, with the impulse response of the truncation function. The ripple problem in the window interval will be reduced, and thus less distortions to the time waveform could be achieved. The Hamming and Hanning window are used for performance comparison because they are easy to implement, provide adequate frequency resolution [Harr 78], and preliminary tests on other types of more complicated windows indicated that there exists negligible performance difference when compared to the Hanning window. The Hamming window is [Oppe 89]

$$w(i) = .54 - .46 \cos\left(\frac{2\pi i}{L-1}\right) \quad i = 0, \dots, L-1 \quad (2.34)$$

$$= 0 \quad \text{otherwise.}$$

The peak side lobe is -41 dB from the main lobe level, but the fall off rate of the remaining side lobes is 6 dB per octave which is same as that of the rectangular window [Jack 96]. We note here that both ends of this window are not zero, thus we see some small fluctuations at both ends of window interval after the convolution with the impulse response in Eq (2.29), which is much smaller than those with the rectangular window case. Fig. 2.15 (a) and (b) shows the example of convolution result when the window length is 64 and 256, respectively. The truncation length was $L/16$ of the FFT length. The solid line is the window, and the circles represent convolution result. Note the smoothness of the convolution result, except the slight fluctuations at both ends of the window. The sum of the absolute error of middle $L/2$ points between the window and the convolution result for the case (b) is .0292, whereas the sum of the absolute error for the rectangular window case in Fig. 2.14 (b) is .3732.

Although the amount of ripple is much decreased compared to the rectangular window case, and therefore the distortion in the middle of the window is decreased

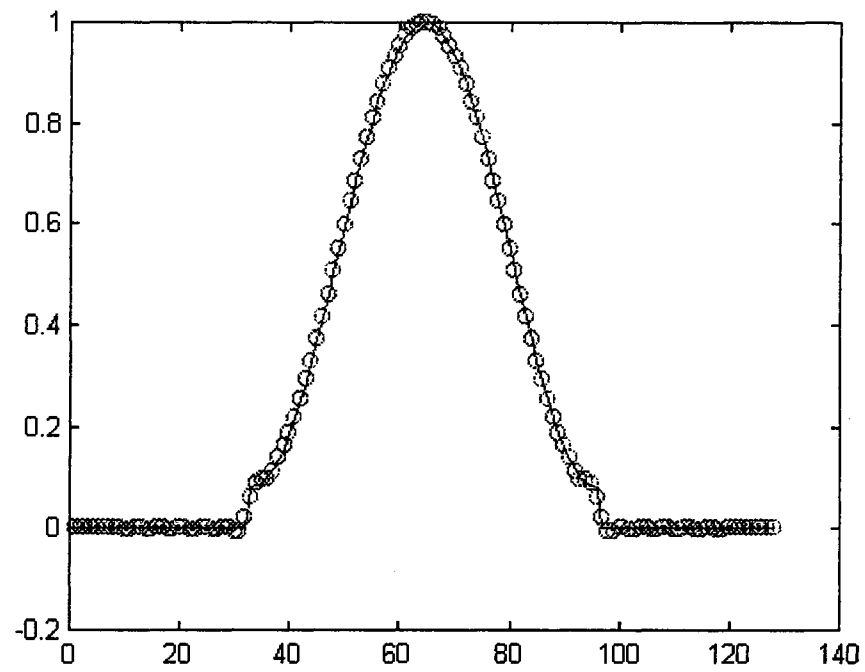


Fig. 2.15 (a) Convolution result with 64 length Hamming window.

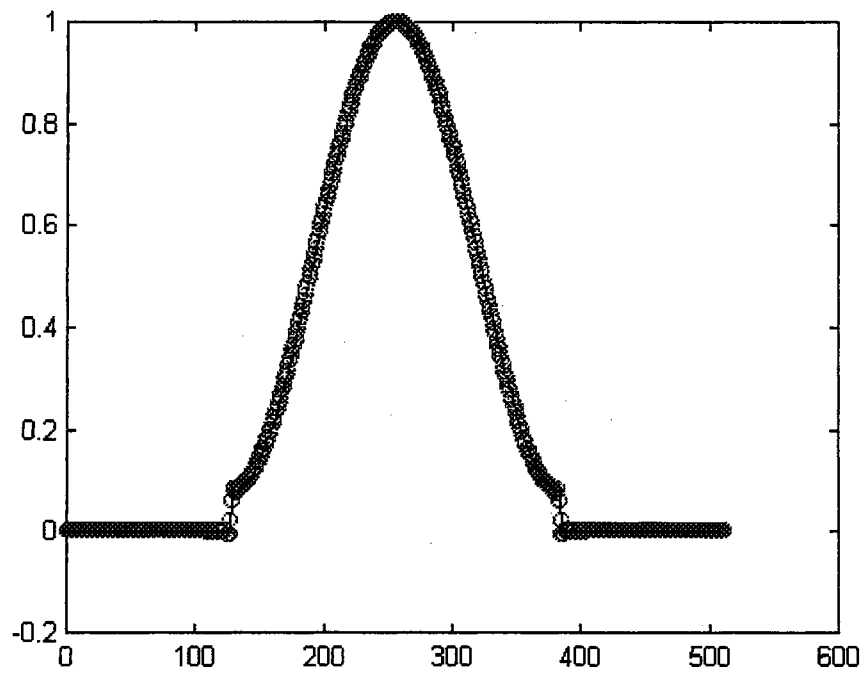


Fig. 2.15 (b) Convolution result with 256 length Hamming window.

considerably, it could be reduced further. It was noticed that the Hanning window yields less error in the middle of the window, compared to the Hamming window case. The Hanning window is [Oppe 89]

$$w(i) = .5 - .5 \cos\left(\frac{2\pi i}{L-1}\right) \quad i = 0, \dots, L-1 \quad (2.35)$$

$$= 0 \quad \text{otherwise.}$$

The peak side lobe level is -31 dB from the main lobe level, but the fall off rate of the remaining side lobes is 18 dB per octave [Jack 96], which is the best fall off rate among the compared windows. Fig. 2.16 (a) and (b) shows the convolution result with the impulse response of the truncation function. Note that both ends are much smoother than the Hamming window case. The sum of the absolute error, with a same truncation length, between the Hanning window and the convolution result for this case is .0016, which is the smallest value. Therefore, we may say this type of window function yields the least amount of distortion to the time domain waveform after the convolution. Note here that the ideal window would be the one which does not introduce any distortion after the convolution, and at the same time it should not have a broad main lobe or a higher rate of side lobe fall off rate in the frequency domain as well. It would be impossible to obtain such window because of the limitation by the Uncertainty Principle. Table 2.1 shows the examples of sum of absolute error of middle $L/2$ points between the convolution result and the window function. The truncation length was $L/8$ for (a) and $L/16$ for (b), where L is the FFT length. We observe from the table that the Hanning window apparently produces least amount of error. This observation enables us to predict that the performance of the ETHG algorithm would be at its best with the Hanning window when the simulation data and the real audio signals are processed. We will see that the Hanning window, which

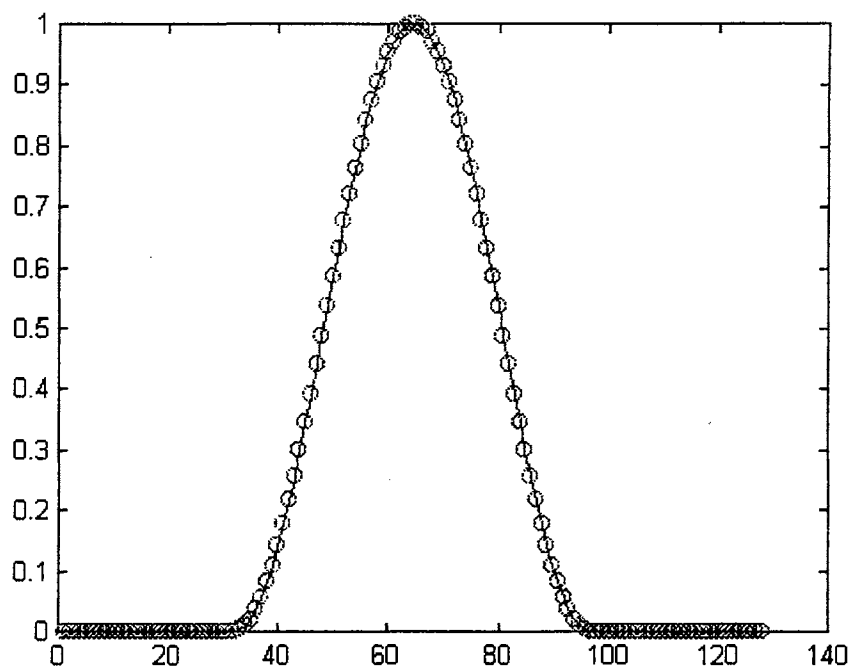


Fig. 2.16 (a) Convolution result with 64 length Hanning window.

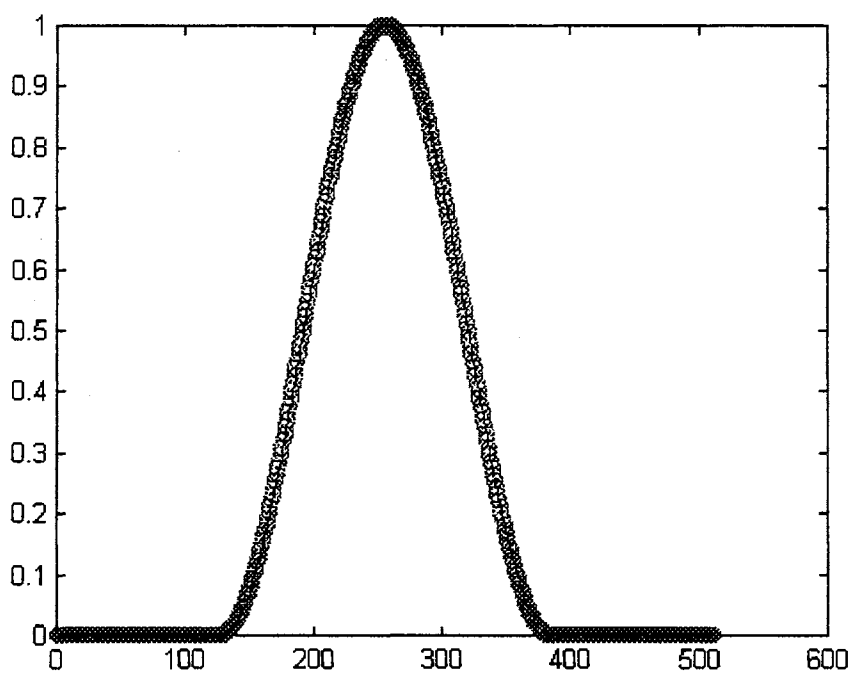


Fig. 2.16 (b) Convolution result with 256 length Hanning window.

has the smallest fall off rate of side lobes in the frequency domain and the least amount of distortions to the waveform after the convolution in time domain, yields the lowest error in section 2.3.4.4, section 2.3.5, and section 2.3.6, which in turn results in acceptable amount of distortion level when the real audio signal is processed in Chapter IV.

<i>Window Length</i>	<i>Rectangular</i>	<i>Hamming</i>	<i>Hanning</i>
64	.2085	.0154	.0035
128	.1981	.0155	8.2766×10^{-4}
256	.1931	.0154	2.0174×10^{-4}
512	.1907	.0152	4.9879×10^{-5}
1024	.1895	.0152	1.2407×10^{-5}

(a)

<i>Window Length</i>	<i>Rectangular</i>	<i>Hamming</i>	<i>Hanning</i>
64	.4228	.0301	.0305
128	.3865	.0285	.0068
256	.3732	.0292	.0016
512	.3675	.0292	4.005×10^{-4}
1024	.3648	.0291	9.95×10^{-5}

(b)

Table 2.1 Examples of sum of absolute error of middle $L/2$ points between the window and the convolution result.

- (a) Truncation length is $L/8$ of FFT length.
- (b) Truncation length is $L/16$ of FFT length.

2.3.4.4 Error Pattern As stated in the beginning of section 2.3.4, the squared error is not always minimum in the middle of the processed window with some signals. This phenomenon will be examined in the following discussion. For this analysis, a 512 length rectangular window was used.

The squared error pattern of the processed output window appears to depend on the position of the input signal energy in the window interval. When the envelope amplitude of an input signal is nearly constant within the window interval, or the majority energy is located in the middle of the window, the error appears to be minimum in the middle of the processed window. Fig. 2.17 (a) shows a signal with nearly constant envelope, and Fig. 2.17 (b) shows the squared error between the desired signal and the processed signal. The horizontal axis represents positions in the window interval. Fig. 2.18 shows a case when the majority energy section is located in the middle of the window. In both figures, we see the squared errors are small in the middle of the window.

When the major energy section is not located in the middle of the window, the minimum error position tends to be off center. Fig. 2.19 shows the squared error when the major energy section of the input signal is located in the right side of the center of the window. The minimum error appears to be located in the right side of the middle of the window. Likewise, Fig. 2.20 shows the squared error when the major energy section of the input signal is in the left side. It is obvious that a low error position is moved to the left of the center.

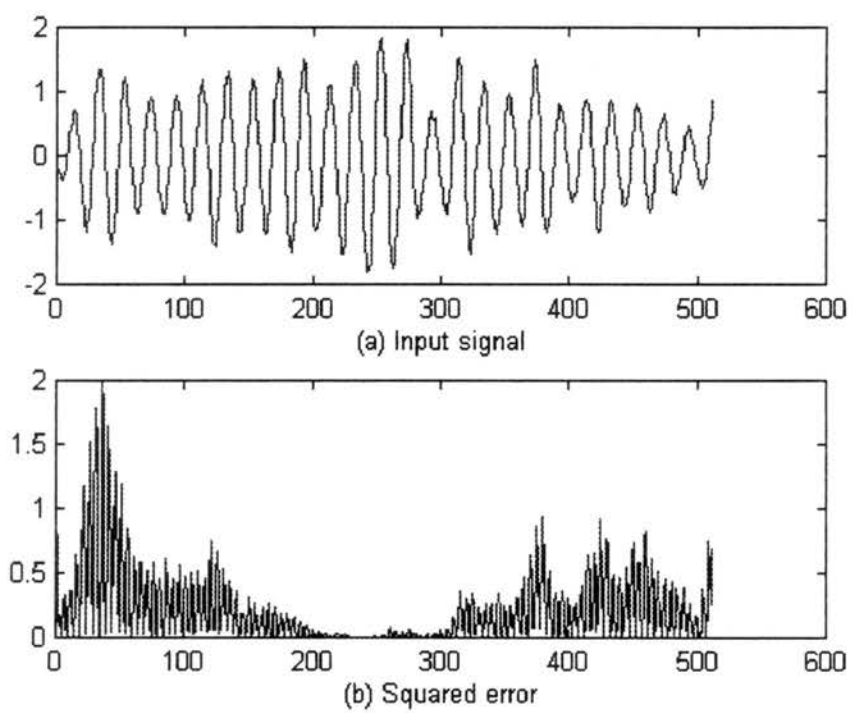


Fig.2.17 Minimum squared error position (middle).

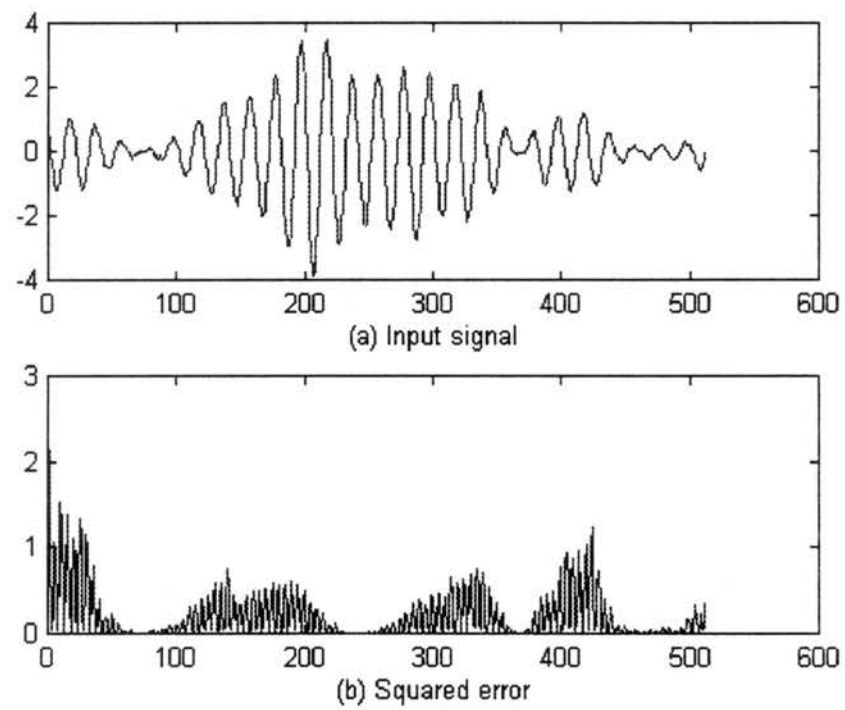


Fig. 2.18 Minimum squared error position (middle)

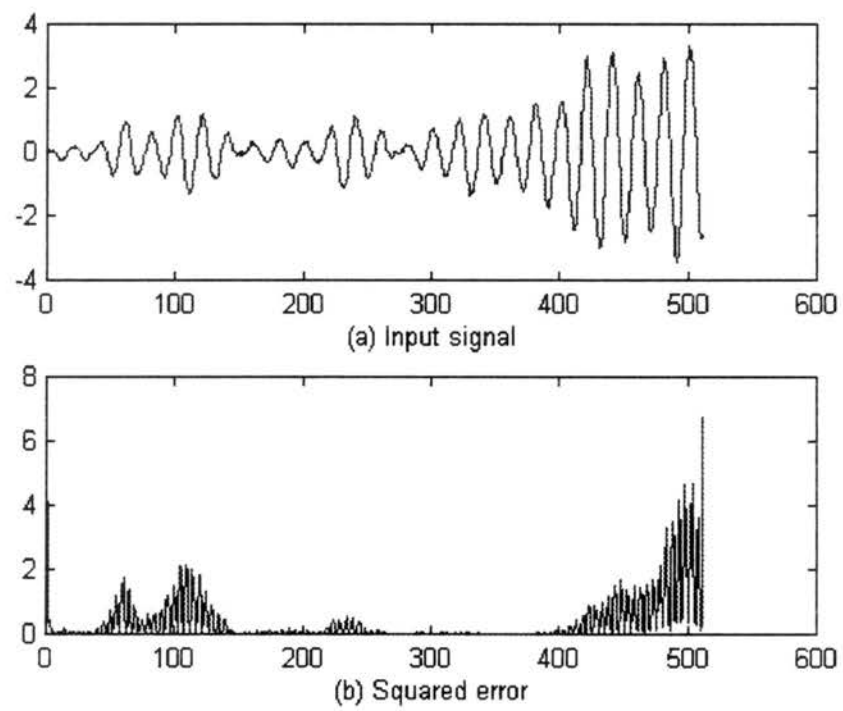


Fig.2.19 Minimum squared error position (right of middle).

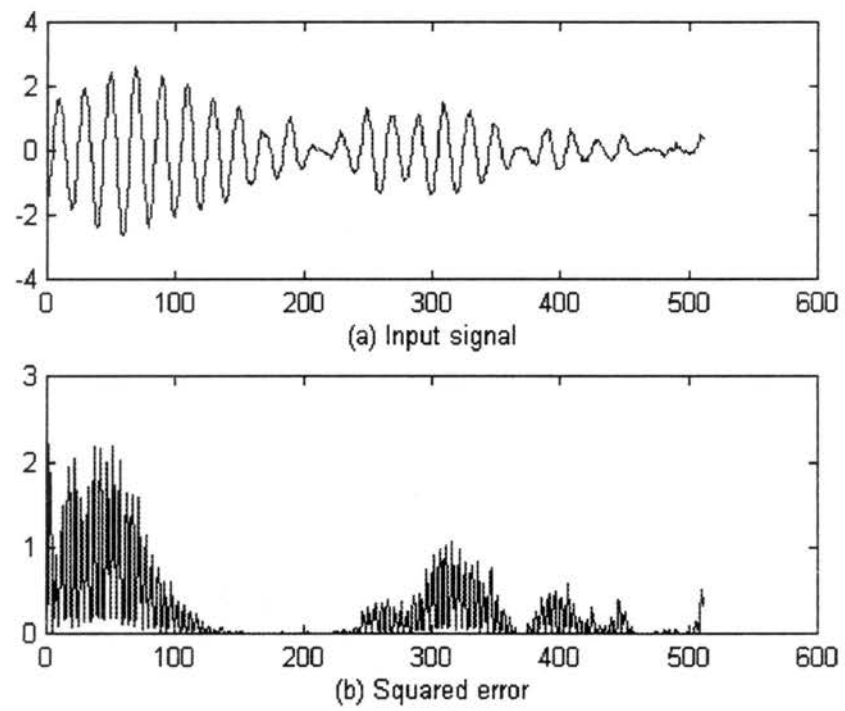


Fig.2.20 Minimum squared error position (left of middle).

The shifting of the minimum squared error position noted above brings a question whether the middle section is the best place to extract output estimates or not. To find an answer for this, the average squared error pattern is investigated.

To examine the mean squared error pattern, computer simulations were done as follows. The rectangular, Hamming, and Hanning windows were used for this simulation. Also, zero padding was implemented to all three window types to see the effects on error. The zero padding increases the resolution of the frequency spectrum, at the expense of increased computation load. The input signals are sinusoids with frequencies of .05 ~ .225, which modulate the envelopes that are generated by the envelope generation scheme in Fig. 2.3. Fig. 2.21 (a) shows the mean squared error (MSE) pattern at eight normalized frequencies, from .05 to .225, and the average MSE over all frequencies. The upper plot shows the MSE of 1000 trials at each normalized frequency, with a 512 length rectangular window, and the lower plot shows the average MSE, where the horizontal axis represents a position in the window. It appears that the average MSE is minimum in the middle of the window, and the middle section shows a shape of approximately quadratic curve. Likewise, Fig. 2.21 (b) shows the mean squared error and the average of MSE when 512 zeros are padded. Note the average MSE has decreased compared to Fig. 2.21 (a). Fig. 2.22 (a) shows plots of MSE and average MSE when the Hamming window is used. Fig. 2.22 (b) shows same plots when 512 zeros are padded. They also have minimum MSE in the middle of the window. Note the average MSE for the zero padded case decreased more, compared to Fig. 2.22 (a) and Fig. 2.21 (b). Likewise, Fig. 2.23 (a) and (b) shows MSE and average MSE when the Hanning window and zero padded Hanning window is used, respectively. Note that the zero padded Hanning window yields the

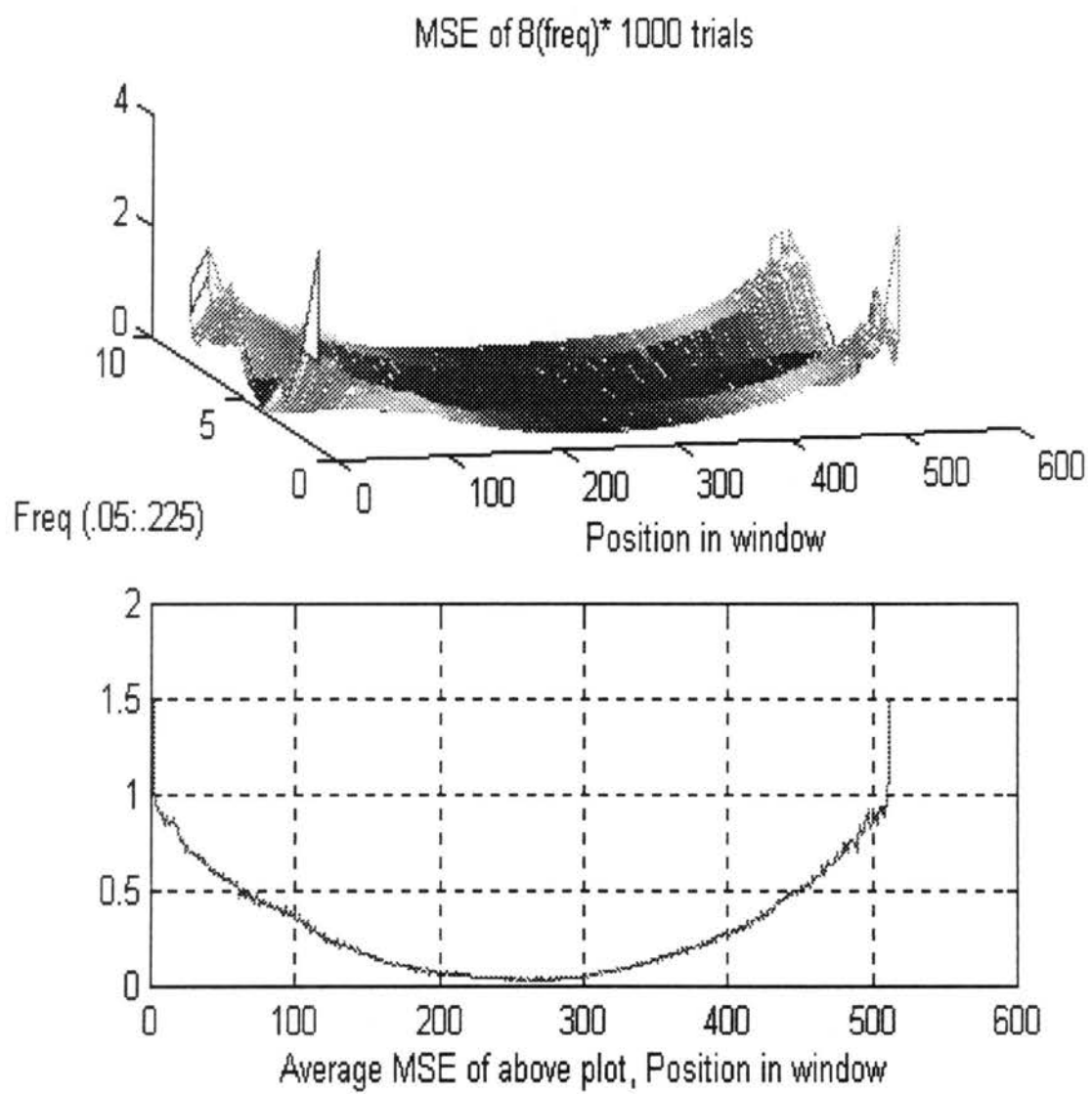


Fig.2.21 (a) MSE and average MSE, Rectangular window.

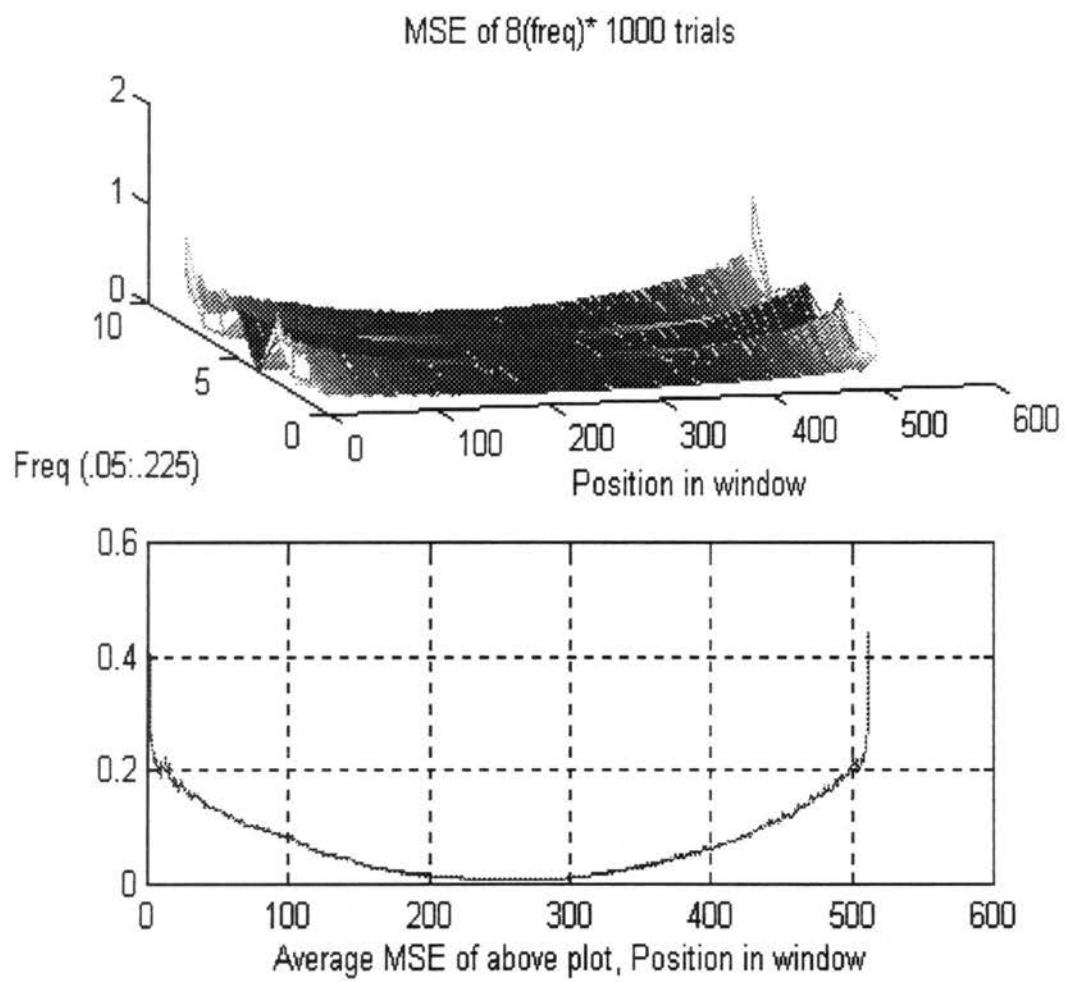


Fig.2.21 (b) MSE and average MSE, Rectangular window, 512 zeros padded.

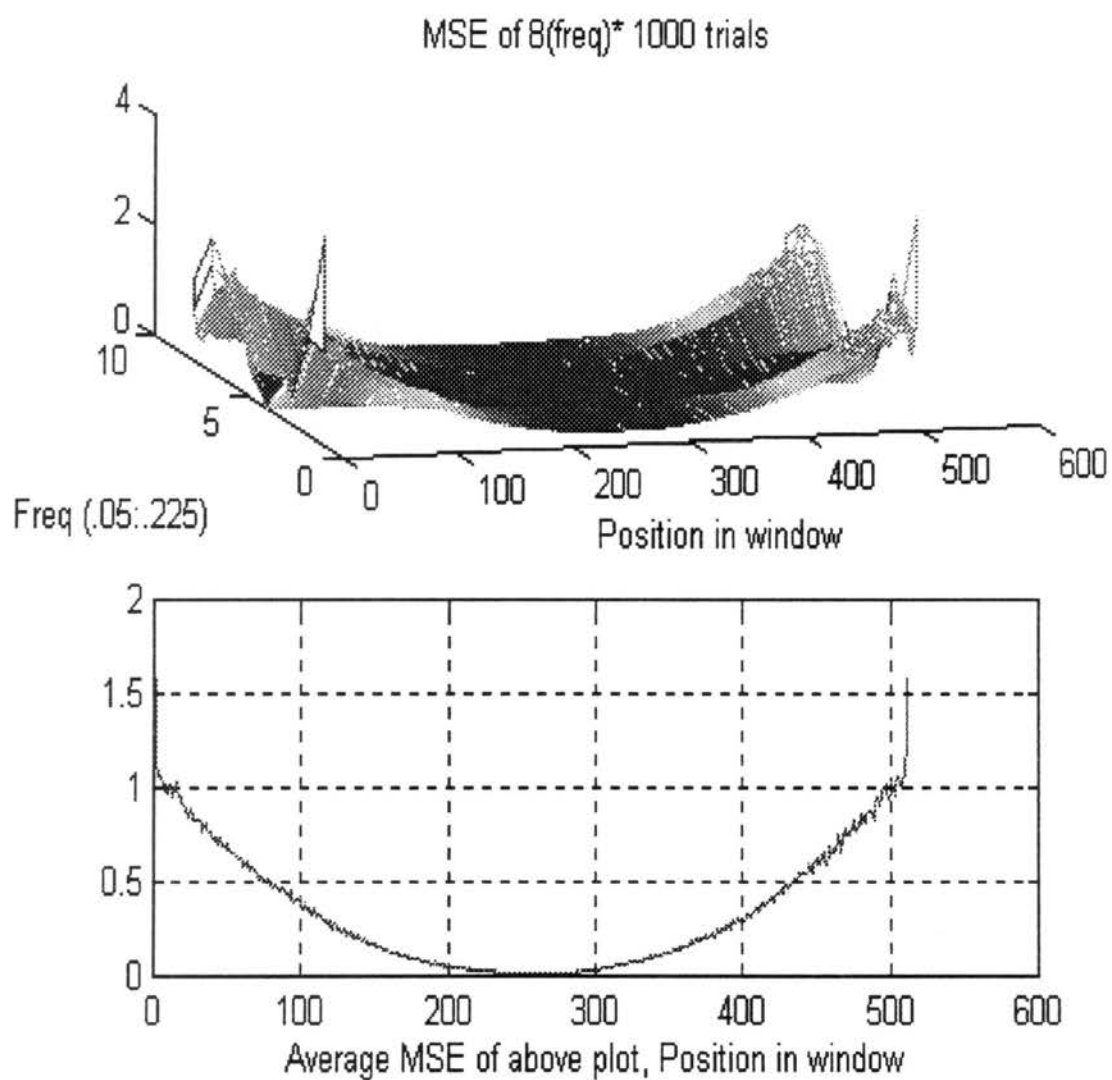


Fig. 2.22 (a) MSE and average MSE, Hamming window.

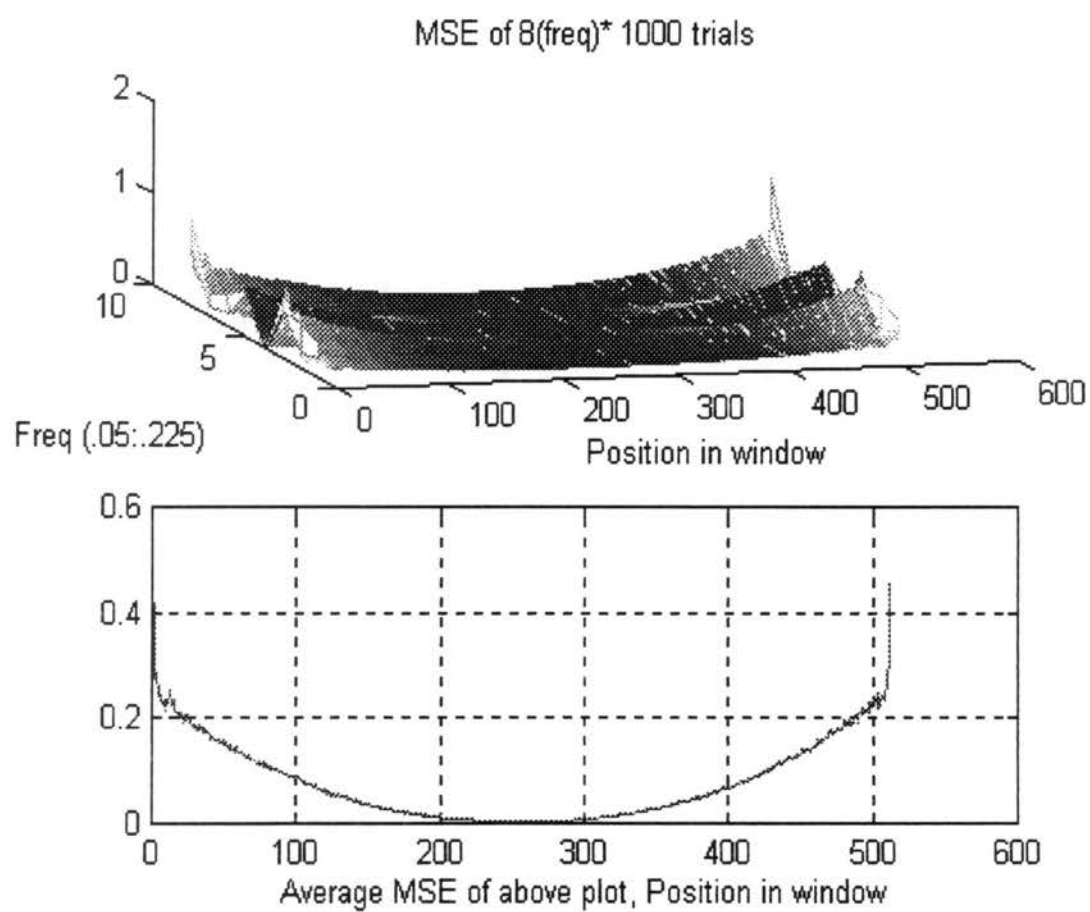


Fig. 2.22 (b) MSE and average MSE, Hamming window, 512 zeros padded.

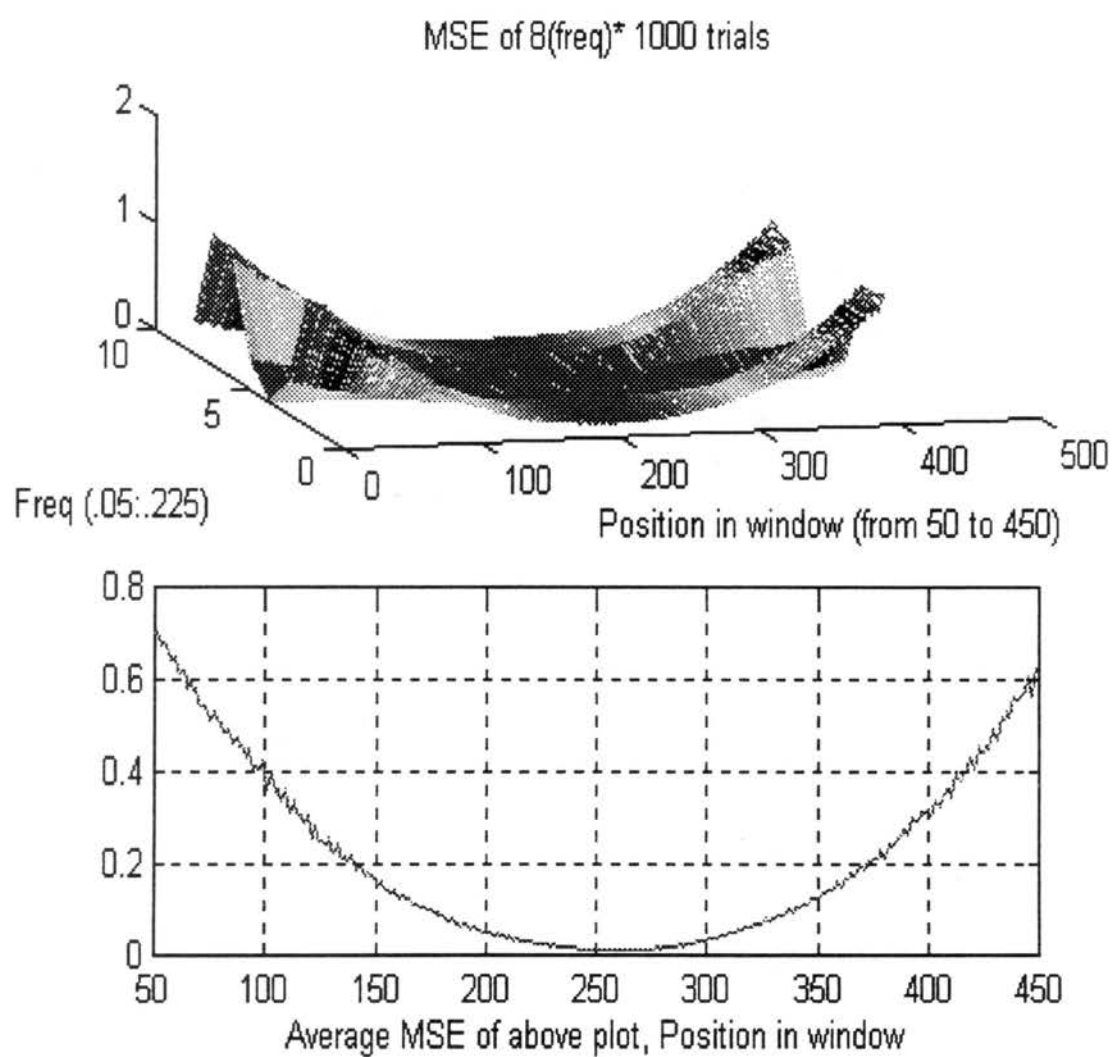


Fig. 2.23 (a) MSE and average MSE, Hanning window.

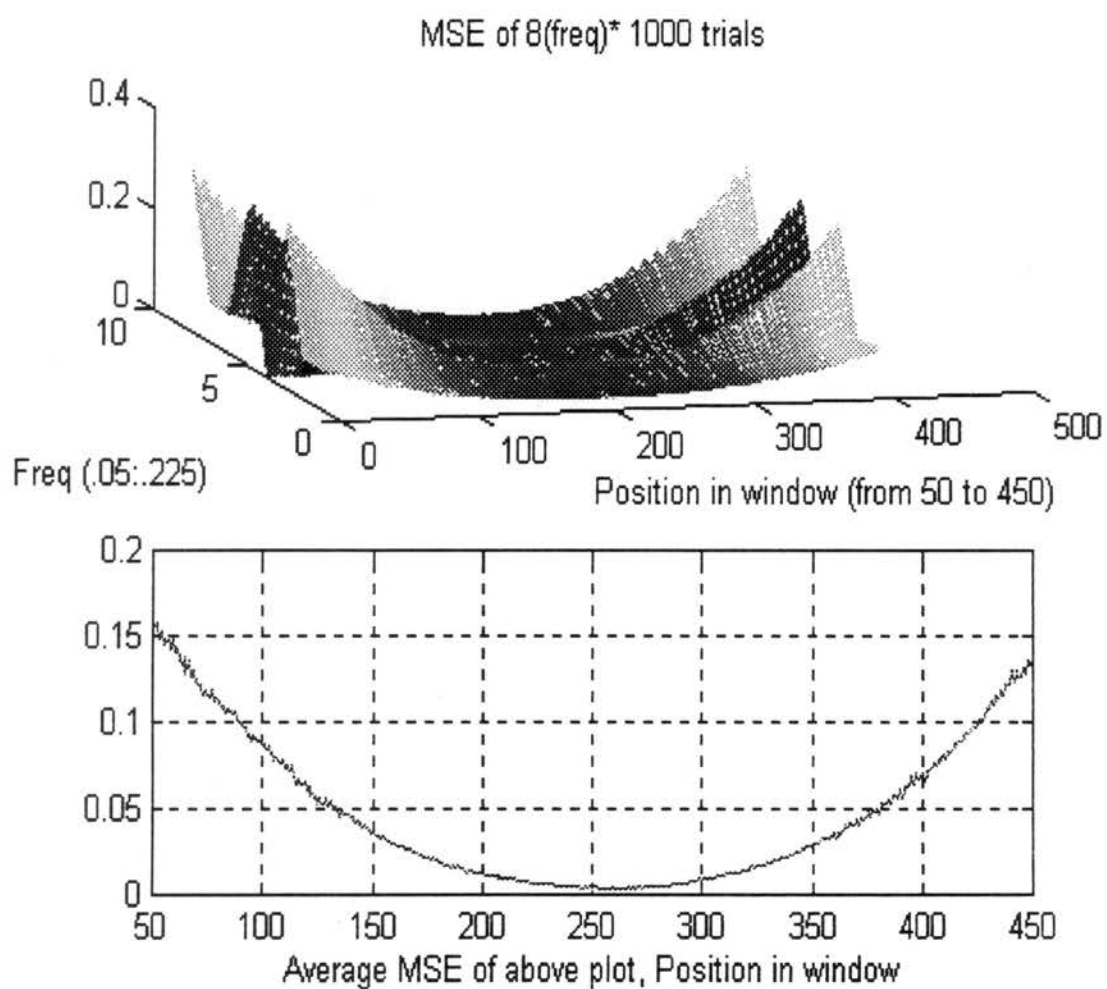


Fig. 2.23 (b) MSE and average MSE, Hanning window, 512 zeros padded.

lowest average MSE, and the rectangular window yields the highest average MSE among the compared windows. These error differences will be intensively investigated in section 2.3.5, by computer simulations.

These simulation results appear to indicate that the squared error of the processed signal is not only, on average, minimum in the middle of the processed window, but also it could be reduced by implementing window functions other than the rectangular window, and by applying zero padding. Therefore, the middle of the window seems to be the best place to draw the estimated output signals, in the mean squared error sense.

We note here that even though the author in [Sche 93] assumed the envelope amplitude is constant, and derived that the error is minimum in the middle of the processed window, that assumption generally holds, in the mean squared error sense, for the sinusoid with time varying envelope, which are assumed to be slowly varying in the window interval.

We observed in this section that the average error appears to be minimum in the middle of the processed output window. A remaining question is whether this average error pattern can be mathematically described or not. In next section, it will be shown mathematically that the error is, on average, minimum in the processed output window.

2.3.4.5 Error Analysis : Theoretical Derivation

We observed in the previous section that the error is minimum in the middle of the window when the envelope within the window interval is nearly constant or when the majority energy section is located in the middle of the window. When the signal's major energy section is not located in the middle of the window, the minimum error point tends to slightly off the center. Nevertheless, we noticed that the mean squared error is minimum in the middle of the window.

In this section it will be mathematically shown that, on average, the error is minimum in the middle of the window, i.e., the error function in Eq (2.20) approaches zero in the middle of the window for a single frequency signal with a time varying envelope. The derivation here follows that in [Sche 93].

We observed in section 2.3.4.3 that the convolution between window and the impulse response of truncation function results in some amount of distortion to the waveform. It was also noted that as the window length increases the distortion amount in the middle of the window decreased to a smaller level, as shown in Table 2.1. The Hanning window produced the least amount of distortion after the convolution among the compared windows. Therefore with an assumption that the window length is long, the following equation in the right hand side of Eq (2.20) may be written as,

$$\frac{1}{w(i)} [w(i)v(i) \cos(N \hat{w}_d i + N \hat{\theta}) * h_t(i)] \quad (2.36)$$

$$\approx v(i) \cos(N \hat{w}_d i + N \hat{\theta}) \quad i = 0, \dots, L - 1. \quad (2.37)$$

We assume that Eq (2.37) also holds for the rectangular window case, under the condition that the window length is long enough, to simplify mathematical derivation, even though we are aware that there exists Gibbs phenomenon on both sides of the rectangular window

after the convolution. Since we are taking estimates of the desired signal from the middle of the processed output window, we may make such assumption as in Eq (2.37), in the limited sense for the rectangular window case, to simplify the mathematical derivation.

Then Eq (2.20) may be written as,

$$e(i) \approx 2v(i) \sin\{.5(\hat{w}_d - w_d)i + .5N(\hat{\theta} - \theta)\} \sin\{.5(\hat{w}_d + w_d)i + .5N(\hat{\theta} + \theta)\},$$

$$i = 0, \dots, L - 1. \quad (2.38)$$

Assuming the envelope is independent of the sinusoid terms as described in section 2.3.2, and taking expected notation to Eq (2.38) yields

$$E[e(i)]$$

$$\approx 2E[v(i)] E[\sin\{.5(\hat{w}_d - w_d)i + .5N(\hat{\theta} - \theta)\} \sin\{.5(\hat{w}_d + w_d)i + .5N(\hat{\theta} + \theta)\}],$$

$$i = 0, \dots, L - 1. \quad (2.39)$$

Eq (2.39) becomes zero when the low frequency term is equal to zero, i.e., when the frequency and the phase estimate is precise. It will be shown in the following discussion that, on average, the low frequency term approaches zero in the middle of the window, thus making the error signal approach zero for signals with time varying envelope case.

The frequency and the phase estimates are found by using the following methods.

The first step to estimate the input frequency is to locate the local peak of the voltage spectrum. The local peak bin B_{peak} is where the spectrum has the maximum magnitude.

It is found from

$$B_{peak} = \text{INT}[L \frac{fin}{fs}], \quad (2.40)$$

where $\text{INT}[\bullet]$ operator indicates rounding off to the nearest integer, and L is the FFT

length. There are L bins, and the frequency separation of each bin is

$$\Delta f = \frac{f_s}{L}. \quad (2.41)$$

The frequency estimate \hat{f}_{in} is found by,

$$\hat{f}_{in} = B_{peak} \frac{f_s}{L}. \quad (2.42)$$

This gross estimate can be off by as much as $\frac{f_s}{2L}$ Hertz, however as the FFT length

$L \rightarrow \infty$ the spectrum interval Δf decreases, and the estimated frequency \hat{f}_{in} converges to the real value f_{in} .

$\hat{\theta}$ is the phase angle of the FFT at the estimated peak frequency bin B_{peak} .

The Discrete Fourier Transform definition when the rectangular window is used is

$$X(f) = \sum_{i=0}^{L-1} x(i) e^{-j2\pi f i / L}. \quad (2.43)$$

Evaluating at $f = B_{peak}$ yields

$$X(B_{peak}) = \sum_{i=0}^{L-1} v(i) \cos(2\pi \frac{f_{in}}{f_s} i + \theta) e^{-j2\pi B_{peak} i / L} \quad (2.44)$$

$$= \sum_{i=0}^{L-1} v(i) \cos(2\pi \frac{f_{in}}{f_s} i + \theta) [\cos(2\pi B_{peak} i / L) - j \sin(2\pi B_{peak} i / L)], \quad (2.45)$$

where $j = \sqrt{-1}$.

In addition to the assumption that the envelope is independent of the sinusoidal frequency, we also assume that the envelope is slowly varying within a short window interval, as discussed in section 2.3.2, and the sinusoidal frequency is a constant with a constant phase. The mean voltage spectrum at $f = B_{peak}$ Hertz is then,

$$\begin{aligned}
& E[X(B_{peak})] \\
&= E\left[\sum_{i=0}^{L-1} v(i) \cos(2\pi \frac{fin}{fs} i + \theta) \{\cos(2\pi B_{peak} i/L) - j \sin(2\pi B_{peak} i/L)\}\right] \quad (2.46)
\end{aligned}$$

$$\begin{aligned}
&= E\left[v(0) \cos(2\pi \frac{fin}{fs} 0 + \theta) \{\cos(2\pi B_{peak} 0/L) - j \sin(2\pi B_{peak} 0/L)\} + \dots + \right. \\
&\quad \left. v(L-1) \cos(2\pi \frac{fin}{fs} (L-1) + \theta) \{\cos(2\pi B_{peak} (L-1)/L) - j \sin(2\pi B_{peak} (L-1)/L)\}\right] \quad (2.47)
\end{aligned}$$

$$\begin{aligned}
&= E[v(0)] E[\cos(2\pi \frac{fin}{fs} 0 + \theta) \{\cos(2\pi B_{peak} 0/L) - j \sin(2\pi B_{peak} 0/L)\}] + \dots + \\
&\quad E[v(L-1)] E[\cos(2\pi \frac{fin}{fs} (L-1) + \theta) \{\cos(2\pi B_{peak} (L-1)/L) - j \sin(2\pi B_{peak} (L-1)/L)\}] \quad (2.48)
\end{aligned}$$

We observed in section 2.3.2 that the envelopes of audio signals have a bell shaped probability density function. Therefore, assuming ergodicity, each envelope data point in the window has the same probability density function, and the expected value is

$$E[v(i)] = \sum_{k=1}^N v_k P(v_k) = \frac{A}{2}, \quad (2.49)$$

where A is the maximum amplitude of audio signal, k is the number of trials at i th envelope position, and $i = 0, \dots, L-1$. Therefore, Eq (2.48) may be simplified as,

$$\begin{aligned}
& E[X(B_{peak})] \\
&= \frac{A}{2} E[\cos(2\pi \frac{fin}{fs} 0 + \theta) \{\cos(2\pi B_{peak} 0/L) - j \sin(2\pi B_{peak} 0/L)\}] + \dots + \\
&\quad \frac{A}{2} E[\cos(2\pi \frac{fin}{fs} (L-1) + \theta) \{\cos(2\pi B_{peak} (L-1)/L) - j \sin(2\pi B_{peak} (L-1)/L)\}] \quad (2.50)
\end{aligned}$$

$$= \frac{A}{2} E \left[\sum_{i=0}^{L-1} \cos(2\pi \frac{fin}{fs} i + \theta) \{ \cos(2\pi B_{peak} i / L) - j \sin(2\pi B_{peak} i / L) \} \right],$$

$$i = 0, \dots, L - 1. \quad (2.51)$$

The phase angle $\hat{\theta}$ of Eq (2.46) is therefore

$$\hat{\theta} = \tan^{-1} \left\{ \frac{-E \left[\sum_{i=0}^{L-1} (A/2) \cos(2\pi \frac{fin}{fs} i + \theta) \sin(2\pi B_{peak} i / L) \right]}{E \left[\sum_{i=0}^{L-1} (A/2) \cos(2\pi \frac{fin}{fs} i + \theta) \cos(2\pi B_{peak} i / L) \right]} \right\} \quad (2.52)$$

$$= \tan^{-1} \left\{ \frac{-(A/2) E \left[\sum_{i=0}^{L-1} \cos(2\pi \frac{fin}{fs} i + \theta) \sin(2\pi B_{peak} i / L) \right]}{(A/2) E \left[\sum_{i=0}^{L-1} \cos(2\pi \frac{fin}{fs} i + \theta) \cos(2\pi B_{peak} i / L) \right]} \right\} \quad (2.53)$$

$$= \tan^{-1} \left\{ \frac{-E \left[\sum_{i=0}^{L-1} \cos(2\pi \frac{fin}{fs} i + \theta) \sin(2\pi B_{peak} i / L) \right]}{E \left[\sum_{i=0}^{L-1} \cos(2\pi \frac{fin}{fs} i + \theta) \cos(2\pi B_{peak} i / L) \right]} \right\},$$

$$i = 0, \dots, L - 1. \quad (2.54)$$

The sinusoidal frequency, phase, and the peak bin B_{peak} are constants, thus

$$\hat{\theta} = \tan^{-1} \left\{ \frac{-\sum_{i=0}^{L-1} \cos(2\pi \frac{fin}{fs} i + \theta) \sin(2\pi B_{peak} i / L)}{\sum_{i=0}^{L-1} \cos(2\pi \frac{fin}{fs} i + \theta) \cos(2\pi B_{peak} i / L)} \right\} \quad i = 0, \dots, L - 1. \quad (2.55)$$

Obtaining the sum and difference frequency terms of the sinusoids in Eq (2.55), letting

$B_{peak} = B$, and then approximating the summation as an integral and integrating, as

described in [Sche 93] for constant envelope case, yields

$$\hat{\theta} = \tan^{-1} \left\{ \frac{\frac{\cos[(L-1)2\pi\alpha + \theta] - \cos\theta}{2\pi\alpha} - \frac{\cos[(L-1)2\pi\beta + \theta] - \cos\theta}{2\pi\beta}}{\frac{\sin[(L-1)2\pi\alpha + \theta] - \sin\theta}{2\pi\alpha} + \frac{\sin[(L-1)2\pi\beta + \theta] - \sin\theta}{2\pi\beta}} \right\} \quad (2.56)$$

$$= \tan^{-1} \left\{ \frac{\beta \cos[(L-1)2\pi\alpha + \theta] - \beta \cos\theta - \alpha \cos[(L-1)2\pi\beta + \theta] + \alpha \cos\theta}{\beta \sin[(L-1)2\pi\alpha + \theta] - \beta \sin\theta + \alpha \sin[(L-1)2\pi\beta + \theta] - \alpha \sin\theta} \right\}, \quad (2.57)$$

where

$$\alpha = \frac{fin}{fs} + \frac{B}{L} = \frac{fin}{fs} + \frac{\hat{fin}}{fs} \quad (2.58)$$

and

$$\beta = \frac{fin}{fs} - \frac{B}{L} = \frac{fin}{fs} - \frac{\hat{fin}}{fs} \quad (2.59)$$

Noting that $\alpha = \beta + \frac{2B}{L}$, and assuming the window length $L \gg 1$, we may rewrite

Eq (2.57) as

$$\hat{\theta} \approx \tan^{-1} \left\{ \frac{\beta \cos[(L-1)2\pi(\beta + 2B/L) + \theta] - (\beta + 2B/L) \cos[(L-1)2\pi\beta + \theta] + (2B/L) \cos\theta}{\beta \sin[(L-1)2\pi(\beta + 2B/L) + \theta] + (\beta + 2B/L) \sin[(L-1)2\pi\beta + \theta] - 2(\beta + B/L) \sin\theta} \right\} \quad (2.60)$$

\approx

$$\tan^{-1} \left\{ \frac{\beta \cos[2\pi(\beta L + 2B) + \theta] - (\beta + 2B/L) \cos[2\pi L\beta + \theta] + (2B/L) \cos\theta}{\beta \sin[2\pi(\beta L + 2B) + \theta] + (\beta + 2B/L) \sin[2\pi L\beta + \theta] - 2(\beta + B/L) \sin\theta} \right\} \quad (2.61)$$

$$= \tan^{-1} \left\{ \frac{\beta(\cos[2\pi\beta L + \theta] - \cos[2\pi L\beta + \theta]) + 2B/L(\cos\theta - \cos[2\pi L\beta + \theta])}{\beta(\sin[2\pi\beta L + \theta] + \sin[2\pi L\beta + \theta] - 2\sin\theta) + 2B/L\{\sin[2\pi L\beta + \theta] - \sin\theta\}} \right\} \quad (2.62)$$

$$= \tan^{-1} \left\{ \frac{2B/L(\cos\theta - \cos[2\pi L\beta + \theta])}{2\beta(\sin[2\pi\beta L + \theta] - \sin\theta) + 2B/L(\sin[2\pi L\beta + \theta] - \sin\theta)} \right\} \quad (2.63)$$

$$= \tan^{-1} \left\{ \frac{(\cos\theta - \cos[2\pi L\beta + \theta])}{(\beta L/B + 1)(\sin[2\pi L\beta + \theta] - \sin\theta)} \right\} \quad (2.64)$$

$$= \tan^{-1} \left\{ \frac{\sin[\pi L\beta + \theta]\sin\pi L\beta}{(\beta L/B + 1)\cos[\pi L\beta + \theta]\sin\pi L\beta} \right\} \quad (2.65)$$

$$= \tan^{-1} \left\{ \frac{\sin[\pi L\beta + \theta]}{(\beta L/B + 1)\cos[\pi L\beta + \theta]} \right\}. \quad (2.66)$$

Note that $\beta L/B \ll 1$ because the value of β is usually less than $1/2L$ and the peak bin B is usually greater than 1. So Eq (2.66) becomes

$$\hat{\theta} \approx \pi L\beta + \theta. \quad (2.67)$$

Note in Eq (2.67) that the expected value of $\hat{\theta}$ asymptotically converges to θ as the length of FFT increases. Therefore, $\hat{\theta}$ is an unbiased estimator of phase.

Substituting Eq (2.67) into Eq (2.38) yields

$$\begin{aligned} E[e(i)] & \approx 2E[v(i)]E[\sin\{.5(\hat{w}_d - w_d)i + .5N\pi L\beta\} \sin\{.5(\hat{w}_d + w_d)i + .5N(\pi L\beta + 2\theta)\}], \\ i &= 0, \dots, L-1. \end{aligned} \quad (2.68)$$

The low frequency sinusoid in Eq (2.68) makes $E[e(i)]$ equals to zero when

$$[.5(\hat{w}_d - w_d)i + .5N\pi L\beta]_{MOD2\pi} = 0, i = 0, \dots, L - 1, \quad (2.69)$$

where MOD 2π indicates Modulo 2π arithmetic. Further simplifying Eq (2.69) yields,

$$[N2\pi\Delta i / fs + N\pi L\beta]_{MOD2\pi} = 0 \quad (2.70)$$

$$= [2\Delta i / fs + L\beta]_{MOD2\pi} \quad (2.71)$$

$$= [2\Delta i / fs - L\Delta / fs]_{MOD2\pi} \quad (2.72)$$

$$= [2i - L]_{MOD2\pi} \quad (2.73)$$

When $i = L/2$, Eq (2.73) equals to zero, hence Eq (2.38) approaches zero in the mean sense, which was also observed in the previous section via computer simulations.

The mathematical derivation for non- rectangular window case is complicated due to the fact that the complexity of convolution operator does not lend a closed form solution for Eq (2.52). However, we may say the result of mathematical derivation holds stronger for the Hamming and Hanning window cases, based on empirical analysis in section 2.3.4.4. Discussions on analysis of the output error via computer simulation when parameters such as window type, window size, and the zero padding are adjusted for single and multiple frequency case in the following sections support above theoretical result.

2.3.5 An Analysis of the Single Frequency Case

In this section, a general description of the simulation, an introduction of the window selection criterion (WSC), and computer simulation results will be presented. Through the analysis of the simulation results, we will verify that the theoretical prediction, which was based on theoretical analysis of window function in section 2.3.4.3, agrees with the results of computer simulation. In section 2.3.5.1, the simulation results from a constant envelope, single sinusoid model are presented. Section 2.3.5.2 contains a discussion of the simulation results from the time varying envelope, single sinusoid model, which was described in section 2.3.2. A discussion of the WSC based on the results of the time varying envelope, single sinusoid model is then presented.

Following is a general description of the simulations.

- The window lengths L are 32, 64, 128, 256, 512, and 1024. Window lengths less than 1024 points were selected, because of the time varying nature of audio signals. Brown suggested that the time resolution of less than 25 msec is desirable for analyzing computer music [Brow 93]. 25 msec corresponds to 1102 points at 44.1 KHz sampling rate, thus 1024 point window is the maximum window length for this simulation.

- L length rectangular, Hamming, Hanning, and zero padded Hanning window with $L/2$ data points and $L/2$ zeros are used.

- Normalized frequencies of .05 to .225 are used with an increment of .025.

There are a total of 8 normalized frequencies used.

- The number of output points in the middle of the output window K is varied from 2, 4, ..., $L/2$.

The Normalized Squared Error (NSE) was chosen as the performance figure of merit for this estimator, despite the possibility that other error functions may be better suited for audio signals. NSE was chosen to both facilitate the analysis and because no known objective error function precisely describes the relationship between low and high errors, and what sounds good. A small NSE will have a strong correlation with good perceptive quality of voice or music, however a high NSE does not necessarily mean poor sound quality. The NSE is defined as,

$$NSE = \frac{\sum_{i=0}^M e^2(i)}{\sum_{i=0}^M d^2(i)}, \quad (2.74)$$

where M is the number of data sample points, and $e(i)$ is the error signal defined in Eq (2.19). Note that a large value of NSE could result in Eq (2.74) when the desired signal is near zero, i.e., a silence portion in between speech or music frames. This problem can be alleviated by excluding the silence portions from error calculations.

A mean NSE, denoted as the MNSE, is calculated for each combination of L , K , window type, and normalized frequency. The desired MNSE is set below .01, at every frequency, for the constant envelope, single frequency case, and .02 for the time varying envelope, single frequency case. These numbers (.01 and .02) are based on experience, and represent 1% and 2 % error power with respect to the signal power. For example, in the constant envelope, time varying case, let the window length $L = 32$ and $K = 2$, and we observe the computer generated MNSE at each normalized frequency. If the MNSE is below .01 at all frequencies then we can take 2 points from the output as a part of the

generated signal with higher harmonic frequency. If the MNSE is over .01 in any of the normalized frequencies, then the estimate is not assumed to be suitable for output because the error is higher than the threshold value. We then observe the MNSE at $K = 4$ for all frequencies, and so on.

2.3.5.1 Result 1: Constant Envelope, Single Sinusoid The constant envelope single sinusoid with random phase is used as an input to the ETHG. For this experiment, the desired output is the second harmonic frequency of this sinusoid. As discussed before, the estimation error is minimum in the middle of the output window. Therefore, K output points are extracted from the middle of each output window. Table 2.2 shows a comparison of obtainable K when the MNSE is desired below .01 at all eight normalized frequencies, which are varying from .05 to .225, for this constant envelope single sinusoid case. The MNSE is the mean NSE of 2000 trials at each frequency.

OBTAINABLE K WHEN MEAN NSE IS BELOW 0.01 AT ALL NORMALIZED FREQUENCIES (2000 TRIALS AT EACH FREQUENCY)

<i>FFT LENGTH(L)</i>	<i>RECTANGULAR</i>	<i>HAMMING</i>	<i>HANNING</i>	<i>HANNING (L/2 ZEROPAD)</i>
32	0	0	0	0
64	0	8	8	4
128	8	16	16	16
256	32	32	32	32
512	64	64	64	64
1024	128	128	128	128

Table 2.2. Obtainable K for constant envelope, single sinusoid.

The MNSE of rectangular, Hamming, Hanning, and Hanning window with $L/2$ zero padding are compared as a performance measure. Note here that the FFT length is L . In Table 2.2, no window of length 32 yielded an acceptable MNSE. At $L = 64$, the Hamming and Hanning windows allow a larger K than a 32 point Hanning window with 32 point zero pad. The rectangular window yields unacceptable results.

At $L = 128, 256, 512$, and 1024 all window types yield same number of K , except at $L = 128$ where the rectangular window shows less K . It was observed that the rectangular window yields higher MNSE. Fig. 2.24 shows a comparison of the average MNSE, which is the average of eight MNSEs at eight normalized frequencies and 2000 random phases, of each window type with $L = 256, 512$, and 1024 . The rectangular window appears to yield the highest average MNSE.

Generally, the number of K is proportional to $L/8$, which means $K = 128$ when $L = 1024$, $K = 64$ when $L = 512$, and so on.

Fig. 2.25 and Fig. 2.26 show the average MNSE with $K = 4$ and $K = 8$, respectively. The horizontal axis represents the window length L . Fig. 2.27 shows the case with $K = 16$ and L from 128 to 1024. Likewise Fig. 2.28 is for $K = 32$ and L of 256 through 1024, and Fig. 2.29 is for $K = 64$ with L of 512 and 1024. Fig. 2.30 shows the average MNSE for $K = 128$ when $L = 1024$. Here the Hamming and Hanning windows show the lowest MNSE, while the rectangular window shows a higher MNSE.

We note that as the window length increases, the average MNSE decreases, as shown in Fig. 2.25 through Fig. 2.29.

In general, the performance of the windows described above do not show a significant difference for the constant envelope single sinusoid case.

Comparison of Average MNSE (dB)

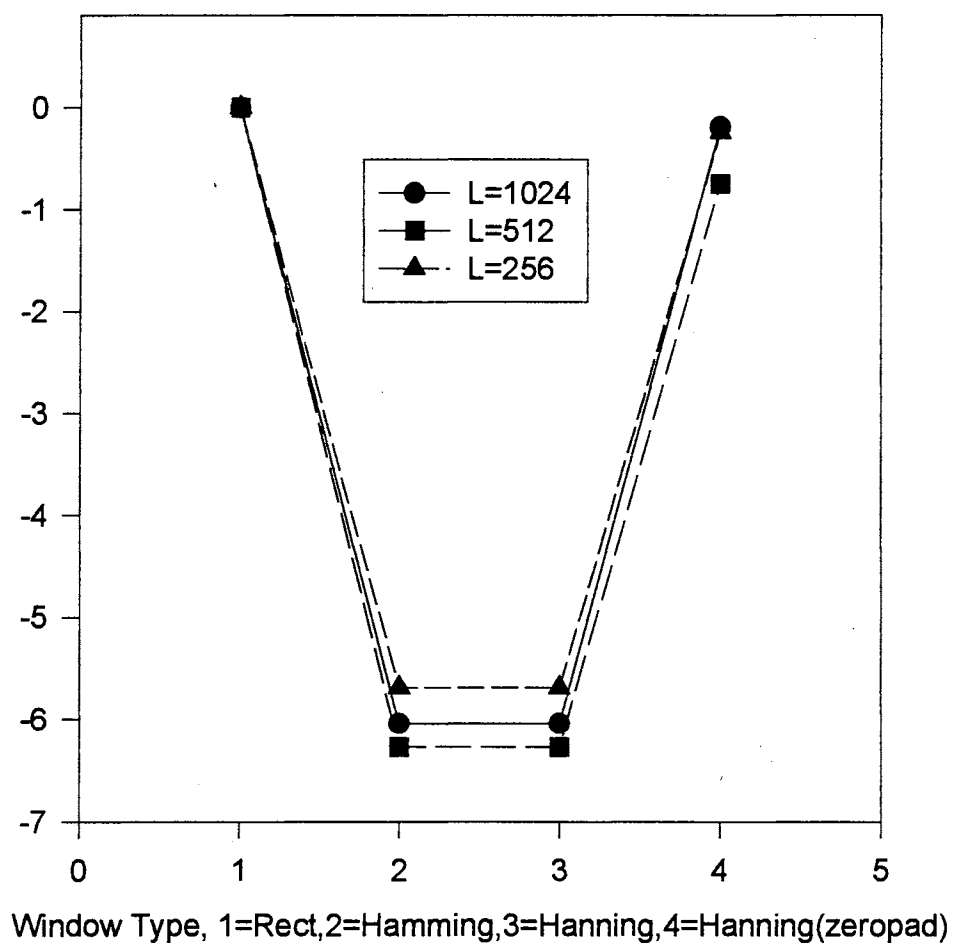


Fig. 2.24 Comparison of Average MNSE.
Normalized to highest value.

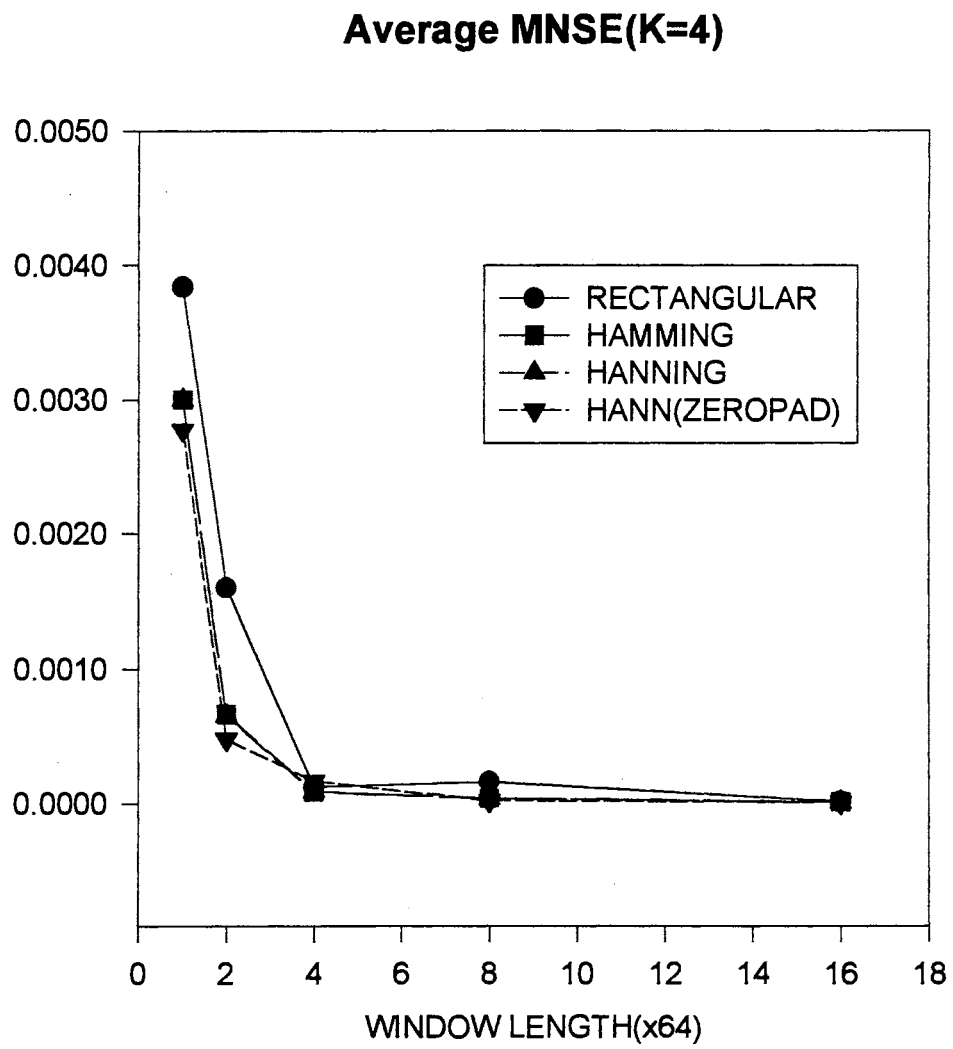


Fig. 2.25 Average MNSE for $L = 64, 128, 256, 512, 1024$, $K = 4$

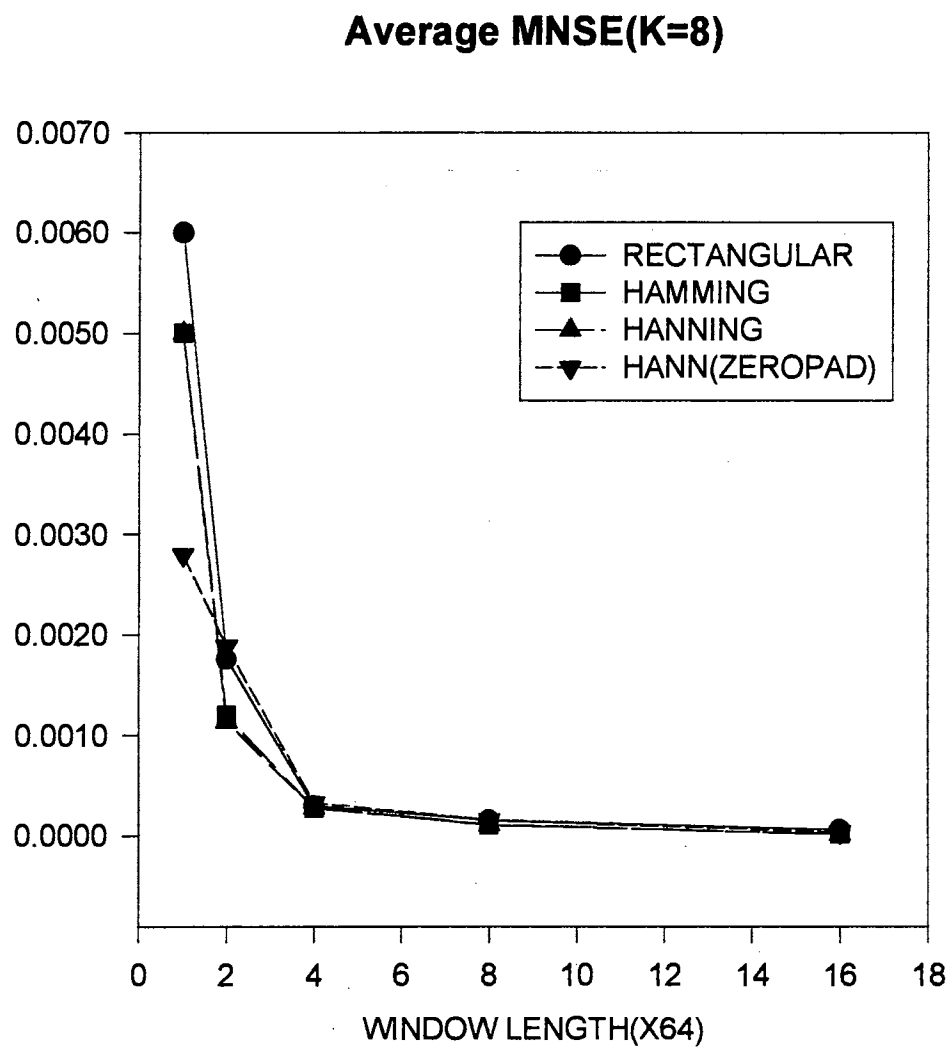


Fig. 2.26 Average MNSE for $L = 64, 128, 256, 512, 1024$, $K = 8$

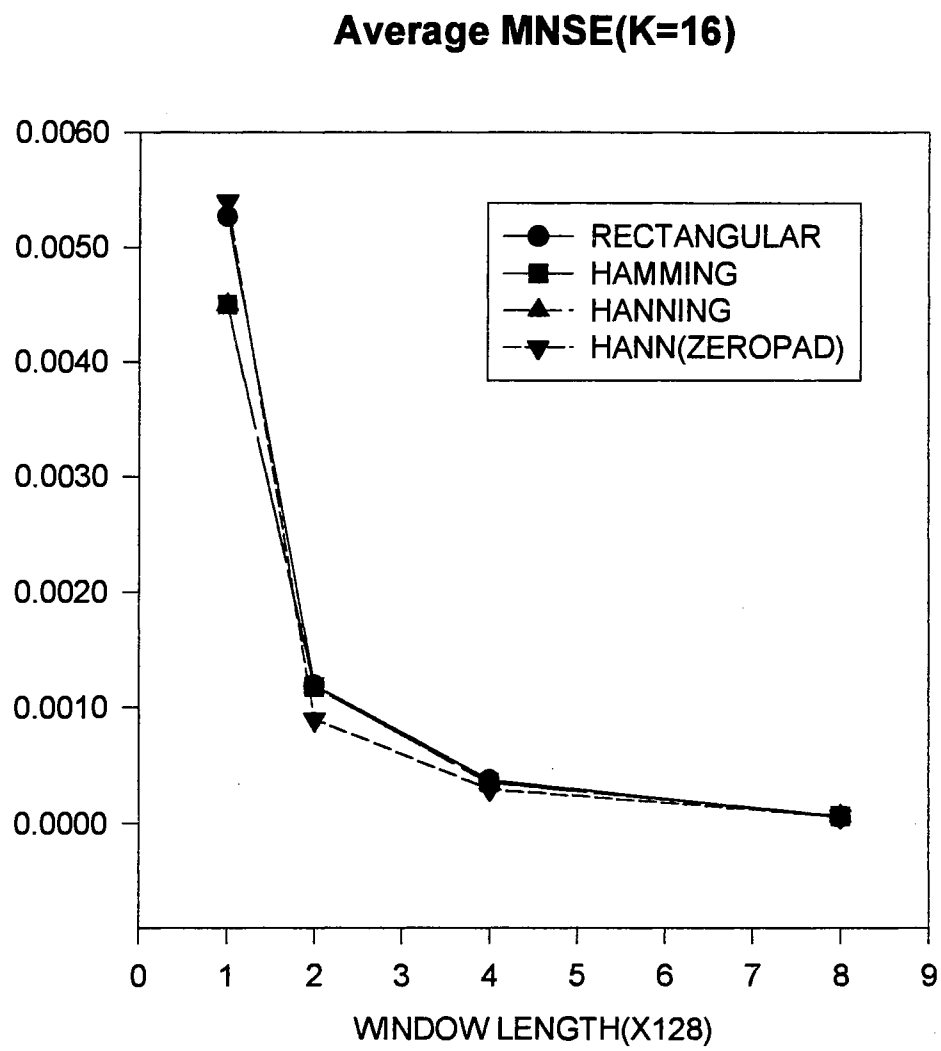


Fig. 2.27 Average MNSE for $L = 128, 256, 512, 1024$, $K = 16$

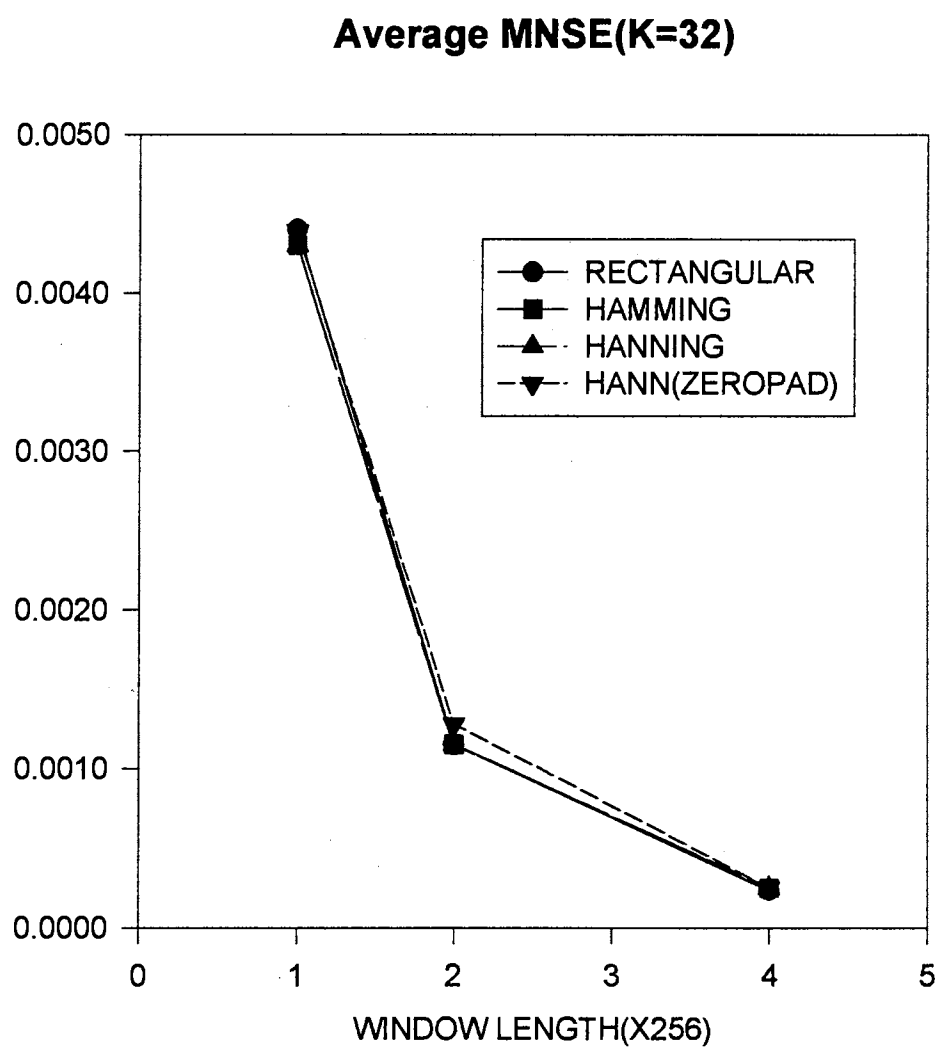


Fig. 2.28 Average MNSE for $L = 256, 512, 1024$, $K = 32$

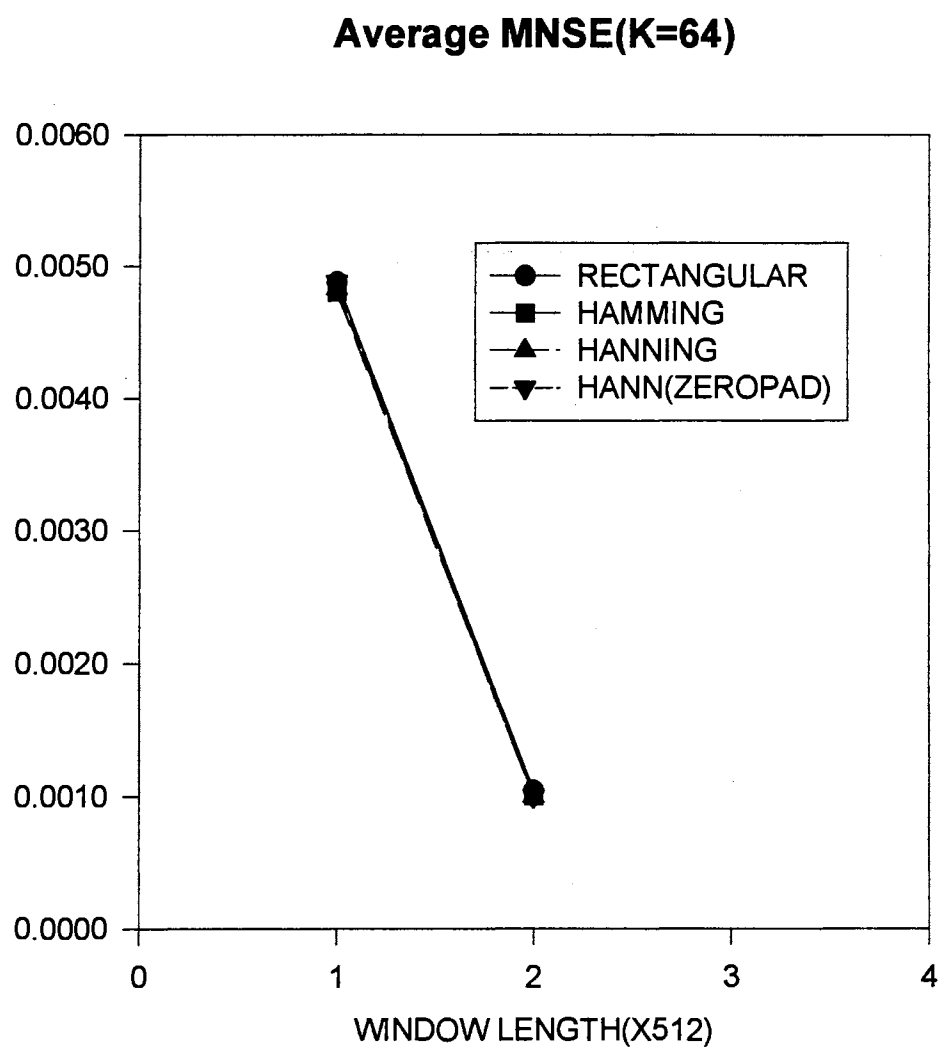


Fig. 2.29 Average MNSE for L = 512, 1024, K = 64

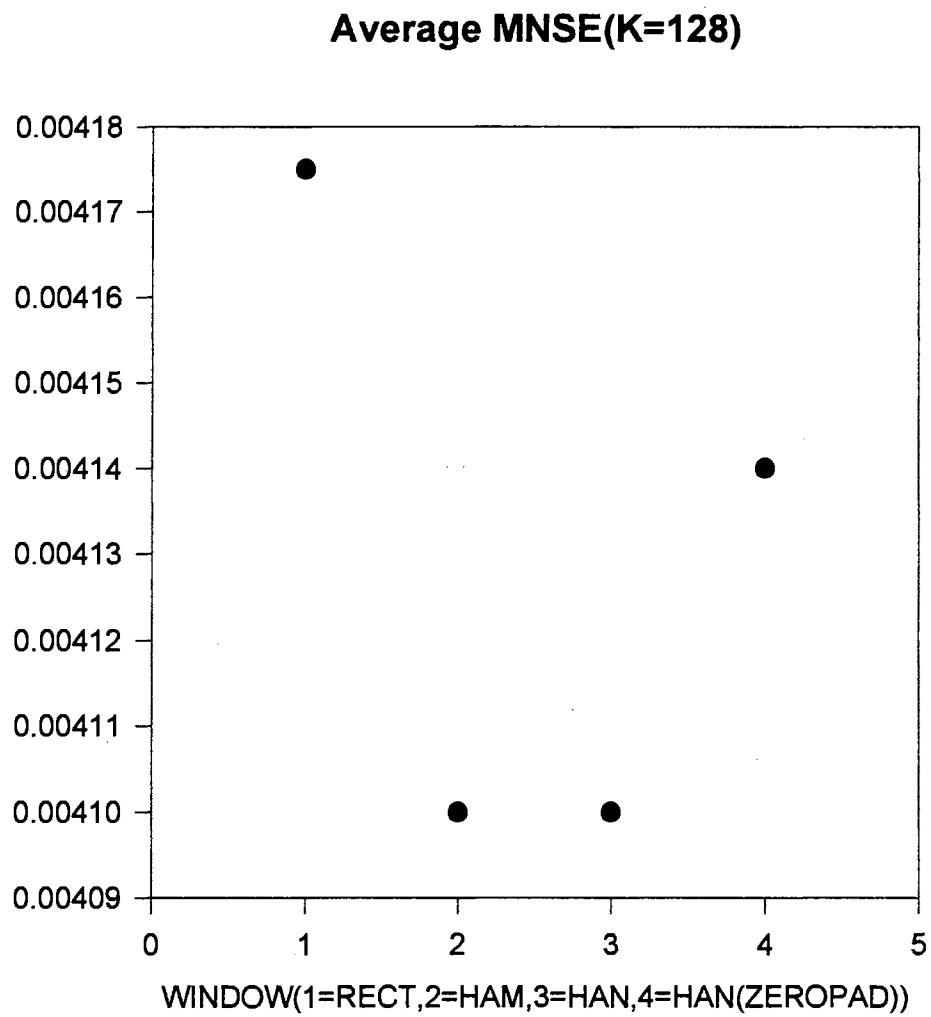


Fig. 2.30 Average of MNSE for L =1024, K=128.

The constant amplitude sinusoid is not a realistic model for real audio, because the envelope of an audio signal is generally dynamically varying. But a general assumption is that the amplitude is slowly varying within short window interval [Smit 87, Brow 93, Port 81]. It was noticed in section 2.3.4 that when there is a significantly sharp transition point in the window, a significant amount of error may be found in the middle section, depending on the position of the transition. Nevertheless, we observed in section 2.3.4, by theoretical derivation and computer simulation, that the average errors in the middle section of the processed output window were minimal.

In the next section, a discussion of computer simulation results from time varying envelope sinusoids will be presented. It will be shown that the average MNSE is decreased as the window length increased, and the zero padded Hanning window yields the lowest average MNSE, for the input signal with a time varying envelope case.

2.3.5.2 Result 2: Time Varying Envelope, Single Sinusoid This section discusses the experimental results from a single input sinusoid with a time varying envelope model. Similar to the previous section, the normalized frequency is varied from .05 to .225. Fig. 2.5, in section 2.3.2, shows examples of input signals generated by the model in Fig. 2.2. Input signals were generated by the waveform generation scheme which was discussed in section 2.3.2. In order to make a more exact comparison, the initial seed was fixed, and the consecutive seeds of the random number generator were adjusted in a controlled manner to generate exactly the same envelope shapes for different window types. Again, the MNSE is a mean NSE of 2000 trials at each normalized frequency and K, however, the desired MNSE was set below .02.

It was observed that there are larger errors in the middle of the output window compared to the constant envelope, single frequency case. Only the zero padded Hanning window achieved the MNSEs below .02 at all normalized frequencies, under specific K values. The last column in Table 2.3 shows that the Hanning window with zero padding has obtainable K at every window length, except $L = 32$, while other windows yielded none. Table 2.4 shows the MNSE of the rectangular, Hamming, Hanning, and Hanning window with zero padding at each K and normalized frequency with 1024 FFT length. The Hanning window with zero padding shows the lowest MNSE among the window types. For example, note the highlighted rows at $K = 128$. Fig. 2.31 is a plot of the MNSE with 1024 length rectangular, Hamming, Hanning, and 512 length Hanning window zero padded with 512 zeros at 8 normalized frequencies. The MNSE is shown when $K = 128$ points are extracted from the middle of estimated signal at eight different normalized frequencies. It shows that the 512 length Hanning window zero padded with

OBTAINABLE K WHEN MEAN NSE IS BELOW 0.02 AT ALL NORMALIZED FREQUENCIES (2000 TRIALS AT EACH FREQUENCY)

<i>FFT LENGTH(L)</i>	<i>RECTANGULAR</i>	<i>HAMMING</i>	<i>HANNING</i>	<i>HANNING (L/2 ZEROPAD)</i>
32	0	0	0	0
64	0	0	0	4
128	0	0	0	8
256	0	0	0	16
512	0	0	0	32
1024	0	0	0	128

Table 2.3. Obtainable K for time varying envelope, single sinusoid.

Window Type	K	Normalized frequency							
		.05	.075	.1	.125	.15	.175	.2	.225
Rectangular	2	.025	.016	.0432	.0238	.0389	.033	.0241	.0816
	4	.0158	.015	.0461	.0193	.0428	.0268	.0151	.0551
	8	.0217	.0148	.0459	.0164	.0428	.0222	.0145	.0469
	16	.0313	.0216	.0482	.0118	.0469	.0202	.0206	.0548
	32	.0265	.0219	.0492	.009	.047	.0154	.0176	.0558
	64	.0186	.0156	.0459	.0049	.0441	.0127	.0137	.0482
	128	.0152	.0131	.0455	.0025	.043	.0115	.0118	.0472
	256	.0179	.0163	.0616	.0017	.061	.0154	.0156	.0627
	512	.0372	.0362	.1382	.0013	.1379	.0355	.0357	.1383
Hamming	2	.0117	.0108	.0199	.021	.0159	.0251	.0124	.0352
	4	.0106	.0093	.0206	.0167	.0171	.0189	.0092	.0274
	8	.0166	.0097	.022	.0142	.0193	.0157	.0094	.0248
	16	.0264	.0167	.0243	.0105	.0229	.0138	.0152	.0295
	32	.0215	.0172	.0252	.009	.0224	.0104	.0129	.0313
	64	.0134	.0107	.0212	.0049	.0194	.0076	.0086	.023
	128	.0101	.0082	.0213	.0025	.0206	.0066	.007	.0226
	256	.0125	.011	.0367	.0017	.0361	.0101	.0103	.0375
	512	.0313	.0303	.113	.0013	.1127	.0296	.0298	.113
Hanning	2	.0111	.0105	.0184	.0209	.0143	.0246	.0118	.0339
	4	.0104	.0089	.0189	.0166	.0152	.0184	.0088	.0266
	8	.0163	.0094	.0204	.0142	.0178	.0153	.0091	.0235
	16	.0261	.0165	.023	.0104	.0215	.0134	.015	.0279
	32	.0212	.017	.0239	.009	.0209	.0101	.0126	.0297
	64	.0131	.0104	.0198	.0049	.0179	.0073	.0083	.0214
	128	.0098	.0079	.0199	.0025	.0192	.0063	.0067	.0211
	256	.0122	.0107	.0352	.0017	.0347	.0098	.01	.036
	512	.0308	.0299	.1114	.0013	.1111	.0292	.0294	.1115
Hanning w/ zero pad	2	.0103	.0057	.0084	.0098	.0064	.009	.007	.0131
	4	.0109	.0065	.0089	.008	.0073	.0076	.0054	.0164
	8	.0153	.0094	.011	.0079	.0097	.0076	.007	.0152
	16	.021	.0155	.0161	.0089	.0138	.0099	.0117	.0182
	32	.0154	.012	.0138	.007	.0113	.007	.0088	.0167
	64	.0101	.0076	.0109	.0042	.0095	.0051	.0059	.0122
	128	.008	.0062	.0124	.0026	.0117	.0047	.0051	.0131
	256	.0106	.0094	.0293	.0017	.0288	.0084	.0086	.0298

Table 2.4 Comparison of MNSE (FFT length = 1024).

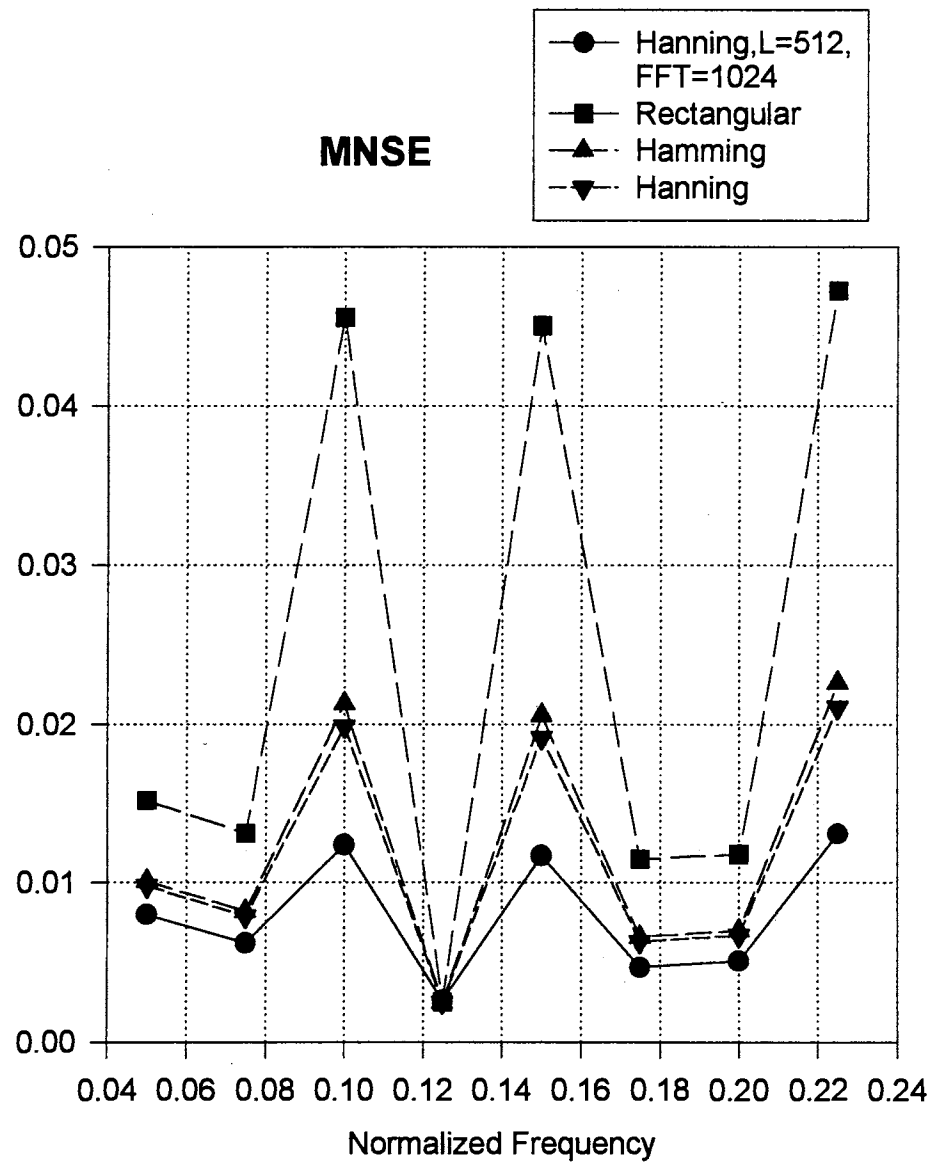


Fig. 2.31 MNSE for FFT length = 1024, $K = 128$.

512 zeros yields the smallest estimation error at every normalized frequencies.

Similar results were also observed at other window lengths. Fig. 2.32 shows the average of MNSEs at eight normalized frequencies when $K = 4$ points are taken with L of 64, ..., 1024. We see that the rectangular window produces the highest estimation error on average, and the zero padded Hanning window yields the lowest average estimation error. Also, in Fig. 2.33 through Fig. 2.36, we note that the zero padded Hanning window yields the lowest MNSE at every window length.

The average MNSE generally seems to decrease and more K can be taken as the window length increases, as shown in Fig. 2.32 through Fig. 2.35. The allowable K is proportional to $L/16$ for window lengths of 64, ..., 512, and $L/8$ for 1024 length.

However, increasing the window length also increases the computation load, which results in a longer time to generate the estimate. Therefore, we should find an optimal window length which satisfies the required minimum MNSE, which is .02, and optimizes the computation speed.

The window selection criterion (WSC) was developed to find an optimal window length in the sense of maximizing the computation speed under the desired MNSE. There are three factors to consider. First, the FFT computation load which is proportional to $L \log_2 L$ additions and multiplications. We will use $L \log_2 L$ for simplicity. The number of available K points also affects the speed of the whole process of generating an estimate of the signal with higher harmonics. The third factor is the difference of MNSE, however, it is not clear how to calibrate the MNSE difference into the WSC calculation, because it depends on psychoacoustic perception. Therefore, based on the K value, and

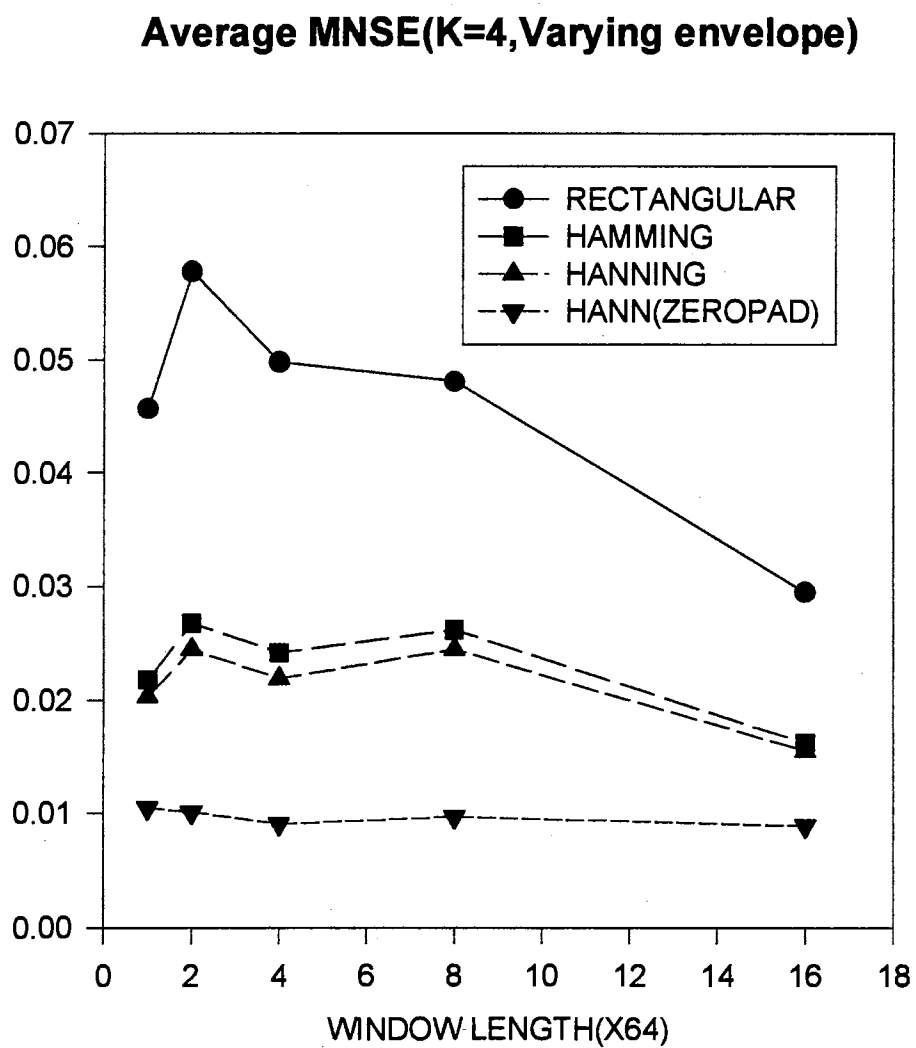


Fig. 2.32 Average MNSE for $L = 64, 128, 256, 512, 1024$, $K = 4$.
Time varying envelope case.

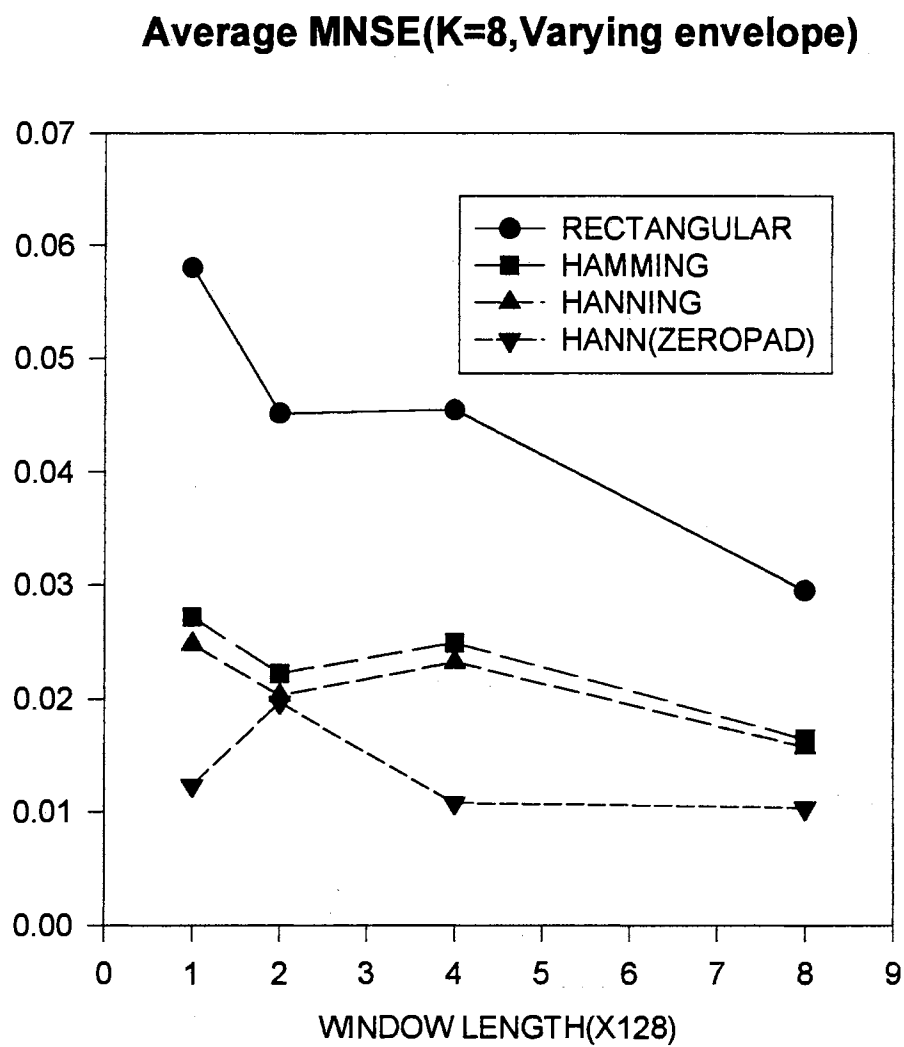


Fig. 2.33 Average MNSE for $L = 128, 256, 512, 1024$, $K = 8$.
Time varying envelope case

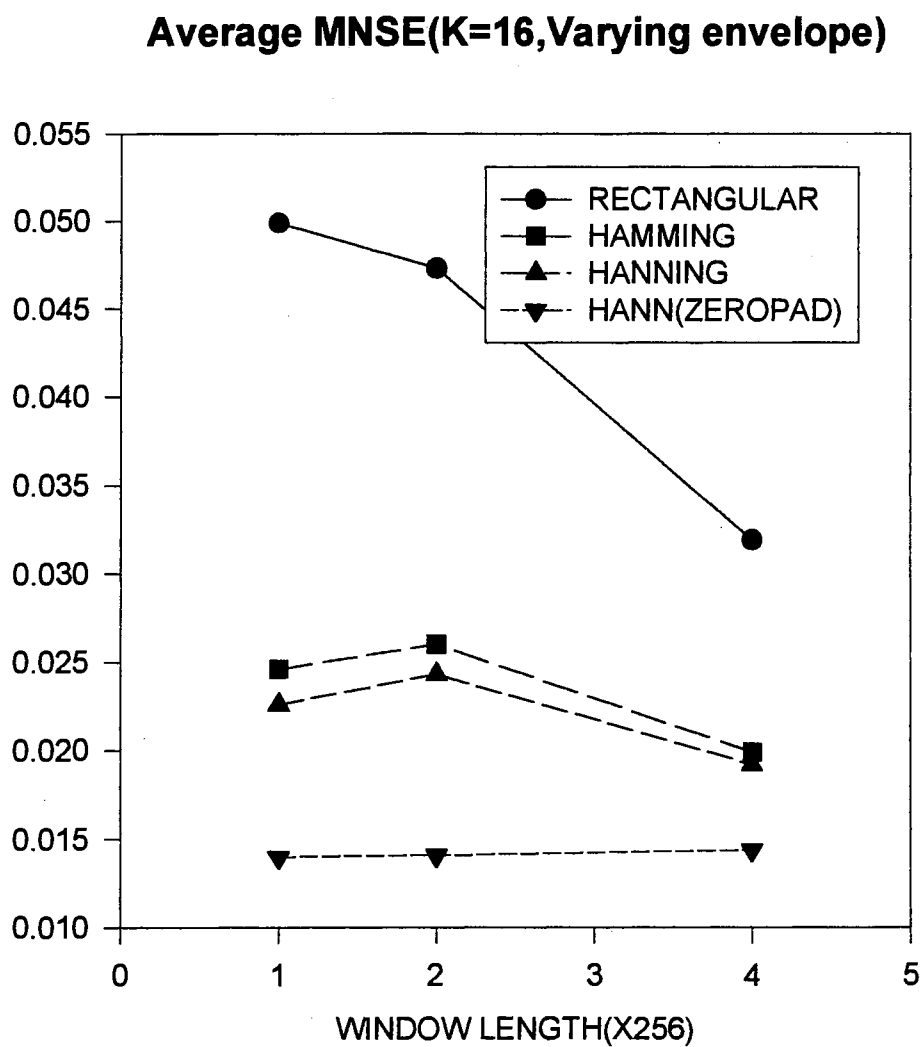


Fig. 2.34 Average MNSE for $L = 256, 512, 1024$, $K = 16$.
Time varying envelope case.

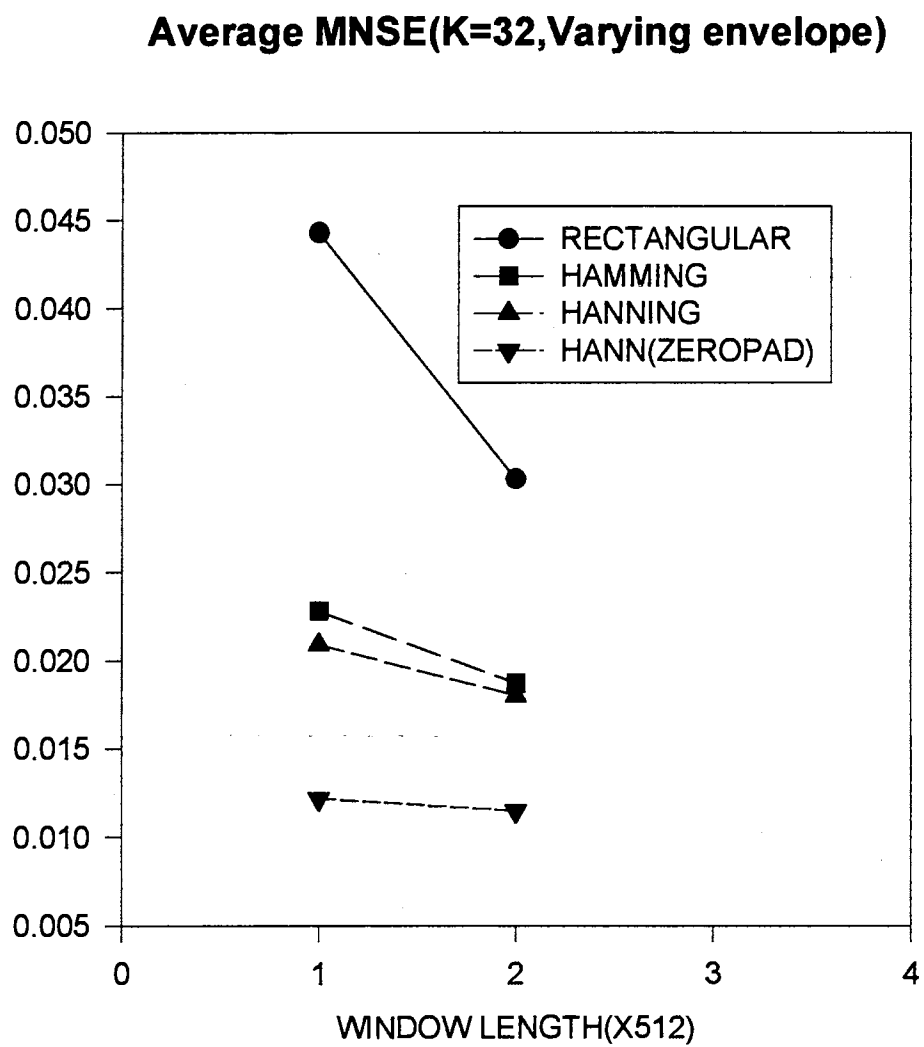


Fig. 2.35 Average MNSE for $L = 512, 1024, K = 32$.
Time varying envelope case.

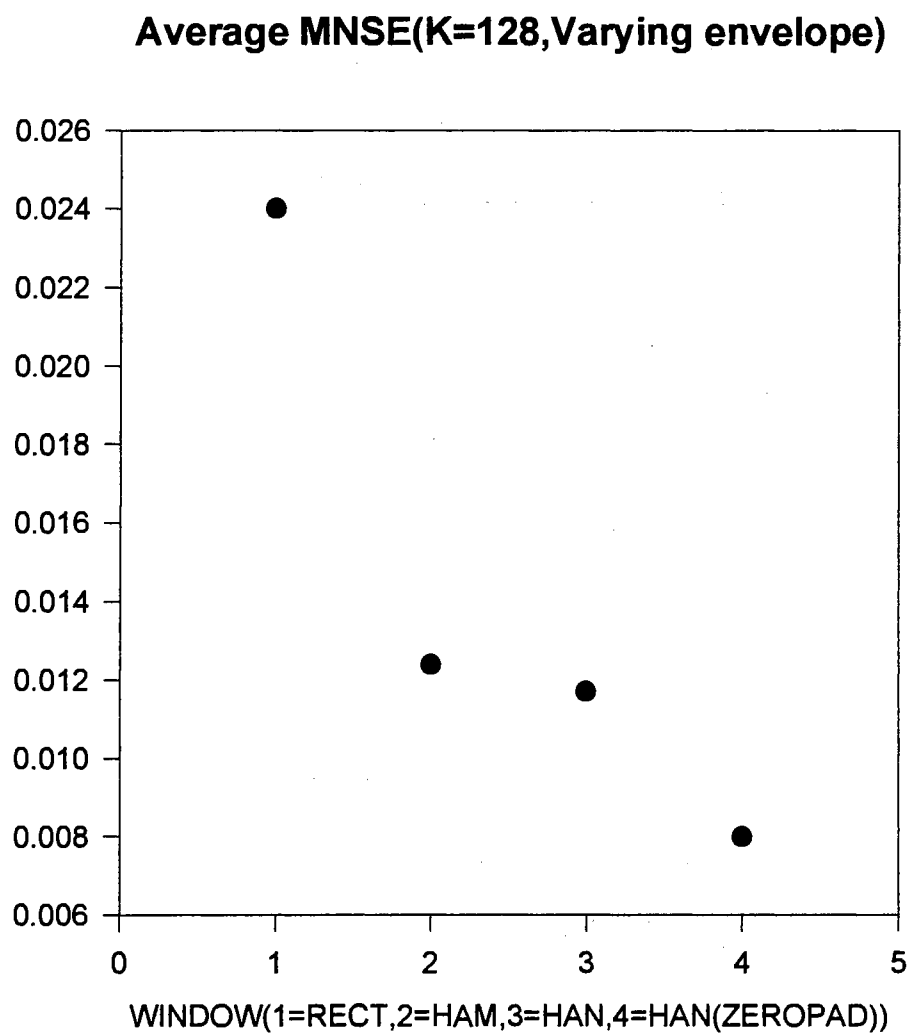


Fig. 2.36 Average MNSE for $L = 1024$, $K=128$.
Time varying envelope case.

the window length, the optimal window size and type where the FFT computation load and number of output points K produce an optimal computation speed under the desired MNSE will be determined.

The window selection criterion (WSC) is defined as

$$WSC = \frac{FFT_{load}}{K}, \quad (2.75)$$

where

FFT_{load} is the FFT computation load,

$$FFT \text{ computation load} = L \log_2 L \text{ [Oppe 89]}, \quad (2.76)$$

and K is the obtainable K at window length L . The smaller the value of the WSC the smaller the overall computational load will be.

Table 2.5 shows the WSC values at various FFT lengths using the Hanning window with zero padding technique, based on results found in Table 2.3. The 2048 length FFT (1024 zero padded) is added for comparison purpose. The minimum WSC value is achieved with $L = 1024$ with 512 zeros padded.

Note here that with 2048 FFT, $K = 256$ is obtained. The 2048 length FFT also yielded K of 128 with MNSEs less than .01 at all normalized frequencies. Table 2.6 shows the average of MNSEs at eight normalized frequencies, with K values from 2 to 128. It is obvious that the 2048 length FFT yielded less average MNSE.

Even though the minimum WSC value is located at 1024 FFT length, it should be used with caution because the difference of MNSE between other FFT length was not considered. In other words, the minimum WSC value does not guarantee the best perceptive sound quality, since it is only a computation speed measure based on MNSEs

<i>FFT LENGTH</i>	<i>K</i>	<i>FFT LOAD</i>	<i>WSC</i>
64	4	384	96
128	8	896	112
256	16	2048	128
512	32	4608	144
1024	128	10240	80
2048	256	22528	88

Table 2.5 Window Selection Criterion (WSC).

<i>K</i>	<i>AVERAGE MNSE</i>	
	<i>1024 FFT (512 data, 512 zeros)</i>	<i>2048 FFT (1024 data, 1024 zeros)</i>
2	.0087	.0084
4	.0089	.0072
8	.0104	.0087
16	.0144	.0138
32	.0115	.0132
64	.0082	.008
128	.008	.0057

Table 2.6 Comparison of average MNSE of 1024 FFT and 2048 FFT.

below the threshold value .02. How to allocate a weight for MNSE difference between window lengths on the WSC calculation is not clear at this moment, thus it was not used in the WSC value. The effects of error difference on the subjective quality should be analyzed by hearing test, etc., and it will be discussed in chapter 4.

We note here that the zero padding enhances the resolution of the frequency spectrum, but the window length should not be too small. For example, at $L = 32$ the zero padding was not observed to make any improvement. The window length should be minimum of 64 in this case. Note in the constant envelope case, we observed that zero padding does not make any performance improvement over Hamming and Hanning window. However, zero padding achieved more K values compared to the rectangular window as shown in Table 2.2. In the next section, some simulations with multiple input frequencies are done to see whether the zero padded Hanning window consistently yields the lowest MNSE.

2.3.6 The Multiple Frequency Case

In this section, the average error pattern will be investigated when the input signal has multiple frequencies, and will be followed by a discussion of differences of MNSEs between window types.

The average squared error pattern is investigated in the following discussion. For this experiment, two constant frequency sinusoids modulated by an envelope which was generated by the envelope generation scheme discussed in section 2.3.2 were generated. The frequencies of these sinusoids were independent and uniformly distributed between normalized frequencies of 0 and .23 Hertz. Fig. 2.37 shows a block diagram of the generation scheme for a typical window. It shows that sum of the two sinusoids are modulated by envelope function. It appears that the estimation error is, in general, also minimum in the middle of the window for this case. Fig. 2.38 shows an example plot of mean squared error between the desired signal and the generated output when the 512 length Hanning window is used. This plot is a mean of 2000 trials. The minimum mean squared error is shown to be in the middle of the processed window. This behavior was consistently observed with other window types and lengths. Note the minimum error is larger than the minimum error in Fig. 2.23, which is a plot for single frequency case. This increased error is not unexpected, because, unlike the single frequency case, two overlapping spectral magnitude should be separated in this case. For this simulation, spectral peaks were separated at the minimum spectral magnitude point in between them. In general, separating two spectral peaks by the minimum magnitude point or by the middle point did not cause significant differences.

Computer simulations were performed to examine the MNSE difference between the rectangular, Hamming, Hanning, and the zero padded Hanning windows, using the signal generated by the scheme in Fig. 2.37. It was also observed that the Hanning window with zero padding consistently yields more K points with less errors. Table 2.7 shows a comparison of obtainable K when the MNSE from 2000 trials is less than .02, for two sinusoid input case.

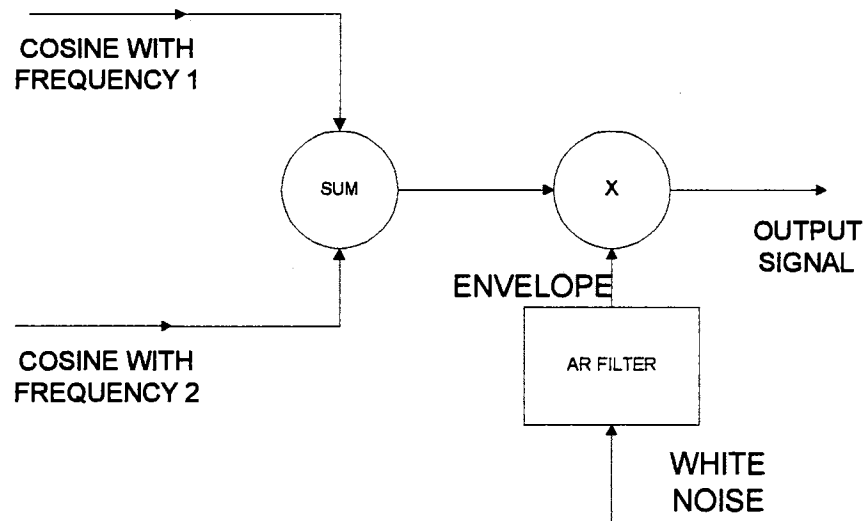


Fig. 2.37 Waveform with double frequencies generation for typical window.

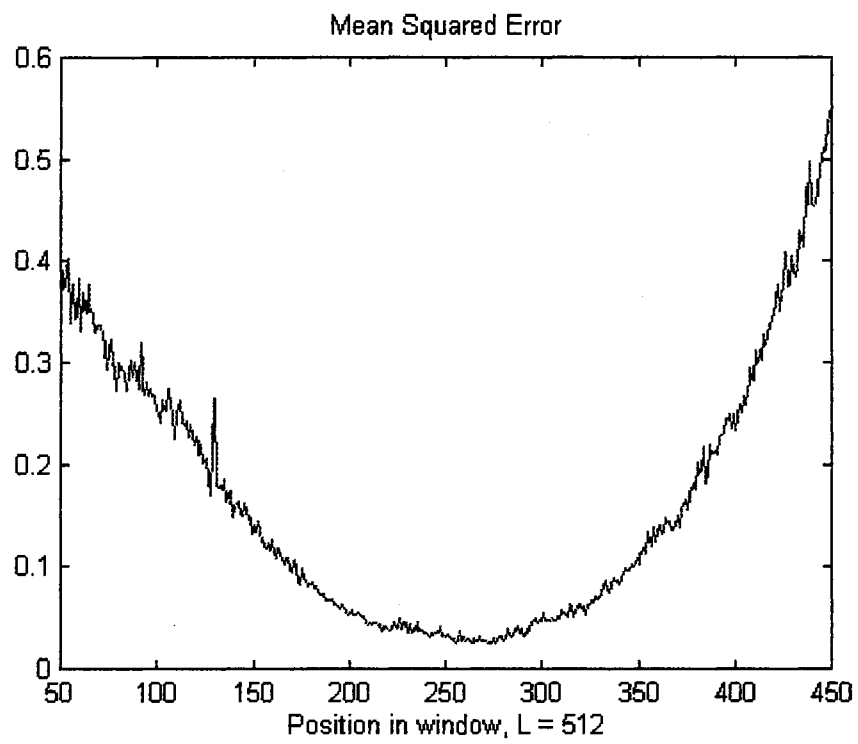


Fig. 2.38 Mean squared error, 512 length Hanning window.

OBTAINABLE K WHEN MEAN NSE IS BELOW 0.02
(2000 TRIALS)

<i>FFT LENGTH(L)</i>	<i>RECTANGULAR</i>	<i>HAMMING</i>	<i>HANNING</i>	<i>HANNING (L/2 ZEROPAD)</i>
32	0	0	0	0
64	0	0	0	0
128	0	0	0	0
256	0	0	0	0
512	0	0	0	0
1024	0	0	0	4

Table 2.7. Obtainable K for time varying envelope, double sinusoid.

Each entry in Table 2.7 represents a mean of 2000 trials. Note the K points are decreased to 4 at FFT length $L = 1024$ and no K points are shown for other windows. This result shows that the zero padded Hanning window consistently yields lower estimation error, even though the error amount is larger compared to the single frequency case. Another look at this two sinusoid case is presented in the following discussion.

Fig. 2.39 shows a comparison of MNSE when $K = 32$, for window length of 256, 512, and 1024, using rectangular, Hamming, Hanning, and zero padded Hanning window. The zero padding length is same as the window length for this plot, where as $L/2$ zero padding was used for plots in section 2.3.5. Again, each point in the plot represents a

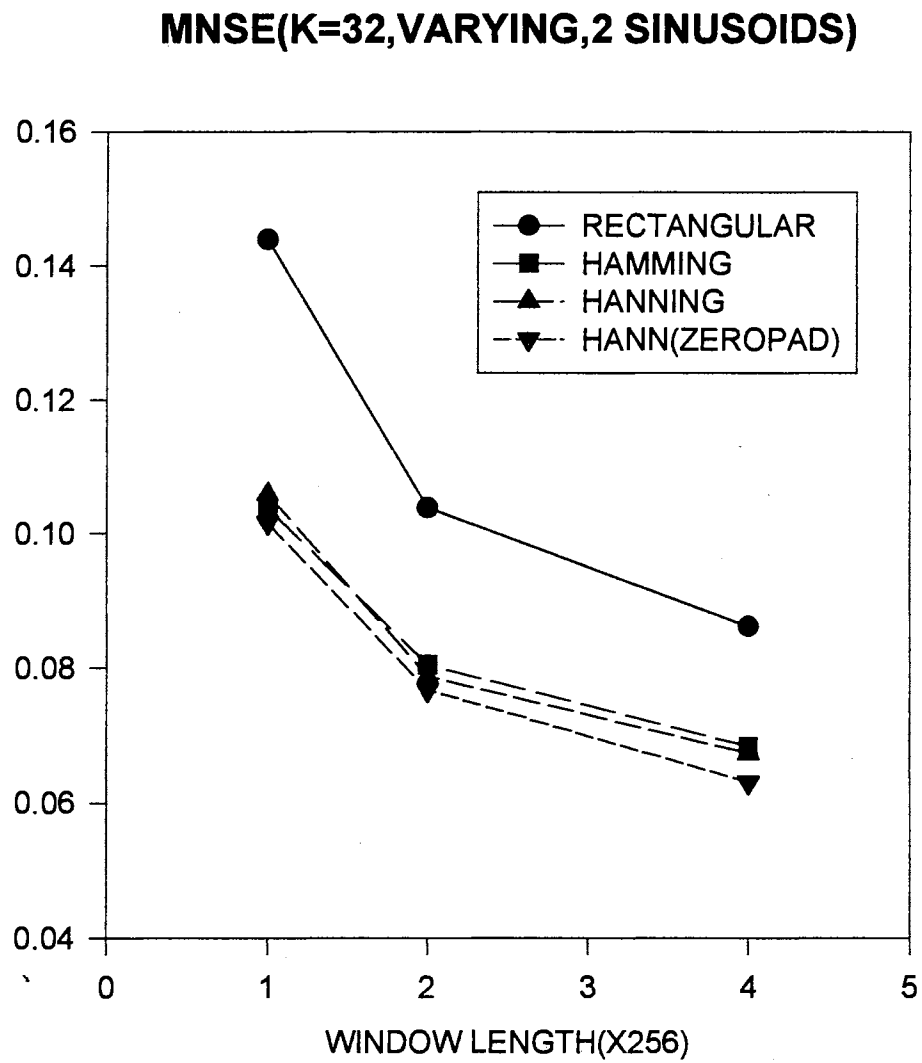


Fig. 2.39 Average MNSE for $L = 256, 512, 1024, K = 32$
Time varying envelope, 2 sinusoid case

mean of 2000 trials. Likewise Fig. 2.40, and Fig. 2.41 shows when $K = 64$, and 128, respectively. The MNSE is decreased as the window length goes up, and the zero padded Hanning window shows the lowest MNSE. In Fig. 2.41, the MNSE from 2048 length FFT using 1024 zero padded Hanning window shows the lowest MNSE. We also noticed in single frequency case (Table 2.6) that the 2048 length FFT using Hanning window with 1024 zeros yielded the lowest MNSE. The MNSE of 2048 length FFT is .0291, and non-zero-padded 1024 FFT using Hanning window is .0346. The MNSE from zero padded 2048 FFT is 16 % lower than that from 1024 FFT.

We observed in this section that the zero padded Hanning window consistently yields lower MNSE. Note this result shows conformity to the prediction made in the section 2.3.4, which predicted improved performance from the Hanning window. It was noticed that the zero padded 2048 length FFT yields 16 % lower MNSE than the non-zero-padded 1024 FFT. The effect of this difference on subjective listening quality will be tested in chapter 4.

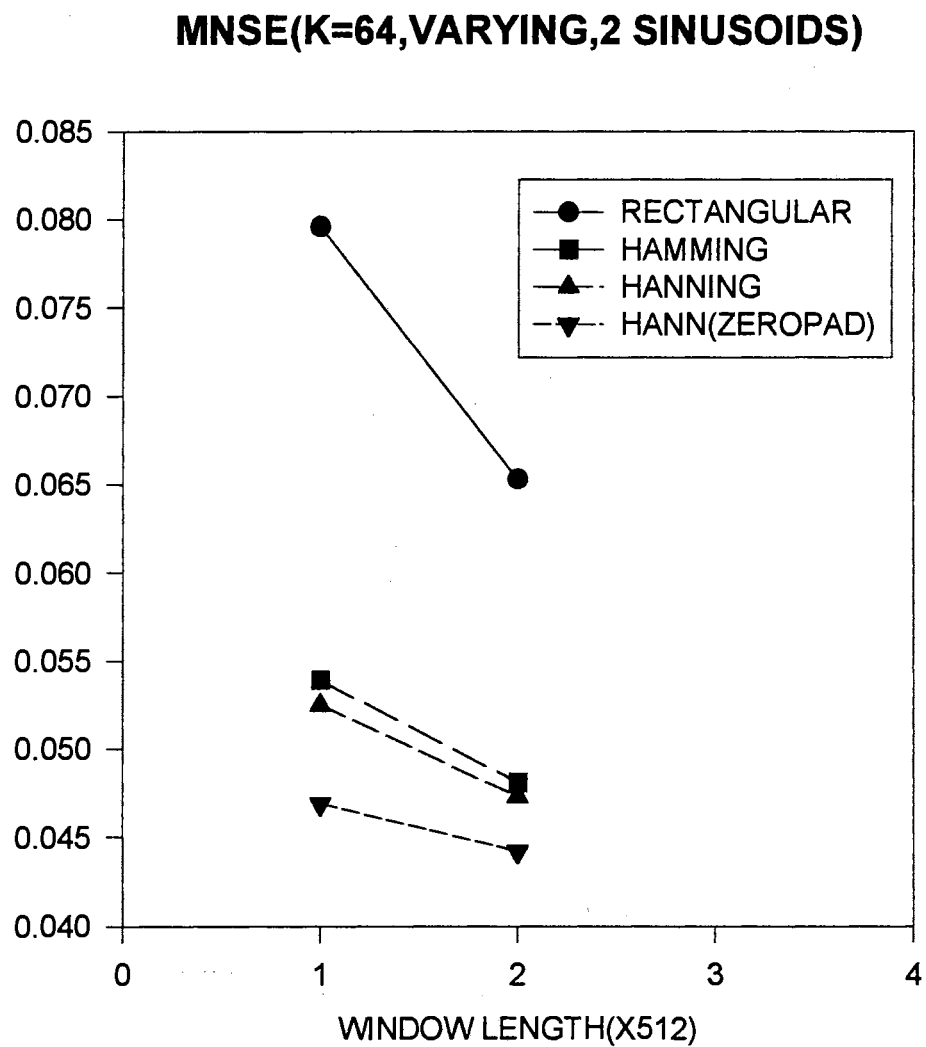


Fig. 2.40 Average MNSE for $L = 512, 1024, K = 64$
Time varying envelope, 2 sinusoid case

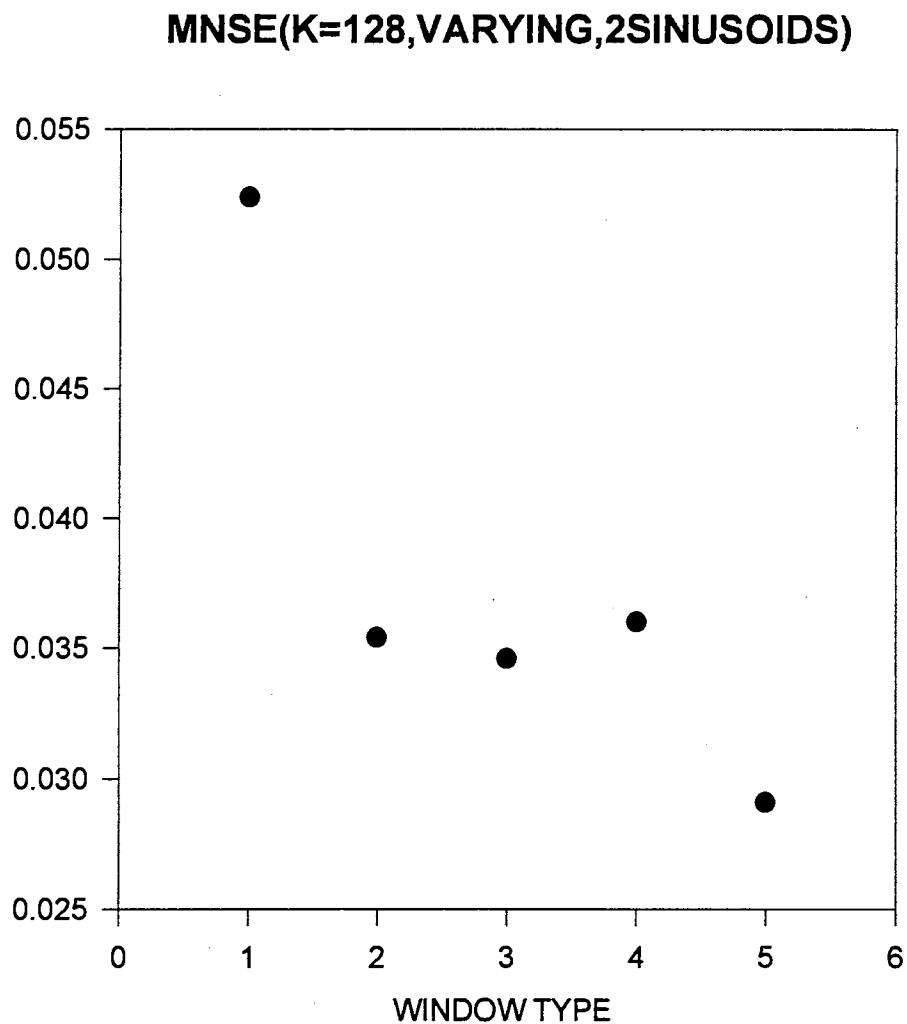


Fig. 2.41 Average MNSE for FFT length = 1024, K = 128.

Time varying envelope, 2 sinusoid case.

X-axis represents window type as,

- 1: Rectangular
- 2: Hamming
- 3: Hanning
- 4: Hanning (1024 FFT with 512 zero pad)
- 5: Hanning (2048 FFT with 1024 zero pad).

2.3.7 Summary

We have discussed that there is no algorithm designed specifically for enhancing the quality of severely bandlimited audio, and possible applications of the ETHG algorithm in section 2.2 and 2.3, respectively.

An envelope waveform model was introduced in section 2.3.2. The waveform model is used for simulation purposes, and matches the first and second order statistics of real audio envelopes.

In section 2.3.3, the general description and sources of error for ETHG algorithm were presented. An important convolution relationship between the impulse response of the frequency truncation function and the window function was discussed in section 2.3.4.3. The Hanning window produced the least amount of distortion among the compared windows. A prediction, based on theoretical analysis of the convolution result, was made that the Hanning window would yield the best performance. It was also shown, by computer simulations in section 2.3.4.4 and by theoretical derivation in section 2.3.4.5, that the mean squared error is minimum in the middle of the processed window.

Computer simulation results in section 2.3.5 and section 2.3.6 appear to indicate that the Hanning window with zero padding yields the least amount of MNSE, which agrees with the prediction made in the section 2.3.4.3.

The window selection criterion was discussed as a reference to find an optimal window size for optimal computation speed. Also, the need to do listening tests to find out the effect of MNSE difference on subjective listening quality was addressed.

CHAPTER III

DISTORTION MEASURES

3.1 Introduction

The signal after processed by the ETHG algorithm is different from the desired signal due to the errors discussed in section 2.3.4. The amount of difference can be evaluated either by objective distortion measures or by subjective quality tests. Subjective tests are, in general, time consuming, difficult to reproduce, and may be inconsistent. On the other hand, objective measure can be repeated with consistency, and reliability [Beer 92, Pali 92, Quack 88]. It is generally believed that an objective distortion measure should not be computationally intensive so that the actual distortions can be easily computed, and have some correlation with the subjective perception of quality [Gray 80, Quac 88]. We note that there is no universal objective quality measure that can explain the characteristics of *any* processed output signal. They are usually designed for a specific application.

This chapter serves as a literature survey of some objective distortion measures in use today. In the following section, several types of time domain and frequency domain distortion measures will be briefly discussed. The discussion includes their usefulness for ETHG algorithm with advantages and disadvantages.

3.2 Time Domain Distortion Measure

Time domain distortion measures typically evaluate the difference of certain set of compared signals. Usually, when there is a small error, it is highly correlated with a good subjectively perceived quality [Beer 92].

The main drawback of time domain distortion measures is the inability to explain psychoacoustic perception. For example, even though a slight phase misalignment between two signals may cause a high distortion measurement, the human auditory system may not perceive any difference. This is due to the fact that human ear is insensitive to phase distortion [Wang 82, Kubi 91]. We note that there are cases when the phase information is critical, such as noise, or echo suppression by an adaptive signal processing technique [Hayk 96, Widr 85].

Following is a discussion of some selected time domain distortion measures.

3.2.1 Signal to Noise Ratio

The signal to noise ratio (SNR) has been widely used in many signal processing areas due to its simplicity and tractability. The original unprocessed signal is needed to calculate the SNR, so it is useful for measuring the quality of audio enhancing or coding systems [Dell 93, Kubi 91, Quac 88]. SNR is defined as [Dell 93]

$$\text{SNR} = 10 \log_{10} \frac{\sum_n d^2(n)}{\sum_n [d(n) - \hat{d}(n)]^2}, \quad (3.1)$$

where $d(n)$ is the original signal, and $\hat{d}(n)$ is the processed output signal.

SNR tells us the ratio of signal power versus the error power. Therefore, it is meaningful when the degree of differences between the original signal and a processed signal is compared. There are cases, however, that small SNR does not necessarily mean poor subjective perception quality. For example, even though a processed signal with a phase distortion shows a smaller SNR, the human ear may not detect the difference. This is a major drawback of SNR when phase distortions exist in the processed signal.

Despite this problem, SNR is an important distortion measure because the ear is very sensitive to amplitude and frequency distortions [RMoo 90]. A higher SNR between the original and a processed signal indicates that the two signals have little differences, because the SNR is based on sample-by-sample differences. Therefore, a processed signal that yields high SNR may be said to be a good estimate of the original signal.

Many variations of SNR have been proposed so far, such as segmental SNR, and frequency weighted segmental SNR. Among them, the frequency weighted segmental SNR is discussed as follows.

3.2.1.1 Frequency Weighted Segmental SNR The frequency weighted segmental SNR ($\text{SNR}_{\text{fw-seg}}$) puts weight on frequency bands which is proportionally spaced to the ear's critical bands. The purpose of the frequency weight is to closely approximate psychoacoustic perception. Therefore, a knowledge of the frequency-dependent critical bands is needed. Among many different types of $\text{SNR}_{\text{fw-seg}}$, Tribollet's formula is introduced here. It is expressed as [Dell 93]

$$\text{SNR}_{\text{fw-seg}} = \frac{1}{M} \sum_{j=0}^{M-1} 10 \log_{10} \left[\frac{\sum_{k=1}^K w_{j,k} 10 \log_{10} [E_{s,k}(m_j) / E_{\varepsilon,k}(m_j)]}{\sum_{k=1}^K w_{j,k}} \right], \quad (3.2)$$

where M is the number of frames, K the number of frequency bands, w is the weight, $E_{s,k}(m_j)$ is the short term signal energy contained in the k th frequency band for original signal frame, and $E_{\varepsilon,k}(m_j)$ is the noise energy in the k th frequency band. $\text{SNR}_{\text{fw-seg}}$ measures the perceptive quality of a processed signal better than the standard SNR and SNR_{seg} [Dell 93, Quac 88]. However, due to the complexity of frequency-dependent weighting function, it was not implemented for the ETHG.

3.2.2 Mean Square Error

The mean square error (MSE) between the original signal $d(n)$ and the processed output signal $\hat{d}(n)$ is defined as

$$\text{MSE} = E[\{d(n) - \hat{d}(n)\}^2], \quad (3.3)$$

where E is an expected value operator.

MSE has been widely used in many areas of signal processing such as Wiener filtering and Widrow's adaptive Least Mean Square (LMS) algorithm [Hayk 96, Widr 85].

MSE was used in section 2.3.4 to show that the error was minimum in the middle of the processed output window. Since it does not show the ratio of error amount with respect to the desired signal power, which gives a clear indication of error percentage for both large and small signal power sections, its use was limited to section 2.3.4.

3.2.3 Normalized Squared Error

The normalized squared error (NSE) is an energy ratio of the desired signal and error signal between the desired signal and processed signal. It was used in [Sche 93] and [Sche 94] to facilitate the performance analysis of the ETHG algorithm. NSE is defined in Eq (2.32), and rewritten as

$$\text{NSE} = \frac{\sum_{n=0}^M e^2(n)}{\sum_{n=0}^M d^2(n)} \quad (3.4)$$

where M is the number of data sample points, and $e(n)$ is the error signal between the desired signal and processed signal. We note that the desired signal is available only for simulation, and is not available in reality. The NSE weights the error measured with respect to the signal power, so that large errors occurring during periods of large signal power are not so important as the same amount of error occurring during a period of small signal power. This ratio of error power versus desired power is very important in subjective quality, because the human auditory system tends to be more sensitive to the errors for small signal power period than the errors at large signal power period, due to the masking effect. So, a smaller NSE is highly correlated with better subjective quality [Beer 92]. However, we note that a higher NSE value does not necessarily mean a poorer audio quality, since NSE does not wholly correlate with the perceptive nature of the human ear.

NSE measure can be erroneous with a silence segment of the desired signal, which does not affect the accuracy of the error analysis. This erroneous portion may be avoided

by not including the silence segment of signal into calculation of NSE.

The mean normalized squared error (MNSE) is a mean value of K trials with M length signals, and it is statistically more reliable than the single NSE.

MNSE is defined as

$$\text{MNSE} = \frac{1}{K} \sum_{T=1}^K \left[\frac{\sum_{n=0}^M e^2(n)}{\sum_{n=0}^M d^2(n)} \right], \quad (3.5)$$

where K is the number of trials. In the computer simulation section in chapter 2, we used a target MNSE of .01 for constant envelope, single frequency case, and .02 for time varying envelope, single frequency case, to compare the performance of windows. As stated before in Chapter 2, these numbers were borne out of experience. We noticed that the MNSE is a convenient tool to compare different windows with different lengths, with good mathematical tractability.

3.3 Frequency Domain Distortion Measure

Frequency domain distortion measures generally use the second order statistics of the compared signal. The power spectrum is estimated by many techniques such as periodogram, modified periodograms, and linear prediction (LP) [Haye 96, Kay 88]. In general, a frequency domain distortion measure has more ability to explain the psychoacoustic perception characteristics [Kubi 91, Itak 70, Dell 93].

3.3.1 Distance Measure

The distance measure is the most common spectral distortion measure, and it shows the difference between the input power spectrum and output power spectrum, based on frames. A unweighted L_p norm is defined as [Kubi 91],

$$D = \left[\frac{1}{L} \sum_{i=0}^{L-1} \{PY_i - PX_i\}^p \right]^{\frac{1}{p}}, \quad (3.6)$$

where PY_i and PX_i represent the output and input power spectrum at the i th frequency, L is the number of frequency points, and p is the distance norm. The most common p selections are 1, 2, and ∞ , and 2 (root mean square) is most popular because of analytical tractability [Gray 80]. However, $p = 8$ provides better correlation to subjective quality than $p = 2$ [Kubi 91]. Note a larger value of p puts more emphasis on large spectral distances [Quac 88, Kubi 91].

This distance measure is insensitive to a slight misalignment of input signal and output signal frames because a slight misalignment between the signals causes negligible differences in the power spectrum. Therefore, it is a more reliable measure of the subjective quality than the time domain distortion measures [Kubi 91].

A frequency weighted L_p norm is

$$D = \left[\frac{\sum_{i=0}^{L-1} PX_i |PY_i - PX_i|^p}{\sum_{i=1}^{L-1} PX_i} \right]^{\frac{1}{p}}. \quad (3.7)$$

Eq (3.7) puts more weight on the distortion in the region of greater spectral energy [Kubi 91, Quac 88].

3.3.2 Other Spectral Distance Measures

Many spectral distance measures are generally designed to describe the subjective quality of specific application areas [Dell 93, Kubi 91, Quac 88]. Following are brief descriptions of some selected distortion measures.

The Information index uses the human auditory model to describe the transmission loss, circuit noise, room noise, attenuation, frequency distortion, and side tone [Kubi 91]. It divides the spectrum into 16 critical bands, and applies frequency weights and hearing thresholds to compute the signal-to-distortion (SDR) ratio. The SDR is then used for calculation of Information index. This process is complicated due to the weighting functions for frequency channels, and the hearing threshold functions.

The Coherence function describes a measure of signal-to-distortion ratio (SDR) weighted to account for hearing sensitivity, noise threshold effects, and the receiver sensitivity. It divides the speech frames into four quartiles based on amplitude [Kubi 91]. The Coherence function also is complicated due to the weighting functions, scale factor, and nonlinear mapping functions.

The Itakura distance measure, and Itakura-Saito (IS) distortion measure are based on the distance of AR parameters between input and output signal. The AR parameters are found by minimizing the IS distortion measure, which is equivalent to maximizing the likelihood function [Itak 70, Dell 93]. They are widely used in speech signal processing area.

3.4 Cepstral Distance Measure

This measure is based on the cepstral distance between input and output signal. The cepstral coefficients are derived from the parameters of the linear predictive coding (LPC), which uses an AR model described in Eq (1.15). The cepstral distance is defined as [Kubi 91],

$$CD = \frac{10}{\text{Log}_e 10} \left[2 \sum_{i=1}^m [C_x(i) - C_y(i)]^2 \right]^{\frac{1}{2}}, \quad (3.9)$$

where $C_x(i)$ and $C_y(i)$ are the i th cepstral coefficients of the input and output signal, and m is the number of the coefficients used. The cepstral coefficients are obtained from

$$\text{Log} \left(\frac{1}{A(z)} \right) = \sum_{k=1}^{\infty} c(k) z^{-k}, \quad (3.10)$$

where $c(k)$ is the cepstral coefficients, and $A(z)$ is the LPC model polynomial [Dell 93].

The cepstral distance measure is used in speech signal processing, where the LPC generated parameters provide an estimate of the smoothed speech spectrum [Dell 93].

3.5 Summary

We have discussed several time domain and frequency domain distortion measures in this section. Any one of them could have been chosen as the performance measure, but the NSE was chosen for the following reasons;

- 1) It offers good mathematical tractability,
- 2) A low error is highly correlated with good perceptual quality, and

3) it was also used in preliminary investigations by Scheets [Sche 93 Sche 94].

The SNR is basically an inverse of NSE, thus it could be used for the ETHG as well, but the NSE was chosen in part to maintain continuity with previous work.

We note that the NSE is not a good choice for the subjective quality measure if signal phase misalignment exists between input and output signals, as a large NSE may not necessarily correspond to bad perceptive quality. However, when the NSE amount is small, and it is small when an appropriate window type and length is applied which we observed in chapter 2, it has high correlation to good audio quality. Therefore, the NSE was selected as the major performance analysis tool for the ETHG algorithm.

CHAPTER IV

SIMULATION WITH AUDIO SIGNAL

4.1 Introduction

This chapter describes computer simulation results with real audio signals, and is composed as follows. Section 4.2 contains a brief introduction of sound generating mechanisms, and sound synthesis techniques. A review of the psychoacoustic nature of the human auditory perception system is in section 4.3, and a statistical analysis of selected audio signals is presented in section 4.4. Section 4.5 contains discussions about selecting the proper length of the Hanning window with zero padding, and the effects of the injected magnitude of the harmonics generated for 5 ~ 10 KHz interval on the subjective quality. Extension of harmonic generation to 10 ~ 15 KHz region, and the analysis of the perceptive quality is discussed in section 4.6. Section 4.7 is a summary of important findings in chapter 4.

4.2 Sound Generating Mechanism and Synthesis Techniques

Sound is a pressure wave propagated by disturbing air molecules. The disturbance is originated by vibrations of objects such as musical instruments, vocal cords, etc. The vibration can be either periodic or aperiodic. In order to properly design the sound

processing system, it is necessary to understand the generating mechanism and modeling of sounds from human and instruments. Following is a brief explanation of voice and musical sound analysis and synthesis techniques in use today.

4.2.1 Definitions

Before we discuss the generating mechanism and synthesis of sound, some definitions are in order.

Harmonic

The sinusoidal frequency components of a periodic sound which are integer multiples of the fundamental frequency.

Partial

The sinusoidal frequency components of a sound which are not integer multiples of the fundamental frequency.

Inharmonic

The sinusoidal frequency components of a sound which are not harmonically related.

Pitch

A subjective perception of sound quality (bass or treble) based on frequency.

Pure Tone

A tone for which the air pressure varies in a sinusoidal pattern with time.

4.2.2 Speech Model and Synthesis

The generally accepted speech production model by Schafer and Rabiner [Scha

75] describes speech as an output of a linear time varying filter which simulates the resonance characteristics of the vocal tract. For voiced speech, the filter is driven by a quasi-periodic unit sample generator which approximates the quasi-periodic flow of air from the lungs through the glottis, causing vibrations in the vocal tract. The filter is driven by a stationary white noise sequence for unvoiced speech [Scha 75]. Voiced speech is, in general, periodic, and the consonants (especially the fricatives) are inharmonic. The vocal tract is moving relatively slow compared to the input and output waveforms, so the system is approximately stationary for the duration of its memory, and it is called a “quasi-stationary” system [Port 81].

4.2.3 Analysis and Synthesis of Musical Sound

The sound from musical instruments is generally modeled and synthesized by additive, subtractive, and nonlinear methods. The additive method is a spectrum based harmonic analysis technique which can produce remarkably good instrument sounds. The basic idea of the additive method is that musical sounds are the result of summation of multiple sound components. In the additive model, the signal is assumed to be composed of multiple sinusoids with different amplitude and harmonic frequencies. The signal is expressed as,

$$x(n) = \sum_{k=1}^M A_k(n) \sin\{nT[kw + 2\pi F_k(n)]\} \quad (4.1)$$

where $x(n)$ is the signal at time nT , n is the sample number (time index), T is the time between consecutive samples, w is the radian fundamental frequency of the note, k is the harmonic number, $A_k(n)$ is the amplitude of harmonic k at time nT , $F_k(n)$ is the frequency

deviation of harmonic k at time nT , and M is the number of harmonics. The amplitude $A_k(n)$ and frequency deviation $F_k(n)$ are assumed to be slowly time varying [Moor77].

The additive method is well described by Fourier theorem in that a periodic waveform can be expressed as a sum of harmonically related sinusoids, each with a particular amplitude and phase. The phase vocoder which was developed by Flanagan and Golden in 1966, and applied to the analysis of musical sound by Moorer, is a very useful tool for analysis and manipulation of the parameters of the additive method. The additive method's main drawback is that it can be applied only to the isolated tones of nearly constant frequency, and the dedicated oscillators become quite expensive even on the single sound. For example, the number of dedicated oscillators required to cover all the partials for a single sound can easily be more than 100 [Moor 77, RMoo 90, Puck 95].

The subtractive method is based on the complementary idea that a sound is formed after subtracting unnecessary components from the complex sound block. The resulting time varying filter is driven by an excitation source to formulate the sound. This model has been extensively used in speech synthesis. Linear prediction, which uses an all pole model, is one of the popular methods to select the coefficients of a time varying filter. The subtractive method is not constrained to the isolated tones of constant frequency, but it has less fidelity in generated sound compared to that of the additive method [RMoo 90].

The nonlinear method is based on intuition and requires trial and error. This method is much simpler and needs less memory compared to the previous methods. The summation formula synthesis technique, especially the frequency modulation (FM) synthesis technique devised by Chowning [Chow 73] is most appropriate method for systems with limited resources [Moor 77]. Typical examples of non linear synthesis

methods are frequency modulation and nonlinear distortion. The generated sound does not exactly match to an arbitrary sound, but it provides a convenient and highly efficient method to control the sound parameters. Determining the parameters automatically by using an algorithm called Genetic Algorithm was proposed to overcome the drawback of trial and error nature [Chan 96].

4.3 Psychoacoustics

Via the human ears, sound is processed through a narrow 20 Hz to 20 KHz bandwidth auditory system, and perceived as a meaningful information by the mental judgment of the brain. For most people, the hearing bandwidth is even narrower. According to Hall, healthy young persons' upper limit is usually 17-18 KHz, and gradually decreasing to 12 KHz (women) or 5 KHz (men) by retirement age [Hall 91, Enne 74]. The ear is most sensitive in the region of 1 to 5 KHz, and has a logarithmic response to sound pressure levels. The human auditory organ is depicted in Fig. 4.1 [Pohl 95].

The sound waves pass through the ear canal, which resonates at about 3 KHz to provide extra sensitivity for speech intelligibility, and causes vibrations to the ear drum. These vibrations are transmitted by the middle ear organs to the cochlea, causing motion of the basilar membrane. The hair cells along the basilar membrane detect the vibrations and convey audio information via electrical impulses to the brain. These hair cells respond to the strongest vibrations in their local region, and are unable to distinguish nearby vibrations. This is the critical band, which works analogous to the spectrum analyzer with variable center frequencies. The hair cells send frequency components using

separate nerve fibers when they are separated more than the critical bandwidth, but send the frequency components that are within the critical bandwidth over same nerve fibers.

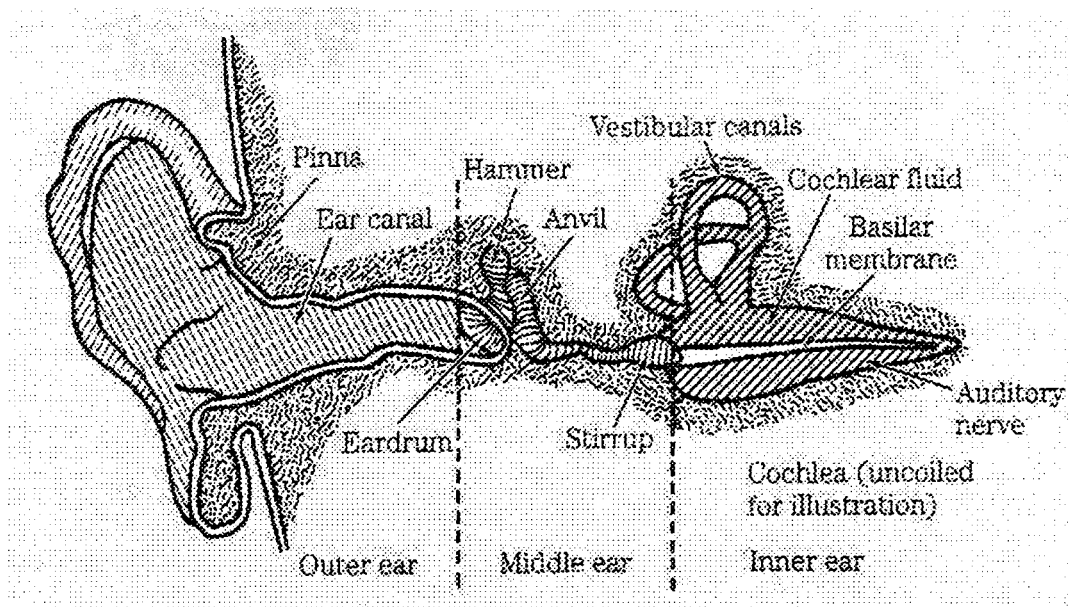


Fig. 4.1 The human auditory organ [Pohl 95].

The critical bands are much narrower at low frequencies, approximately 100 Hz wide between 20 Hz and 400 Hz, and approximately 1/5 octave wide for frequencies from 1 to 7 KHz. This means that the ear gets more information from low frequencies. This critical band concept is used in conceptual coding, which uses the masking phenomenon of the ear [Pohl 95, Ghit 94].

The ear perceives sound with higher frequency harmonics as bright or shrill, and a sound with dominant low frequency harmonics as dull, or not bright. Also, sound which is not periodic is neither clear nor bright. It will generally sound impure or unsteady [Pier 83, Hall 91]. These aspects justify an effort to develop the ETHG algorithm which has the potential to regenerate the missing higher harmonics due to the narrow bandwidth of current systems such as AM radio or the telephone system.

4.4 Audio Characteristics

In this section, general characteristics of an audio signal are addressed. Knowledge of audio characteristics such as harmonic structure, role of harmonic components, and the effects of manipulating harmonic structure are essential for designing the generated harmonic structure by the ETHG algorithm. Following are some of important aspects of audio signals.

Most musical signals and voiced speeches, on a short time basis, are nearly periodic with harmonically related frequencies, except unvoiced speeches and certain musical signals such as piano, bells, drums, and gong. The harmonic frequencies of the periodic signals are exactly or nearly equal to integer multiples of fundamental frequency [Brow 96, Port 81, Ando 93].

The intensity of the harmonics are changing with time. As the sound dies away, the higher harmonics have a lower peak amplitude than do the lower harmonics [Pier 83]. Also, a louder note tends to have more significant higher harmonics, or partials than a softer note [Jaff 95].

The shape of an audio signal determines the timbre, or tone quality [Enne 74, Hall 91]. For example, audio signals from two different instruments playing the same note with a same frequency or two people saying same speech with same frequency will not sound the same because of different wave shapes. Therefore, it is very important to follow the envelope shape in the ETHG algorithm to preserve the timbre.

4.4.1 Statistical Analysis of Audio Signals

The statistical characteristics of time varying spectral envelopes of full fidelity audio signals may provide important information for ETHG design, because bandlimited signals do not provide enough information as to the higher harmonic structure. Inserting generated higher harmonics should be done in a proper fashion in order to prevent additional distortion from wrong harmonic amplitudes. One possible method to figure out how to control the amplitude of the generated higher harmonic is a statistical analysis of the harmonic structure, even though we note that the frequency envelope of audio signals are time varying, thus the statistics at one time does not represent other times [Ando 93, Brow 96, Grei 89, Meye 93]. In other words, they are not stationary. Nevertheless, it is important to find general amplitude relationship between harmonics in order to minimize the distortions from outputting the wrong harmonic amplitude. This was done by analyzing the frequency roll-off of several different types of audio.

Sample audio signals are obtained from compact discs with length about 3 seconds long with 44.1 KHz sampling frequency at 8 bit resolution. An 8 bit resolution was used to reduce data amount, even though compact discs use 16 bit resolution. A 256 length FFT is performed with a rectangular window from data point 1 to 256, and the magnitude of each frequency bin is recorded. We continue to calculate 256 point FFTs until the entire 3 second audio clip has been transformed. Then the means of the magnitude of each bin are calculated for each audio signal. The mean of each bin is

$$M_{bin(k)} = \frac{1}{N} \sum_{i=1}^N |A_{k(i)}|, \quad (4.2)$$

where k is the bin number, N is total number of trials, and $A_{k(i)}$ is the amplitude of k th bin at i th trial.

Twenty four audio signals with length of about 3 seconds, at the sampling frequency of 44.1 KHz, were selected from various musical categories. Fig. 4.2 shows typical mean magnitude spectrums of audio signals. Fig. 4.2 (a), (b), (c), and (d) is from a classical music, piano, violin, and popular music, respectively. These various means show that the audio signal's statistics are dynamically varying both within a specific type of audio and in between different audio signals.

Table 4.1 shows the voltage ratios of 2.5 ~ 5 KHz range versus 5.1 ~ 10 KHz range, and 10.1 ~ 15 KHz range of the mean magnitude spectrum. The ratio 1 is a ratio of sum of mean magnitudes between 5.1 ~ 10 KHz divided by sum of mean magnitude between 2.5 ~ 5 KHz. The ratio 2 is a ratio of sum of mean magnitude between 10.1 ~ 15 KHz divided by sum of mean magnitude between 2.5 ~ 5 KHz. The average of ratio 1 and

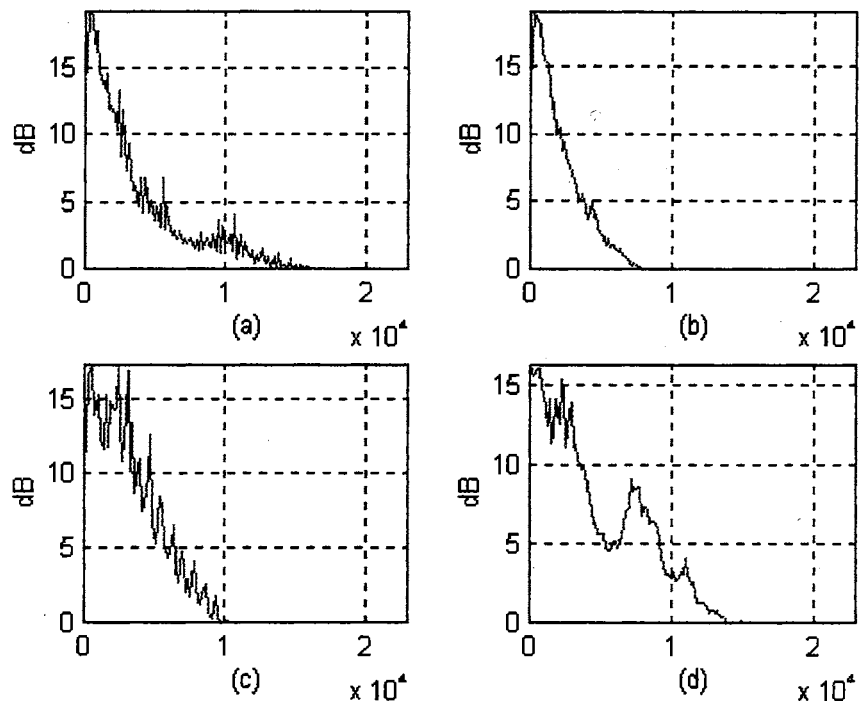


Fig. 4.2 Mean magnitude spectrum

(x-axis : frequency).

(a) : Classical music

(b) : Piano

(c) : Violin

(d) : Popular music

<i>AUDIO SIGNAL TYPE</i>	<i>MAGNITUDE RATIO 1 (5.1~10KHz)</i>	<i>MAGNITUDE RATIO 2 (10.1~15KHz)</i>
Orchestra 1	.9121	.6411
Orchestra 2	.3943	.1816
Orchestra 3	.6906	.4521
Orchestra 4	.5044	.1953
Orchestra 5	.5893	.4028
Orchestra 6	.9281	.6989
Orchestra 7	.536	.3753
Orchestra 8	.5832	.2577
Orchestra 9	1.0841	.7859
Orchestra 10	.8061	.6357
Piano	.5559	.3372
Violin (quiet phase)	.2643	.0875
Flute (solo)	.7924	.5512
Flute (with music)	.8242	.5343
Popular song 1	.5516	.3021
Popular song 2	1.1256	.4079
Popular song 3	.6775	.2329
Popular song 4	.8609	.3532
Popular song 5	.9316	.336
Popular song 6	.6077	.2759
FM Music 1	1.0213	.1521
FM Music 2	.9112	.3501
FM Music 3	.8931	.1001
FM Music 4	.8919	.4701
MEAN OF ALL	.7474	.3757

Table 4.1 Mean spectral magnitude ratios.

ratio 2 are shown at the bottom of Table 4.1, and they are .7474 and .3757, respectively.

The frequency range of 2.5 ~ 5 KHz was chosen, because later it will be assumed that we have an AM signal with 5KHz baseband bandwidth, and harmonics from this interval will be used to generate higher harmonics. Also, the fundamental frequency range of most musical instruments and human voice are within 2.5 KHz [Earg 95].

It appears that the ratio 2 is, on average, approximately half of ratio 1 for all audio signals. Individual ratio 1 appears to show some variation, but the average of ratio 1 from the 24 musical signals is .7474.

Bearing the average ratio 1 and ratio 2 in mind, we manually controlled the generated harmonic's amplitude, and tested the subjective sound quality.

4.5 Subjective Quality Analysis : Generation of Harmonics for 5 ~ 10 KHz Region.

The proper length of the Hanning window with zero padding techniques, and the effect of changing magnitude of generated harmonics will be discussed in this section. As described in section 2.3.5, the WSC (window selection criterion) does not contain any information of psychoacoustic perception quality of the MNSE difference. We observed that even though the 1024 length FFT with 512 length Hanning window padded with 512 zeros yield the lowest WSC value in Table 2.5, it yielded higher average MNSE than the 2048 FFT length, as shown in Table 2.6. Therefore, it is needed to test the effect of the difference in MNSEs. The sound quality was tested with the FFT length of 32, 64, 128, 256, 512, 1024, and 2048 with L/2 Hanning windowed data points and L/2 zeros padded after the window. Among these FFT lengths, the 2048 length FFT appears to generate a sound with the least amount of distortions. It was observed that with the louder portions

of musical passages, the distortion was hardly noticeable. Also, the distortion was slightly noticeable with a careful, repetitive listening for a single, less dynamic portion of musical signal. Hence, a 1024 length Hanning window with 1024 zeros padded after the window was selected for testing the subjective quality when changing the magnitude of generated harmonics. The value of $K = 128$ was selected, after testing several K values. We noticed that as the window length decreases, the sound quality gets degraded, which is closely related to the fact that the MNSE is increased as the window length decreased, which was observed in Chapter 2.

For this analysis, the initial goal is to take an AM quality signal, which is bandlimited to 5 KHz, and generate the harmonics for 5 ~ 10 KHz interval. To do this task, the harmonics in the frequency range of 2.5 ~ 5 KHz are used for generation of higher harmonics for the frequency range of 5 ~ 10 KHz interval. It was stated before, that the harmonic amplitude is generally decreasing as the frequency goes up, as seen in the previous section. Therefore, the effect of attenuating the amplitude of the generated harmonics will be investigated.

The audio signals used in the previous section were used for this analysis. To simulate an AM signal these audio signals were filtered with a 20th order Butterworth low pass filter. The cut off frequency was set to 5 KHz. This bandlimited signal was processed by the ETHG algorithm to generate an enhanced audio signal. The amount of enhancement can be controlled by adjusting the magnitude of generated harmonics. The magnitude of harmonics for 5 ~ 10 KHz interval was set from 1 to .2, on .2 scale. 1 means that the amplitude of generated harmonics are equal to the amplitude of the harmonics in 2.5 ~ 5 KHz interval.

A listening test was performed with several individuals. Among them a musician was included to see there are any distortions related to musical sense, i.e., change of note, unnatural sounds, etc. This test is informal, thus it needs to be done with a certified testing method in the future. For now, checking of any audible distortions, and any unnaturalness are the major concern, therefore this initial test suffices.

With a setting of 1, the enhanced music appears to be more brighter than the bandlimited signal, and it was clear that the sound had increased high frequency components. Noise-like distortions were hardly audible in most audio cases. With less than 1 setting, the degree of generated high frequency components clearly decreased. Some commented that with less than 1 setting (.6 or .4), they felt more comfortable. Note the setting of .6 is close to the average ratio 1 value in Table 4.1, which is .7474. This factor needs to be considered, because often times, hearing audio with significant high frequency energy may cause fatigue. Therefore, the amount of enhancement should be able to be adjusted to suit individual needs. This could be done by using a controlling knob, when the ETHG is implemented in real time, with a default value around .7.

We observed in this section that there were not any significant noise-like distortions, but a careful comparison with the full bandwidth original signal does reveal noticeable audible differences. The enhanced audio appears to be much brighter and crisper than the band limited audio, with tolerable amount of distortion and unnaturalness. Thus, harmonic enhancement extension to 10 ~ 15 KHz region in order to achieve “quasi-FM” quality, which requires 3rd and 4th harmonic generation, is discussed in next section.

4.6 Subjective Quality Analysis : Generation of Harmonics for 10 ~ 15 KHz region

The effect of extending harmonic generation and injection to the 10 ~ 15 KHz region is discussed in this section. In the previous section, we used the harmonics in the frequency range of 2.5 ~ 5 KHz to generate the second harmonics for 5 ~ 10 KHz region, whereas the harmonics 3.5 ~ 5 KHz region is used to generate harmonics for 10 ~ 15 KHz region in this section. The magnitude of generated harmonics was controlled around the ratio2 value, which is shown in Table 4.1. Fig. 4.3 shows an example of the extended harmonic generation. Fig. 4.3 shows original spectrum in the top plot, band limited spectrum in the middle, and the enhanced spectrum in the bottom, respectively. The horizontal axis shows the frequency from zero to 22.05 KHz, and the y axis shows the dB magnitude. Note the 0 ~ 5 KHz portion of the band limited spectrum looks similar to that of the original signal, due to the fact that the Butterworth filter has maximally flat response in the pass band and stop band. We observe from the enhanced spectrum that the harmonic estimates are, in general, located close to the position that corresponds to the original spectrum, especially in the interval of 5 ~ 10 KHz region. We also note that the gaps are evident in the enhanced spectrum, however these are probably not that important to the human perception because of the fact that the ear is sensitive to the spectral peaks rather than to the troughs.

The enhanced sounds, with the harmonic magnitude control values of around .8 for 5 ~ 10 KHz and around .3 for 10 ~ 15 KHz region, appear to be brighter and crisper than the ones with 5 ~ 10 KHz region enhanced only. While the distortion level seems to be increased slightly more, it does not cause a noticeably unpleasant or unnatural

sensation. As we have noticed in the previous section, although the enhanced audio sounded brighter and crisper, the amount of the enhancement seems to be able to be adjusted to satisfy individual preference. Also, this subjective test result is based on informal testing with limited number audio samples and people. It would be necessary to perform extensive testing in the future.

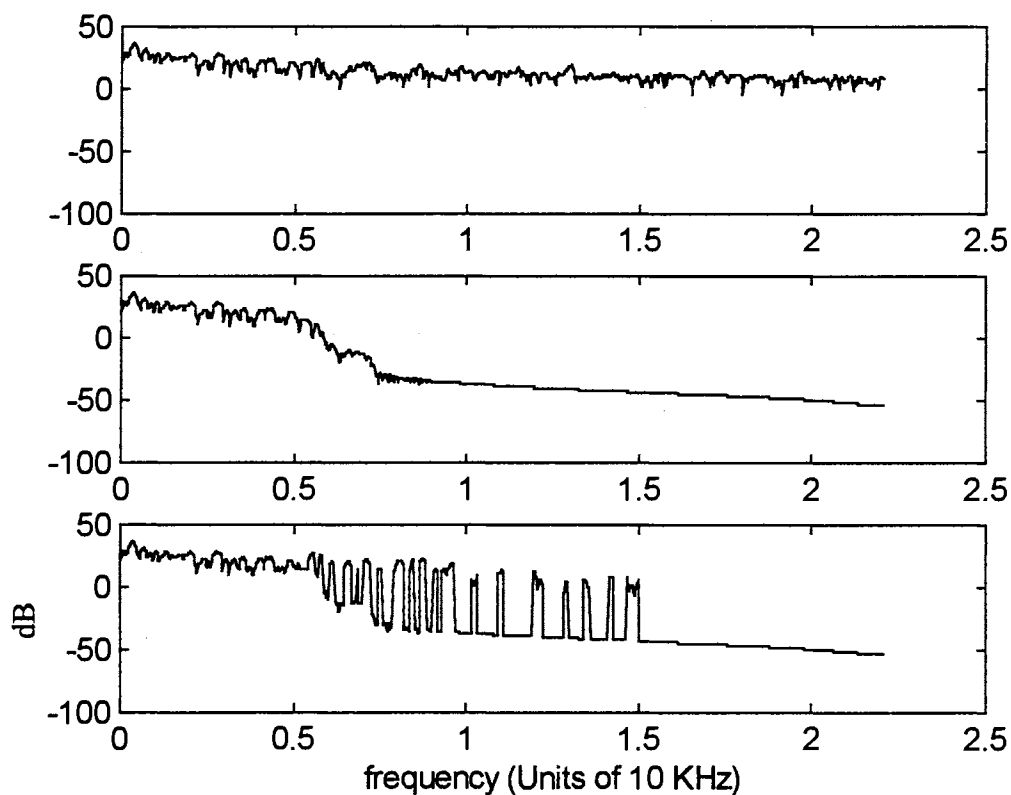


Fig. 4.3. Example of harmonic generation to 5 ~ 15 KHz region.
 Top : Original CD quality audio spectrum.
 Middle : Band limited (AM quality) spectrum.
 Bottom : Enhanced audio (quasi-FM quality) spectrum.

4.7 Summary

A brief introduction of sound generating mechanism, and sound synthesis techniques was presented in section 4.2. The psychoacoustic nature of human auditory perception system was discussed in section 4.3, and statistical analysis of selected audio signals was presented in section 4.4. The mean magnitude spectrums of 24 musical signals were inspected, and the general relationships between regions of 2.5 ~ 5 KHz and 5.1 ~ 10 KHz, and 5.1 ~ 10 KHz and 10.1 ~ 15 KHz were found. In section 4.5, discussions on selecting the proper length of the Hanning window with zero padding, and the effects of magnitude controlling of harmonics generated for 5 ~ 10 KHz interval on the subjective quality were presented. It was noted that a Hanning windowed 2048 point FFT with a 1024 point data signal and zero padding yields acceptable sound quality. This result agrees with the prediction, which was based on the theoretical analysis of the convolution result between the impulse response of the frequency truncation function and the window function in the section 2.3.4. The need to control the magnitude of the enhancement to suit individual preference of the people was discussed. Also, extending harmonic generation for 10 ~ 15 KHz region to achieve quasi-FM quality audio resulted in a brighter sensation. There is slightly increased audible distortion compared to the original. These listening tests are based on a limited number of people, thus extensive subjective quality tests in the future is needed.

CHAPTER V

CONCLUSION

5.1 Summary

The goal of this work is to analyze the ETHG algorithm's behavior and find the parameters which minimize the MNSE by investigation of window type, window length, and zero padding, and which also results in an acceptable computation load. To accomplish the goal, the following key works were completed.

In section 2.3.2, a model envelope waveform generation scheme, which is approximately same as the real audio envelopes in the first and second order statistical sense and lends mathematical tractability, was developed. This model was used for simulations in chapter 2. The AR parameters were obtained by the model fitting procedure of Box and Jenkins [BoxJ 94].

In section 2.3.4, the theoretical convolution relationship between the truncation function's impulse response and the window was investigated. It was noticed that the convolution results in distortion to the waveform due to the Gibbs phenomenon, and the Hanning window produced the least amount of distortion because it smoothes out the Gibbs phenomenon. A theory based prediction was made that the Hanning window would yield the best performance of the tested windows, when the model signal and the real

audio signal are processed by the ETHG algorithm. Also, a discussion about the ideal window type, and the reason why it is unrealizable was included in section 2.3.4. The average mean squared error was shown, by mathematical derivation and computer simulations, to be minimum in the middle of the processed output window for time varying envelopes with a rectangular, Hamming, Hanning, and zero padded Hanning window. The computer simulation results in section 2.3.5 and section 2.3.6 indicate that the zero padded Hanning window produced the least amount of error among the compared windows, which shows agreement with the theoretical prediction.

The window selection criterion (WSC) for selecting optimal window size and type, based on a time domain distortion measure, the MNSE, and the computation load was developed in section 2.3.5.

The Normalized Squared Error (NSE) was chosen as the major performance analysis tool for the ETHG algorithm, based on an examination result of both time and frequency domain distortion measures in Chapter 3. The main advantage of the NSE is that it is very tractable mathematically, and low values correspond to a good perceptive sound quality.

A statistical analysis of selected audio signals was performed, and the general relationship between the harmonics was found in chapter 4. Also, it was found in chapter 4 that the reduced error obtained by using the zero padding technique is directly related to good sound quality. This implies that the lower error corresponds to better perceptive quality, however high error does not necessarily mean a poor quality because the human ear is insensitive to the phase delay. We noticed, by hearing tests with several people, that the audio signals with enhanced harmonics for 5 ~ 15 KHz interval by using

the zero padded Hanning window produced smaller amounts of audible noise-like distortions, compared to Hanning window without zero padding and other window types. The enhanced audio sounded brighter, crisper, and generally appears to have more high frequency range with a tolerable amount of distortion and unnaturalness compared to the band limited input signal which sounded relatively dull and muffled. However, one can notice audible differences between the full fidelity original signal and the enhanced signal when both signals are compared side by side.

Although the enhanced audio appears to be positively brighter and crisper, such sensation may be different to different people, thus it is necessary to include the option of controlling the magnitude of enhancement to suit individual preferences.

5.2 Future Research Considerations

Presently, it seems that no other algorithm specifically designed for enhancing band limited audio signals exists. This dissertation serves as an introduction to this arena, thus there is still much research that can be done toward an optimal algorithm which possesses reasonably small computation load and yields acceptable subjective quality.

We noticed a good improvement of sound quality by artificially generating spectral energy into the 5 ~ 15 KHz range. Although the enhanced audio appears to be positively brighter and crisper than the band limited audio, which sounds dull and muffled, it is necessary to perform extensive subjective analysis of the enhanced audio quality in the future, because the result above is based on the listening test with limited number of people. After that, efforts to implement the ETHG algorithm in real time appear to be worthwhile, in part due to the rapid improvements of the DSP chip technology in recent

years. Further study to reduce the computation load would be worthwhile also. One possible method would be developing and implementing a non-linear FFT algorithm, which is capable of generating the FFT of required spectral regions with high precision and a lower amount of computation complexity in order to expedite the harmonic generating and injecting process.

We also note that the quality evaluation of enhanced audio signal requires subjective perception. It is believed that no known objective measure has the absolute capability to explain the subjective performance of the ETHG algorithm. Therefore, such an objective measure still remains to be found. Further research to include subjective audible distortion into the WSC value, in order to find the best combination of computational complexity versus perceived quality, would be one of the efforts to develop such an objective measure.

Finally, research efforts to find a mathematical equation(s) which can predict the value of the Normalized Squared Error (NSE), given parameters such as window type, length, and zero padding would be worthwhile. Obtaining such predicted outcomes in advance would provide more flexibility to the ETHG design process, and may enable the designers to achieve the best possible subjective quality.

REFERENCES

- [Ando 93] S. Ando, and K. Yamaguchi, "Statistical study of spectral parameters in musical instrument tones," *J. Acoust. Soc. Am.* vol. 94, No. 1, pp.37-45, July, 1996.
- [Beer 92] J.G. Beerends, and J.A. Stemerdink, "A Perceptual Quality Measure Based on a Psychoacoustic Sound Representation," *J. Audio Eng. Soc.*, vol. 40, No.12, pp. 963-978, Dec, 1992.
- [BoxJ 94] G.E.P. Box, G.M. Jenkins, and G.C. Reinsel, *Time Series Analysis: Forecasting and Control*, Prentice Hall, Englewood Cliffs, N.J, 1994.
- [Brow 93] J.C. Brown, and M.S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the Fourier transform", *J. Acoust. Soc. Am.* vol. 94, No. 2, pp. 662-667, Aug, 1993.
- [Brow 96] J.C. Brown, "Frequency ratios of spectral components of musical sounds," *J. Acoust. Soc. Am.* vol. 99, No. 2, pp. 1210-1218, Feb, 1996.
- [Chan 96] S. Chan, and A. Horner, " Discrete Summation Synthesis of Musical Instrument Tones Using genetic Algorithms," *J. Audio Eng. Soc.*, vol. 44, pp. 581-591, July/Aug, 1996.
- [Cohe 95] L. Cohen, *Time-Frequency Analysis*, Prentice Hall, Englewood Cliffs, N.J, 1995.
- [Dell 93] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete-Time Processing of Speech Signals*, Prentice Hall, Upper Saddle River, N.J, 1993.
- [Dorf 93] R.C. Dorf, *The Electrical Engineering Handbook*, IEEE Press, 1993.
- [Earg 95] J.M. Eargle, *Music, Sound, and Technology*, Van Nostrand Reinhold, New York, 1995.
- [Enne 74] H.E. Ennes, *AM-FM Broadcasting Equipment, Operations, and Maintenance*, Howard W. Sams & Co, New York, 1974.
- [Flan 72] J.L. Flanagan, *Speech Analysis, Synthesis and Perception*, 2nd Ed., Springer Verlag, 1972.

- [Ghit 94] O. Ghitza, "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 115-132, Jan 1994.
- [Gray 88] R.M. Gray, A. Buzo, A.H. Gray, JR, and Y.M. Matsuyama, "Distortion for Speech Processing," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 367-376, Aug 1980.
- [Grei 89] R.A. Greiner, and J. Eggers, "The Spectral Amplitude Distribution of Selected Compact Discs," *J. Audio Eng. Soc.*, vol. 37, Apr, 1989.
- [Haga 94] M. Hagan, *ECEN 5523 Estimation Theory Class*, Oklahoma State University, Stillwater, OK, Spring, 1994.
- [Hall 91] D.E. Hall, *Musical Acoustics*, Brooks/Cole Publishing Company, Pacific Grove, CA, 1991.
- [Hans 83] B.A. Hanson, D.Y. Wong, and B.H. Juang, "Speech Enhancement With Harmonic Synthesis," *ICASSP '83*, pp. 1122-1125.
- [Harr 78] F.J. Harris, "On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform," *Proc. IEEE*, vol. 66, pp. 51-83, Jan, 1978.
- [Haye 96] M.H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, New York, 1996.
- [Hayk 90] S. Haykin, *Modern Filters*, Macmillan Publishing Co, N.Y. 1990.
- [Hayk 96] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, Englewood Cliffs, N.J., 1996.
- [Itak 70] F. Itakura, and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," *Electronic and Communication of Japan*, vol. 53-A, pp. 36-43, 1970.
- [Jack 96] L.B. Jackson, *Digital Filters and Signal Processing*, 3rd Ed., Kluwer Academic Publishers, Boston, MA, 1996.
- [Jaff 95] D.A. Jaffe, "Ten criteria for evaluating synthesis techniques," *Computer Music Journal*, 19:1, pp. 76-87, Spr, 1995.
- [Kay 81] S.M. Kay, and L.M. Marple, JR, "Spectrum Analysis- A Modern Perspective," *Proc. IEEE*, vol. 69, pp. 1380-1419, Nov, 1981.

- [Kay 88] S.M. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice Hall, Englewood Cliffs, N.J, 1988.
- [Kubi 91] R.F. Kubichek, "Standards and Technology Issues in Objective Voice Quality Assessment," *Digital Signal Processing*, 1, pp. 38-44, 1991.
- [Kunt 86] M. Kunt, *Digital Signal Processing*, Artech House, Norwood, MA, 1986.
- [Makh 79] J. Makhoul, and M. Beriuti, "High Frequency Regeneration in Speech Coding Systems," *ICASSP 79*, pp. 428-431.
- [Mend 95] J.M. Mendel, *Lessons in Estimation Theory for Signal processing, Communications, and Control*, Prentice Hall, Englewood Cliffs, N.J, 1995.
- [Meye 93] J. Meyer, "The Sound of Orchestra," *J. Audio Eng. Soc.*, vol. 41, Apr, 1993.
- [Moor 77] J.A. Moorer, "Signal Processing Aspects of Computer Music-A Survey," *Computer Music Journal*, 1(1), pp.4-37, 1977.
- [Oppe 89] A.V. Oppenheim, and R.W. Schaffer, *Discrete Time Signal Processing*, Prentice Hall, Englewood Cliffs, N.J, 1989.
- [Pail 92] B. Paillard, P. Mabilieu, and S. Morissette, "PERCEVAL : Perceptual Evaluation of the Quality of Audio Signals," *J. Audio Eng. Soc.*, vol. 40, No. 1/2, pp. 21-30, Jan/Feb, 1992.
- [Pank 83] A. Pankratz, *Forecasting With Univariate Box-Jenkins Models : Concepts and Cases*, John Wiley & Sons, 1983.
- [Park 86] R. Parker, and S.A.T. Stoneman, "On the use of Fast Fourier Transform when high-frequency Resolution is required," *J. Sound and Vibration*, vol. 104, pp. 75-79, 1986.
- [Pier 83] J.R. Pierce, *The Science of Musical Sound*, Scientific American Books, New York, 1983.
- [Pohl 95] K.C. Pohlman, *Principles of Digital Audio*, 3rd Ed, McGraw-Hill, New York, 1995.
- [Port 80] M.R. Portnoff, "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis", *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 55-69, Feb, 1980.

- [Port 81] M.R. Portnoff, "Short-Time Fourier Analysis of Sampled Speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 364-373, June 1981.
- [Puck 95] M. Puckette, "Formant-Based Audio Synthesis Using Nonlinear Distortion," *J. Audio Eng. Soc.*, vol. 43, pp.40-47, Jan/Feb, 1995
- [Quac 88] S.R. Quackenbush, T.P. Barnwell III, and M.A. Clements, *Objective Measures of Speech Quality*, Prentice Hall, Englewood Cliffs, N.J, 1988.
- [Rabi 78] L.R. Rabiner, and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice Hall, Englewood Cliffs, N.J. 1978.
- [Rey 84] R.F. Rey, *Engineering and Operations in the Bell System*, AT&T Bell Laboratories, Murray Hill, N.J, 1984.
- [RMoo 90] F.R. Moore, *Elements of Computer Music*, Prentice Hall, Englewood Cliffs, N.J, 1990
- [Scha 75] R.W. Schafer, and L.R. Rabiner, "Digital representations of speech signals," *Proc. IEEE*, vol.63, pp. 662-677, Apr.1975.
- [Sche 93] G. Scheets, "An All Digital Envelope Tracking Harmonic Generator," unpublished manuscript, 1993.
- [Sche 94] G. Scheets, K. Woolverton, and W. Huang, "An All-Digital Envelope Tracking Harmonic Generator: An Overview," *IEEE Wichita Conference on Communications, Networking and Signal Processing*, pp. 47-51, Apr. 1994.
- [Serr 90] X. Serra, and J. Smith, "Spectral Modeling Synthesis: A Sound Synthesis/Synthesis System Based on a Deterministic plus Stochastic Decomposition," *Computer Music Journal*, vol. 14, No. 4, Winter, 1990.
- [Shan 88] K.S. Shanmugan, and A.M. Breipohl, *Random Signals : Detection, Estimation, and Data Analysis*, John Wiley & Sons, New York, 1988.
- [Smit 87] J.O. Smith, and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," *Proceedings of the 1987 International Computer Music Conference : University of Illinois at Urbana-Champaign*, pp. 290-297, Aug, 1987.
- [Smit 91] J.O. Smith, "Viewpoints on the History of Digital Synthesis", *ICMC-91*, pp. ICMC 1-9. 1991.

- [Soli 90] S.S. Soliman, and M.D. Srinath, *Continuous and Discrete Signals and Systems*, Prentice Hall, Englewood Cliffs, N.J, 1990.
- [Stee 80] R.G.D. Steel, and J.H. Torre, *Principles and Procedures of Statistics, A Biometrical Approach*, 2nd Ed, McGraw-Hill, New York, 1980.
- [Stre 90] F.G. Stremler, *Introduction to Communication Systems*, Addison Wesley, Reading, MA, 1990.
- [Tab 88] M. Tabei and M. Ueda, "A Method of High-Precision Frequency Detection with FFT," *Electronics and Communications in Japan*, vol. 71 pp. 24-32, 1988.
- [Vase 92] S.V. Vaseghi, and R. Frayling-Cork, "Restoration of Old Gramophone Recordings", *J. Audio Eng. Soc.* Vol. 40, No. 10, pp. 791-800, Oct, 1992.
- [Wang 82] D.L. Wang, and J.S. Lim, "The unimportance of phase in speech enhancement," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 679-681, Aug 1982.
- [Widr 85] B. Widrow, and S.D. Stearns, *Adaptive Signal Processing*, Prentice Hall, Englewood Cliffs, N.J, 1985.

APPENDIX I

MODELING PROCESS

1. Introduction

In this section, an overview of system identification process by Box and Jenkins [BoxJ 94], and an example of the process using envelope signal from real audio signal will be presented.

2. Overview of System Identification Process

The system identification procedure for estimating ARMA model was developed by Box and Jenkins [BoxJ 94]. The iterative procedure, which contains four steps, is shown in Fig. A.1 [BoxJ 94]. Fig. A.1 describes the process as follows.

Step1 : Preprocessing

Testing stationarity of the data to be analyzed is the first step of the modeling process. If the data is not stationary, it can be made to be stationary by removing the linear trend or periodic component. This process is referred to as the differencing. First or second differencing suffices in general. The data length should be at least 50 to generate reasonable estimate [BoxJ 94, Pank 83].

Step 2 : Order Identification

In this step we obtain estimate of the order of the data. The order of autoregressive part of the model is p , and the order of moving average part is q . To get p and q , we observe the autocorrelation function (ACF) and the partial autocorrelation

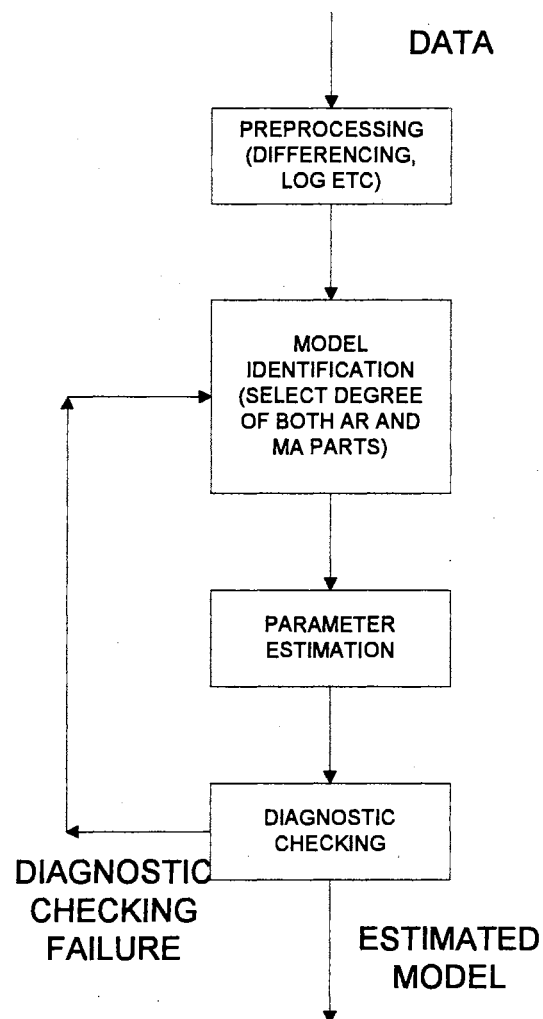


Fig. A.1 Box-Jenkins method of estimation sequence [BoxJ 94].

function (PACF) of the data. We use the fact that the ACF $r(\tau) = 0$ for $\tau > p$ in moving average process, and the PACF $\phi_{mm} = 0$ for $m > p$ in autoregressive process.

ARMA model has a generalized partial autocorrelation (GPAC) pattern of $\phi_{mm}^j = 0$ for $j = q$ and $m > p$, and $\phi_{mm}^j = \phi_p$ for $m = p$ and $j \geq q$, i.e.,

$$\phi_{mm}^j = 0 \text{ for } j = q \text{ and } m > p, \text{ and}$$

$$\phi_{mm}^j = \phi_p \text{ for } m = p \text{ and } j \geq q. \quad (\text{A.1})$$

The pattern in GPAC ϕ_{mm}^j is shown in the Fig. A.2 [Haga 94]. Note the highlighted pattern.

	<i>m</i>								
	0	1	2	3	...	<i>p</i>	<i>p</i> +1	<i>p</i> +2	<i>p</i> +3
<i>j</i>	0	ϕ_{m11}^0	ϕ_{22}^0	ϕ_{33}^0	...	ϕ_{pp}^0
	1	ϕ_{11}^1	ϕ_{22}^1	ϕ_{33}^1	...	ϕ_{pp}^1

	<i>q</i>	ϕ_{11}^q	ϕ_{22}^q	ϕ_{33}^q	...	ϕ_p	0	0	0
	<i>q</i> +1	ϕ_p	0/0	0/0	0/0
	<i>q</i> +2	ϕ_p	0/0	0/0	0/0
	<i>q</i> +3	ϕ_p	0/0	0/0	0/0

Fig. A.2 GPAC [Haga 94].

The PACF can be obtained by using Cramer's rule, or Levinson algorithm [Haga 94, Pank 83, Haye 96]. Using the Cramer's rule, PACF is defined as [Haga 94, Dorf 93],

$$\phi_{mm}^j = \frac{\det \begin{bmatrix} r(0) & \dots & \dots & r(j+1) \\ r(j+1) & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ r(j+m-1) & \dots & \dots & r(j+m) \end{bmatrix}}{\det \begin{bmatrix} r(j) & r(j-1) & \dots & r(j-m+1) \\ r(j+1) & r(j) & \dots & r(j-m+2) \\ \dots & \dots & \dots & \dots \\ r(j+m-1) & r(j+m-2) & \dots & r(j) \end{bmatrix}}, \quad (\text{A.2})$$

where $\det [\bullet]$ denotes a determinant of a matrix.

Step 3 : Parameter Estimation

After estimation of the order of data, which will be ARMA (p,q), AR (p), or MA (q), parameter estimation algorithms are used to generate estimate of the parameters. For example, the least squares (LS) may be used for AR process, and the maximum likelihood estimator may be used for ARMA process [Haye 96, Hayk 96]. A discussion of the relationship between the AR power spectrum and the autocorrelation function (ACF) seems to be in order here. Following is a brief discussion of the relationship between the ACF and power spectrum of AR process, and how the estimates of AR parameters are obtained.

The ACF and power spectrum of AR process is related by the Wiener-Khinchin theorem as [Haye 96, Kay 88],

$$\frac{\sigma_v^2}{\left| 1 + \sum_{k=1}^P \phi_k z^{-k} \right|^2} = \sum_{k=-P}^P r(k) z^{-k}, \quad (\text{A.3})$$

where σ_v^2 is a variance of input signal, which is assumed as white noise,

$$z = e^{j2\pi k/N}, \quad (\text{A.4})$$

and ϕ_k are the true AR parameters we want to estimate. The AR order of a given realization should be estimated, before any attempt to get the estimates of parameters. The order of AR process is obtained by inspection of ACF and partial autocorrelation function (PACF), as described in *Step 2*. By using the symmetry nature of the ACF, we may express Eq (A.3) in matrix notation, which is known as Yule-Walker equation [Haye 96, Hayk 96],

$$\begin{bmatrix} r(0) & r(1) & \dots & r(P-1) \\ r(1) & r(0) & \dots & r(P-2) \\ \dots & \dots & \dots & \dots \\ r(P-1) & r(P-2) & \dots & r(0) \end{bmatrix} \begin{bmatrix} 1 \\ \phi_1 \\ \dots \\ \phi_P \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \dots \\ r(P) \end{bmatrix}. \quad (\text{A.5})$$

The estimate of AR parameters $\hat{\phi}$ may be obtained by solving Eq (A.5) with the Cramer's rule, or the Levinson-Durbin recursion algorithm, using ACF. Another algorithms to get estimates directly from the data set are least squares algorithm (LS), and maximum likelihood estimator (MLE) [Haye 96, Hayk 96, Haga 94]. The result is the same regardless of algorithm chosen.

After parameter estimates are obtained, the stationarity condition should be inspected. For AR (1) or ARMA (1,q) system, the stationarity requirement is that [Pank 83],

$$|\hat{\phi}_1| < 1, \quad (\text{A.6})$$

and for AR (2) or ARMA (2,q) system, following three conditions must be satisfied.

$$|\phi^{\wedge}_2| < 1$$

$$\phi^{\wedge}_2 + \phi^{\wedge}_1 < 1 \quad (\text{A.7})$$

$$\phi^{\wedge}_2 - \phi^{\wedge}_1 < 1.$$

When $p > 2$, the necessary (but not sufficient) conditions is,

$$\phi^{\wedge}_2 + \phi^{\wedge}_1 + \dots + \phi^{\wedge}_p < 1. \quad (\text{A.8})$$

In other words, the stability condition requires all poles inside the unit circle [Oppe 89].

Step 4 : Diagnostic Checking

There is always possibility that the estimated parameters are not correct. Thus it is needed to perform the diagnostic checking. If diagnostic checking results are satisfactory, then the parameters can be used for modeling. Otherwise, steps 1-3 should be performed again.

Diagnostic checking is usually based on residual errors. The variance of the residual should be reasonably small. When making a selection from multiple models, usually the model with the minimum variance is selected. The covariance of the parameter tells the accuracy of the estimated parameters. The 95 % of true value of parameters will be in between 2 square root of covariance [Pank 83, Haga 94], i.e.,

$$\text{True parameter value} = \text{Estimated parameter} \pm 2\sqrt{\text{Co variance}}. \quad (\text{A.9})$$

The whiteness of the residual should also be inspected, to make sure the residual is not correlated in any manner. In theory, if the parameter estimation is perfect, the residual should be white, i.e., there is no correlation between the residual data points. The ACF and PACF are inspected for whiteness test. The ACF should be, except zeroth lag, within 2 standard error defined as [Pank 83],

$$s(r_k) = \sqrt{\frac{1 + 2 \sum_{j=1}^{k-1} r_j^2}{n}}, \quad (\text{A.10})$$

where k is the number of autocorrelation lag, n is the number of the data points.

The *t-statistic test* also shows quantitative measure of how the estimated parameter is far from the hypothesis on a certain confidence level [Pank 83, Stee 80].

It is defined as,

$$t_k = \frac{r_k - \rho_k}{s(r_k)}, \quad (\text{A.11})$$

where ρ_k is the hypothesized autocorrelation value, which is a zero. When t value is greater than absolute 2, we reject the hypothesis that the autocorrelation at that lag is zero, meaning it shows a correlation. The standard error and *t-test* yield the same result, thus the standard error was used in this paper.

The standard error for PACF is [Pank 83],

$$s(\hat{\phi}_{kk}) = \frac{1}{\sqrt{n}}, \quad (\text{A.12})$$

where n is the data points used for calculation of the PACF.

Another method of hypothesis testing is the Portamanteau lack of fit test, which is known as chi-square test [Haga 94, Shan 88]. It calculates the sum of the square of the residual autocorrelation using the formula,

$$Q = N \sum_{\tau=1}^k Ra^2(\tau), \quad (\text{A.13})$$

where k is the number of autocorrelation lag, N is the number of data used for residual autocorrelation calculation, and Ra is the residual autocorrelation. This Q value is used

with degree of freedom $df = (k-p-q)$ to check the whiteness at certain confidence level in the chi-square distribution table.

The modeling procedure described above is explained using an audio envelope, which is an output of low pass filter in Fig. 2.3, as follows.

3. Example

Fig. A.3 shows an example of typical smoothed output envelope from the low pass filter in Fig.2.3, when a segment of a real audio signal with 23 msec long data length at 44.1 KHz sampling frequency is an input.

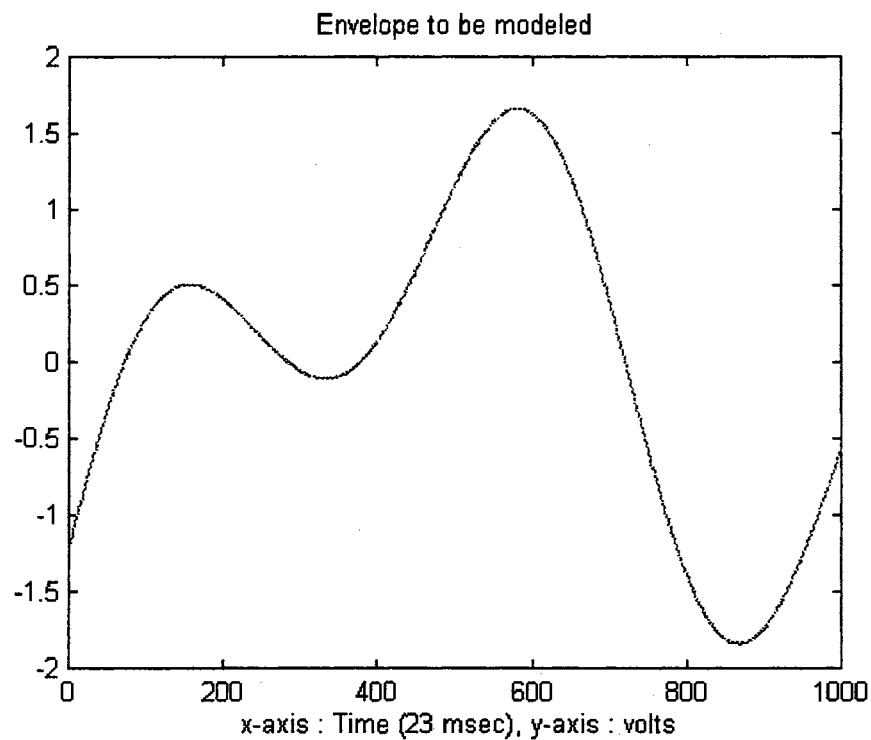


Fig. A.3 An envelope to be modeled.

In step 1, the GPAC in Fig. A.4 is inspected to find a pattern. We see three possible patterns in the GPAC, which are AR (p), ARMA (2,1), and ARMA (3,2). These three possible models were analyzed by using the Box-Jenkins procedure, and it turned out that the AR model describes the frequency spectrum of the envelope more accurately, although details of the comparison of AR (p) model with the ARMA (p,q) is not presented here. Fig. A.5 shows a comparison of AR (p) models' normalized frequency response (magnitude), where $p = 10, 50, 80$, and 100. We see that the AR (100) model's normalized frequency response, denoted as AR (100) in Fig. A.5, closely approximates the frequency spectrum of the envelope in Fig. A.3, which is denoted as FFT(envelope) in Fig. A.5. Following is a discussion of the procedure to obtain the AR (100) parameters, which is a typical example.

0.9991	-0.0520	-0.0494	-0.0470	-0.0449	-0.0429
0.9990	-0.9983	0.0000	0.0000	0.0000	0.0000
0.9989	-0.9983	0.9990	0.0000	0.0000	0.0000
0.9988	-0.9982	0.9989	0.0000	0.0000	0.0000
0.9987	-0.9982	0.9989	0.0000	0.3013	0.0102
0.9986	-0.9981	0.9988	0.0000	0.0109	0.8918
0.9985	-0.9981	0.9988	0.0000	-100.9955	-0.0274
0.9984	-0.9981	0.9988	0.0000	-0.1298	-1.5904
0.9983	-0.9980	0.9988	0.0000	-0.0087	1.5930
0.9982	-0.9980	0.9988	0.0000	6.0622	-0.0596

Fig. A.4 GPAC of the envelope signal in Fig. A.3.

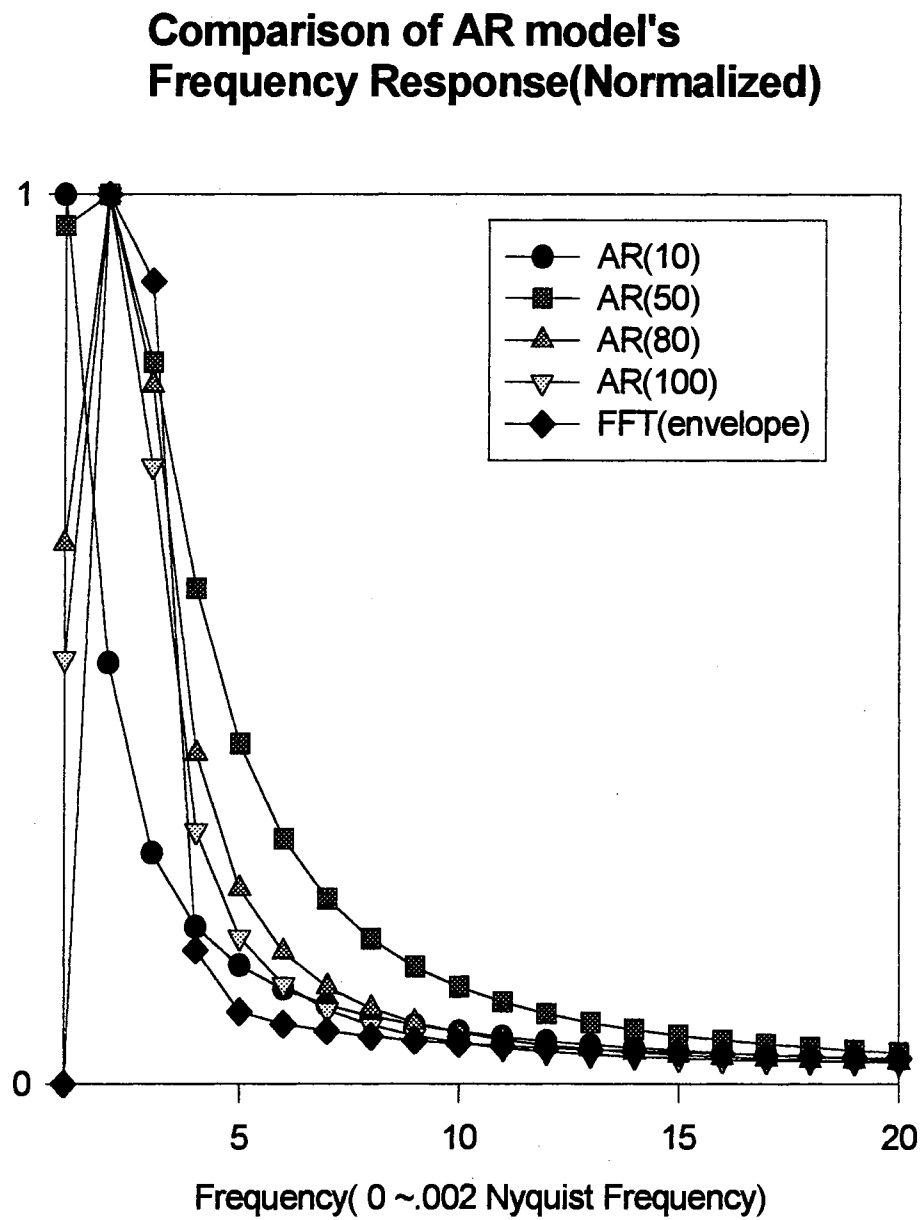


Fig. A.5 Comparison of AR model's normalized frequency response.

The parameters of AR (100) model were obtained by the maximum likelihood estimator (MLE) program, which is attached at the end of Appendix I. Table A. 1 shows the result of the MLE program. It shows the AR parameters, the covariance, the variance of the residual, and the hypothesis testing results.

The estimated AR parameters are

$$\begin{bmatrix} \hat{\phi}_1 \\ \hat{\phi}_2 \\ \dots \\ \hat{\phi}_{100} \end{bmatrix} = \begin{bmatrix} 1.0112 \\ .0006 \\ \dots \\ -.0001 \end{bmatrix}, \quad (\text{A.14})$$

and the variance of the residual is .0016. The covariance matrix of the parameter estimate is a measure of the accuracy. The true parameters will be in between

$$\begin{bmatrix} \phi_1 \\ \phi_2 \\ \dots \\ \phi_{100} \end{bmatrix} = \begin{bmatrix} 1.0112 \pm 2\sqrt{.0011} \\ .0006 \pm 2\sqrt{.0022} \\ \dots \\ -.0001 \pm 2\sqrt{.0011} \end{bmatrix}. \quad (\text{A.15})$$

The standard error test for ACF and PACF show that none of the ACF and PACF lag is more than two standard error. Fig. A.6 shows the ACF and PACF of the residual, and they appear to be white.

The chi-square test value is $q_1 = .0842$. We used $k = 150$ lags for Q value computation, thus with $p = 100$, the degree of freedom $df = (150 - 100) = 50$. From the chi-square distribution table [Stee 80], we get 28.0 with a hypothesis that 95 % of the peaks falls within 2 standard error interval. Since $Q < 28.0$, we accept the hypothesis.

Converged estimated parameters

-- AR parameters --	-- Covariance --
1.0112	0.0011
0.0006	0.0022
-0.0004	0.0022
-0.0001	0.0022
-0.0001	0.0022
-0.0001	0.0022
...	...
-0.0001	0.0022
-0.0056	0.0011

Estimated variance of residual

.0016

----- HYPOTHESIS TESTING -----

q1 =

0.0842

Number of lags exceeding 2 standard error range

ACF : 0

PACF : 0

Table A.1 Result of MLE program

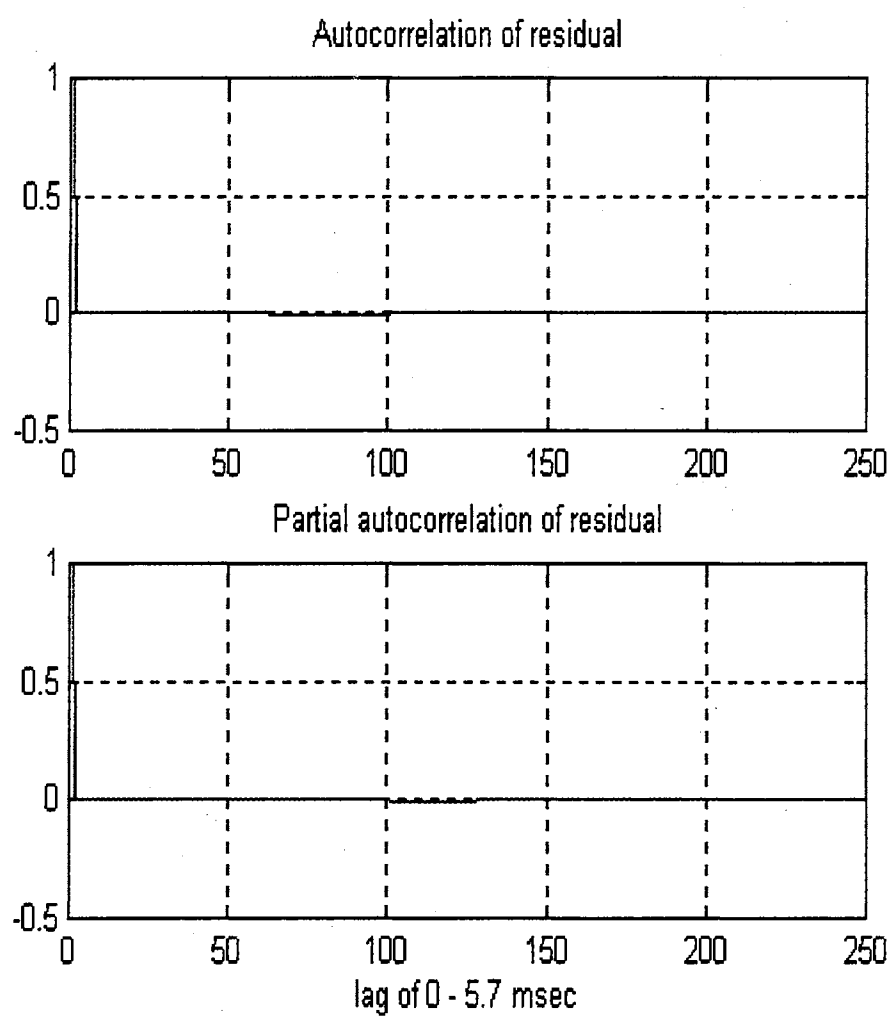


Fig. A.6 Plot of ACF and PACF of residual.

The stationarity of the parameter estimates should also be inspected. The sum of the AR parameters are .9951, which is barely less than 1, and all poles are within the unit circle. Thus, it is a stable model.

So far, we have discussed the modeling procedure using an envelope in Fig. A.3. We observed that the AR (100) model satisfactorily describe the envelope's frequency spectrum, and passed all hypothesis tests. In section 2.3.2, we developed an envelope generating system using AR (100) model, where the parameters were obtained from the ACF which was obtained from inverse Fourier Transform of the mean power spectrum of real audio signals. Since the original envelope signal, such as Fig. A.3, is not available for this case, the hypothesis testing was not performed. Instead, we compared the power spectrum of the AR (100) model with the smoothed power spectrum in Fig. 2.4 in section 2.3.2, and it showed that the AR (100) power spectrum closely approximated the mean power spectrum. The parameters of AR (100) model is shown in Table A.2.

AR (1:20)	(21:40)	(41:60)	(61:80)	(81:100)
1.0400	-0.0009	-0.0005	-0.0001	0.0001
-0.0034	-0.0009	-0.0004	-0.0001	0.0001
-0.0020	-0.0008	-0.0004	-0.0001	0.0001
-0.0019	-0.0008	-0.0004	-0.0001	0.0001
-0.0016	-0.0008	-0.0004	-0.0001	0.0001
-0.0015	-0.0008	-0.0004	-0.0001	0.0001
-0.0014	-0.0007	-0.0004	-0.0001	0.0001
-0.0013	-0.0007	-0.0003	0.0000	0.0001
-0.0013	-0.0007	-0.0003	0.0000	0.0001
-0.0012	-0.0007	-0.0003	0.0000	0.0001
-0.0012	-0.0007	-0.0003	0.0000	0.0001
-0.0011	-0.0006	-0.0003	0.0000	0.0001
-0.0011	-0.0006	-0.0003	0.0000	0.0001
-0.0011	-0.0006	-0.0002	0.0000	0.0001
-0.0010	-0.0006	-0.0002	0.0000	0.0001
-0.0010	-0.0006	-0.0002	0.0000	0.0001
-0.0010	-0.0005	-0.0002	0.0000	0.0001
-0.0010	-0.0005	-0.0002	0.0000	0.0001
-0.0009	-0.0005	-0.0002	0.0001	0.0001
-0.0009	-0.0005	-0.0001	0.0001	-0.0022

Table A.2 Parameters of AR (100) model in Fig.2.4 in section 2.3.2.
 First column : AR(1:20),..., Last column : AR(81:100).

4. Maximum Likelihood Estimator program code

```
% OKLAHOMA STATE UNIVERSITY
% ELECTRICAL AND COMPUTER ENGINEERING
%
% MAXIMUM LIKELIHOOD ESTIMATOR for AR(p) case
% PROGRAMMER: JAE Y. LEE
% APR 11 1995
%
% SUBROUTINE :
%
% IFILTAR : CALCULATE a (= residual) from measurement y
%
%
% PRE : DATA y IS READ INTO PROGRAM. ASSUME THE ORDER OF y WAS
%       ESTIMATED BY GPAC.
%
% POST : THE PARAMETER OF y IS ESTIMATED
%
%-----

clear
clc

load y; % DATA READ
y=y(:);
k=length(y);
y=y-mean(y); % make y zeromean
y= y./std(y); % make variance of y =1

delta=input('Enter delta(typical value is: 0.001) : ');
mu=input('Enter mu (Typical value is :0.01) : ');
f2=10; % CONSTANT TO CONTROL MU WHEN ERROR IS DECREASING
mumax=10; % UPPER LIMIT OF MU
maxiter=input('Enter max iteration.. '); % MAX # OF ITERATION
eps=input('Enter epsilon....'); % ERROR THRESHOLD USED FOR TESTING
CONVERGENCE

% ESTIMATED ORDER OF AR SYSTEM OBTAINED BY INSPECTION OF GPAC
p=input('Enter AR order... (1,2,...) : ');
beta=zeros(1,p); % INITIAL GUESS OF PARAMETER
n1=1; % starting first iteration
a=zeros(k,p+1);

numiter=1;
```

```

while n1==1, % if error is decreasing n1=1
numiter    % tells current iteration number

a(:,1)=ifiltar(beta,y); % GET a vector (=residual) from y
beta1=zeros(1,p)';
beta1(1)=delta;
beta1=beta+beta1;

i=2;
while i < p+1,
a(:,i) = ifiltar(beta1,y);
clear beta1
beta1=zeros(p,1);
beta1(i) = 1 *delta;
beta1=beta+beta1;
i=i+1;
end

a(:,p+1)=ifiltar(beta1,y);

% Error matrix, each column is error of
% a vector from using total beta minus
% a vector from using each individual parameter of beta
%

for i=1:p,
x(:,i)=a(:,1)-a(:,i+1);
end

x = x ./delta;

aij=x' * x;
g=x' * a(:,1);
d=sqrt(diag(aij));
astar1=aij ./ (d * d');
astar2=eye(p)*mu;
astar=astar1+astar2;
gstar=g ./d;
hstar= astar \ gstar;
hj=hstar ./d;

betanew=beta+hj;

```

% Check conditions if error power of a is decreasing...

```
betaold=beta;
sold=a(:,1)' * a(:,1);
a(:,1)=ifiltar(betanew,y);
```

```
snew=a(:,1)' * a(:,1);
```

```
if snew < sold,
```

```
    if max(abs( hj)) < eps,
        fprintf('Converged estimated parameters\n');
        fprintf(' \n');
        fprintf(' -- AR parameters -- \n');
        fprintf(' \n');
        betahat=beta ;% ESTIMATED PARAMETERS
        ar=-betahat(1:p); % AR part

        fprintf('%5.4f\n', ar);
        fprintf(' \n');

        fprintf('Estimated variance of residual \n');
        fprintf(' \n');
        sigahat=snew /(k-p); % ESTIMATED VARIANCE OF RESIDUALS
        fprintf('%5.4f',sigahat);
        fprintf(' \n');
        covariance=sigahat*inv(aij ) % COVARIANCE MATRIX OF BETA
        break;
        return;
```

```
else,
```

```
    mu=mu /f2;
    numiter=numiter+1;
    beta=betanew;
    n1=1;
```

```
end % if abs
```

```
else,
```

```
    mu=mu*f2;
    numiter=numiter+1;
```

```

beta=betaold;
if mu > mumax,

    error('over mumax');
    betanew
    break;
    return;
end % if mu
if numiter > maxiter,
    error(' over maxiter');
    betanew
    break
    return;
end % if numiter
n1=1;
end    % if

end % while n1==1

%----- CALCULATE AUTOCORRELATION AND GPAC OF RESIDUAL-----

for tau=1:k/4,
ra(tau)=(1/(k-tau+1)) *a(1:k-tau+1,1)*a(tau:k,1);
end

ra= ra ./ra(1); % normalize

%----- HYPOTHESIS TESTING -----

fprintf('----- HYPOTHESIS TESTING ----- \n');
fprintf(' \n');

% CHI-SQUARE TESTING

q1=150*sum(ra(2:150) .^2)

% STANDARD ERROR TESTING ON RESIDUAL ACF

count1=0;
for i = 2:150,
sacf(i) = (1+2*sum(ra(1:i-1) .^2 ) ) .^5 ./ sqrt(k); % standard error

```

```

if abs(ra(i)) > 2 * sacf(i)
count1=count1+1;
end
end

```

% CHECK PACF OF RESIDUAL

```

pacf = rtophi(ra);
count2=0;
for i = 2:150,
spacf = 1/ sqrt(k); % standard error for pacf

```

```

if abs(ra(i)) > 2 * spacf
count2=count2+1;
end
end

```

```

fprintf('Number of lags exceeding 2 standard error range \n ')
fprintf(' \n');
fprintf('ACF : ');
fprintf('%2d\n',count1)
fprintf(' \n');
fprintf('PACF : ');
fprintf('%2d\n',count2)

```

```

subplot(2,1,1),plot(ra,'b');
grid;
title('Autocorrelation of residual');
subplot(2,1,2),plot(pacf,'b');
xlabel('lag of 0 - 5.7 msec')
grid;
title('Partial autocorrelation of residual');
whitebg

```

```
% FUNCTION IFILTAR.M
%
% This function calculates a (= residual) from y for AR(p) case
% Jae Y. Lee
% Apr 1995

function a1=ifiltar(beta,y);

a1=filter([1 beta(:,1)'],1,y);

return;
```

VITA

Jae Yong Lee

Candidate for the Degree of

Doctor of Philosophy

Thesis : STFT BASED ENVELOPE TRACKING HARMONIC GENERATOR
DESIGN WITH APPLICATION TO ENHANCING BAND LIMITED
AUDIO SIGNALS

Major Field : Electrical Engineering

Biographical :

Personal Data : Born in Osan, Korea, on Jan 22, 1961, the son of Sun-Hak Lee
and Boo-Young Ahn.

Education : Graduated from Yong Mun High School, Seoul, Korea in Feb 1980;
received Bachelor of Science degree in Electrical Engineering from the
Republic of Korea Air Force Academy, Seoul, Korea in April 1984;
received Master of Science degree in Electrical Engineering from
Oklahoma State University, Stillwater, Oklahoma in 1991. Completed
the requirements for the Doctor of Philosophy degree with a major in
Electrical Engineering at Oklahoma State University in May, 1998.

Experience : Worked as a pilot of search and rescue helicopter with the Korea
Air Force; employed by Oklahoma State University, Department of
Electrical and Computer Engineering as a teaching assistant to
graduate and undergraduate courses, 1994 -1997.