APPLICATION OF SIGNAL ESTIMATION FROM THE MODIFIED SHORT-TIME FOURIER TRANSFORM TO SPEECH PROCESSING

by

Douglas S. Deadrick

B.S.E.E., West Virginia University

(1982)

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 1984

Signature of Author _____

Department of Electrical Engineering and Computer Science
September 27, 1984

Certified by _____

Jae S. Lim
Thesis Supervisor

Accepted by _____

Arthur C. Smith
Chairman, Departmental Committee on Graduate Students

Application of Signal Estimation from the Modified Short-Time Fourier Transform to Speech Processing

by

Douglas S. Deadrick

Submitted to the Department of Electrical Engineering and Computer Science, December 1984, in partial fulfillment of the requirements for the degree of Master of Science.

ABSTRACT

Many short-time Fourier transform (STFT) based speech processing systems modify the STFT magnitude but not the phase. Since the resulting modified STFT is no longer the STFT of any actual signal, the processed speech must be estimated from the modified STFT. Three methods for performing such an estimation are considered here. These methods include the conventional overlap-add algorithm, a weighted overlap-add algorithm, and an iterative algorithm for signal estimation from only the STFT magnitude.

Experiments are conducted to test the performance of these three estimation methods. These experiments involve signal estimation from arbitrarily modified STFT functions. In addition, the three estimation methods are applied to the problems of noisy speech enhancement by spectral subtraction, helium speech enhancement, and time-scale modification of speech.

Two appendices are included to give more insight into the experimental results. First, a model of the STFT of speech is reviewed. Then, a discussion is given on the effects of certain STFT magnitude modifications upon the speech estimated using the overlap-add method.

Thesis Supervisor: Professor Jae S. Lim

Title: Associate Professor of Electrical Engineering

ACKNOWLEDGEMENTS

I would like to thank Professor Jae S. Lim for his supervision of this thesis. His guidance, insight, and special attention were invaluable to the completion of this work. In addition, the helpful comments received from Professor Bruce R. Musicus are greatly appreciated. I also wish to acknowledge the National Science Foundation, without whose financial support my graduate studies would not have been possible.

I wish to thank the members of the Digital Signal Processing Group at MIT for providing a stimulating environment in which to work and for maintaining a superb computing facility. A special thanks goes to Mr. Daniel W. Griffin for providing an excellent foundation from which to start this work.

Finally, words cannot fully express my gratitude to my wife, Robin, for generously providing me with her love, support, and typing skills throughout my tenure at MIT.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Problem Description

Many speech processing systems make use of the short-time Fourier transform (STFT) representation of speech. The effect of the STFT is to segment the speech into a series of frames, each corresponding to the spectral characteristic of an individual piece of the utterance. Modifications can then be made to each of the spectral frames of a given STFT to produce some desired STFT. Thus, it is possible to implement a modification process as a time-varying filter whose properties depend upon the individual parts of speech.

Most current STFT-based speech modification schemes apply only to the spectral magnitude of each frame, called the STFT magnitude. For example, consider the spectral subtraction method for enhancing noisy speech. This process subtracts an estimate of the noise spectral density from the short-time power spectrum of the noisy speech. The square root of the modified short-time power spectrum becomes the desired (modified) STFT magnitude.

In such STFT-based speech modification schemes it is either not known how to modify the STFT phase in a manner consistent with the magnitude modification, or the required STFT phase modification is difficult to implement. Thus, in this type of

speech processing, some estimate of a desirable STFT magnitude has been generated, but there is no new estimate of an appropriate STFT phase.

This situation presents some difficulties when the desired (modified) speech is to be synthesized from the modified STFT. In general, no signal corresponds exactly to the modified STFT magnitude. Thus, a signal must be estimated from the modified STFT, and the estimate should have a STFT magnitude which comes close in some sense to the desired magnitude.

A technique has been developed by Griffin and Lim [1] for estimating the desired signal directly from the modified STFT magnitude. This estimation method is capable of producing a signal whose STFT magnitude is as close as possible in the mean-square error sense to the desired magnitude. However, this approach requires quite a bit of computation and is rather inconvenient to implement on general purpose computers.

Alternatively, techniques such as the overlap-add method [2] and a type of weighted overlap-add method [1] exist for estimating the desired signal from the modified STFT magnitude combined with the unmodified STFT phase. These estimation methods are much faster and computationally less expensive than the method of Griffin and Lim. In fact, in most speech processing applications, the conventional approach to producing a signal estimate is to employ the overlap-add method using the modified magnitude and unmodified phase. However, in some applications, such as time-scale modification of speech, the overlap-add and weighted overlap-add methods fail to produce a signal whose STFT magnitude is very close to the desired magnitude.

Therefore, it is important to determine for some existing speech processing applications the occasions when the LSEE-MSTFTM algorithm gives the only useful performance. In addition, these results will provide some indication of the basic circumstances under which this algorithm is required. On the other hand, this information can also help to identify those cases where the computational savings of one of the overlap-add methods can be exploited in the signal estimation process without a significant loss of the accuracy available with the LSEE-MSTFTM algorithm.

1.2 Past Work

Initially, STFT-based speech processing systems always estimated the desired signal from the modified magnitude combined with the unmodified phase using the overlap-add method, and there was very little basis for determining the success or failure of each case. For example, Richards [3] considered enhancement of speech which is spoken in a pressurized helium atmosphere. He determined that the overlap-add method is acceptable in this case since the enhanced speech synthesized from the appropriately modified STFT magnitude and original STFT phase is intelligible with no obvious defects.

A type of weighted overlap-add method was developed [1] as a more theoretically accurate way to estimate the desired speech from the modified magnitude and unmodified phase. The signal produced from this weighted overlap-add method has the STFT which is as close as possible, in a mean-square error sense, to the modified STFT composed of the modified magnitude and unmodified phase. However, just as with the

standard overlap-add method, it is not desirable in some cases to estimate a signal based on the modified magnitude and unmodified phase. That is, it is the modified STFT magnitude in which we are the most interested, and we are generally unconcerned with the associated phase function. Hence, there are some results [4] which indicate that, in some applications, this method does not produce a significantly better estimate than the standard overlap-add technique.

The least-squares error estimation from modified STFT magnitude (LSEE-MSTFTM) algorithm [1] was developed as an accurate way to estimate the desired speech from the modified STFT magnitude, and it does so without regard to the associated phase function which it produces. The signal produced from the LSEE-MSTFTM algorithm has the STFT magnitude which is as close as possible, in a mean-square error sense, to the desired STFT magnitude.

In one study [5], the results of estimation using the overlap-add method and estimation using the LSEE-MSTFTM algorithm were compared for some specific applications. For the cases of speech enhancement by spectral subtraction and helium speech enhancement, some results indicate that there is no significant difference between the speech estimated from only the modified STFT magnitude, using LSEE-MSTFTM, and that estimated from the modified STFT magnitude and unmodified STFT phase using overlap-add.

1.3 Research Goals

There is much to be saved in terms of time and computational expense if a modified

speech signal can be estimated from the modified STFT magnitude and the unmodified phase (using one of the overlap-add methods) with as much accuracy as the speech signal estimated from only the modified magnitude (using the LSEE-MSTFTM algorithm). Thus, this work is intended to experimentally determine amongst some important speech processing applications, those cases where the above condition exists. The general procedure is to determine for each example the quality of the speech estimated using the two overlap-add methods and the LSEE-MSTFTM algorithm.

1.4 Organisation of the Thesis

In Chapter 2, definitions and terminology of the STFT analysis process will be established. Formulas will be given for the three signal estimation methods of interest: overlap-add, weighted overlap-add, and LSEE-MSTFTM.

In Chapter 3, the criteria are given by which the performance of each of the three estimation methods will be evaluated and compared. The quantitative measure will be the computation and comparison of signal-to-error ratios (SER) between the modified STFT magnitude and the STFT magnitudes of each of the three signal estimates. The qualitative measure will be informal listening tests among the three methods of signal estimation.

In the following four chapters, the algorithms of Chapter 2 will be applied to some examples of speech processing based on STFT magnitude modification. The applications include some arbitrary magnitude modifications. These modifications involve altering the STFT magnitude's envelope or pitch. The other applications are speech

enhancement by spectral subtraction, helium speech enhancement, and time-scale modification of speech. In each case the performance of the estimation algorithms will be evaluated using the methods of Chapter 3.

The results of the presented experiments will be summarized in Chapter 8, and their general applications and limitations will be discussed. In addition, two appendices are included to help give some insight into the results obtained here and the issues they raise. Appendix A summarizes a model of the STFT of voiced speech which was developed by Portnoff [6]. Then this model is used in Appendix B as a basis of overlap-add signal estimation from a magnitude modified STFT. These observations will help to support the experimental results which show the occasions when the accuracy of the LSEE-MSTFTM algorithm is required.

CHAPTER 2

BACKGROUND

2.1 Definition of the STFT

The slowly time-varying characteristics of speech allow us to usefully analyze the spectra of segments of the speech signal. The short-time Fourier transform [2] of signal $x(n)$ can be defined as:

$$X_{\mathrm{w}}(mS,\omega) \equiv \sum_{n=-\infty}^{\infty} x_{\mathrm{w}}(mS,n)e^{-j\omega n} \tag{2.1a}$$

where

$$x_{\mathrm{w}}(mS,n) \equiv \mathrm{w}(mS-n)x(n) \tag{2.1b}$$

and $\mathrm{w}(n)$ is some time domain window function. For all of the analyses performed in this chapter, we will assume that $\mathrm{w}(n)$ is approximately time limited, band-limited, and symmetric about the time origin (zero-phase).

The STFT magnitude is:

$$|X_{\mathrm{w}}(mS,\omega)| = \sqrt{Re^2\{X_{\mathrm{w}}(mS,\omega)\} + Im^2\{X_{\mathrm{w}}(mS,\omega)\}} \tag{2.2}$$

and the tangent of the STFT phase is:

$$\tan\left[\arg[X_{\mathrm{w}}(mS,\omega)]\right] = \frac{Im\{X_{\mathrm{w}}(mS,\omega)\}}{Re\{X_{\mathrm{w}}(mS,\omega)\}}. \tag{2.3}$$

2.2 Signal Reconstruction from the STFT

When a STFT magnitude, $|X_w(mS,\omega)|$, and phase, $\arg[X_w(mS,\omega)]$, are available and have been computed from the same signal, that signal, $x(n)$, can be reconstructed using the overlap-add procedure [2],

$$x(n) = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} X_w(mS,\omega) e^{j\omega n} d\omega \qquad (2.4a)$$

where

$$X_w(mS,\omega) = |X_w(mS,\omega)| e^{j \cdot \arg[X_w(mS,\omega)]}. \qquad (2.4b)$$

Although the overlap-add procedure is valid only for signal reconstruction from the unmodified STFT, this method is frequently used to attempt signal estimation from a modified STFT. In the examples that we are considering, the modified STFT, $\tilde{X}_w(mS,\omega)$, is the modified STFT magnitude, $|\tilde{X}_w(mS,\omega)|$, combined with the unmodified STFT phase, $\arg[X_w(mS,\omega)]$. The estimation procedure is to substitute the desired modified STFT, $\tilde{X}_w(mS,\omega)$, for $X_w(mS,\omega)$ in Equation (2.4). This will produce a signal estimate, $y(n)$.

2.3 Signal Estimation from Modified STFT

A method has been developed for estimating a signal from the modified STFT which is more accurate than the overlap-add method for estimation based on the magnitude *and* phase. This method [1] has the form of a weighted overlap-add procedure, and it is given by the following:

$$y(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n) \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \tilde{X}_w(mS,\omega) e^{j\omega n} d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} \qquad (2.5)$$

where $\tilde{X}_w(mS,\omega)$ is the modified STFT discussed in Section 2.2. This method has been shown [4] to minimize the following distance measure:

$$D\{y(n),\tilde{X}_w(mS,\omega)\} \equiv \sum_{m=-\infty}^{\infty} \frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} |Y_w(mS,\omega)-\tilde{X}_w(mS,\omega)|^2 d\omega. \qquad (2.6)$$

That is, this estimation technique produces the signal whose STFT comes the closest, in a mean square error sense, to the modified STFT.

2.4 Signal Estimation from STFT Magnitude

The Least Squares Error Estimation from Modified STFT Magnitude (LSEE-MSTFTM) algorithm is an iterative procedure which decreases with each iteration the mean square error between the STFT magnitude of the signal estimate and a given STFT magnitude. As such, this algorithm permits estimation of a signal from only a prescribed STFT magnitude. The algorithm has the following iterative form:

$$y^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n)\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}\hat{X}_w^i(mS,\omega)e^{j\omega n}d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} \qquad (2.7a)$$

where

$$\hat{X}_w^i(mS,\omega) \equiv |\tilde{X}_w(mS,\omega)|\frac{Y_w^i(mS,\omega)}{|Y_w^i(mS,\omega)|} \qquad (2.7b)$$

and again $|\tilde{X}_w(mS,\omega)|$ is the desired STFT magnitude. In accordance with its definition, this method has been shown [4] to reduce the distance measure

$$D_M\{y^i(n),|\tilde{X}_w(mS,\omega)|\} \equiv \sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}\left[|Y_w^i(mS,\omega)|-|\tilde{X}_w(mS,\omega)|\right]^2 d\omega \quad (2.8)$$

for increasing i. That is, if we define $D_M^i = D_M\{y^i(n), |\ddot{X}_w(mS,\omega)|\}$, then

$$D_M^i \geq D_M^{i+1} \quad \text{for} \quad i \geq 1. \tag{2.9}$$

For all the speech signals produced by the LSEE-MSTFTM algorithm in this thesis, Equation (2.7) is iterated until D_M^i converges to a limit point. Convergence will be defined as the point where

$$\frac{D_M^{i-1} - D_M^i}{D_M^i} < \Delta.$$

For all of the experiments in this thesis, a convenient choice of Δ is 0.01, and the above convergence condition is easily satisfied after 50 iterations.

2.5 Summary

The three signal estimation methods presented here are the ones which will be used and compared in the upcoming experiments. Each algorithm has its own significance in the speech processing applications. The overlap-add method has historically been the conventional approach to signal estimation. The given version of the weighted overlap-add method is the best approach for trying to simultaneously meet the modified STFT magnitude specifications and the unmodified phase specifications. The LSEE-MSTFTM algorithm is the best signal estimation approach for trying to meet the modified STFT magnitude specifications regardless of the STFT phase function which ultimately results.

CHAPTER 3

EVALUATION OF ALGORITHM PERFORMANCE

3.1 Introduction

Given in this chapter are the bases by which the performance of each of the three signal estimation methods will be evaluated. A signal-to-error ratio formula is the heart of the evaluation process. This formula will be applied to the three estimation methods for a common numerical comparison. Also given in this chapter is the general experimental procedure used in the application chapters.

3.2 Performance Measures

Usually, the goal of STFT-based speech processing schemes is to estimate a signal whose spectral characteristics are close in some way to a desired STFT magnitude without regard to the accompanying STFT phase. Therefore, a natural measure of an estimator's performance is the mean square error between the STFT magnitude of the estimated signal and the desired STFT magnitude. This is the basis of the following numerical performance measure, the signal to error ratio (SER):

$$SER = 10\log\left\{\frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|\tilde{X}_w(mS,\omega)|^2 d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}\left(|Y_w(mS,\omega)|-|\tilde{X}_w(mS,\omega)|\right)^2 d\omega}\right\}. \qquad (3.1)$$

Here $|\tilde{X}_w(mS,\omega)|$ is the modified STFT magnitude and $|Y_w(mS,\omega)|$ is the STFT magnitude of the signal estimate. This measure is applied identically to each of the three signal estimation methods, and the explicit formulas are given next.

First, to evaluate the performance of the overlap-add method, the SER is computed as follows:

$$SER_{OA} = 10\log\left\{\frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|\tilde{X}_w(mS,\omega)|^2 d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}\left(|Y_{OA_w}(mS,\omega)|\cdot N_{OA}-|\tilde{X}_w(mS,\omega)|\right)^2 d\omega}\right\}. \tag{3.2}$$

The subscript OA, meaning overlap-add, is used in the above formula to indicate that those quantities are computed in reference to the overlap-add signal estimate. Accordingly, $Y_{OA_w}(mS,\omega)$ is the STFT of the speech estimated from the modified magnitude and unmodified phase using overlap-add. The SER has the advantage of being independent of signal energy. The normalization constant, N_{OA}, is employed to reduce any contribution to the error due to scale differences between the two STFT magnitudes. N_{OA} is computed by minimizing the error portion of the SER giving

$$N_{OA} = \frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|\tilde{X}_w(mS,\omega)||Y_{OA_w}(mS,\omega)|d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|Y_{OA_w}(mS,\omega)|^2 d\omega}. \tag{3.3}$$

Second, to evaluate the performance of the weighted overlap-add method, the specific formula for the SER becomes:

$$SER_{WO} = 10\log\left\{\frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|\tilde{X}_w(mS,\omega)|^2 d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}\left(|Y_{WO_w}(mS,\omega)|\cdot N_{WO}-|\tilde{X}_w(mS,\omega)|\right)^2 d\omega}\right\} \tag{3.4}$$

where, again, minimization of the error portion of the SER gives

$$N_{WO} = \frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|\tilde{X}_w(mS,\omega)||Y_{WO_w}(mS,\omega)|d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|Y_{WO_w}(mS,\omega)|^2 d\omega}. \tag{3.5}$$

The subscript WO, meaning weighted overlap-add, is used in the above two formulas to indicate that those quantities are computed in reference to the weighted overlap-add signal estimate. Accordingly, $Y_{WO_w}(mS,\omega)$ is the STFT of the speech estimated from the modified magnitude and unmodified phase using the weighted overlap-add method.

The third SER, used to evaluate the performance of the LSEE-MSTFTM algorithm, is computed as follows:

$$SER_{MO} = 10\log\left\{\frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|\tilde{X}_w(mS,\omega)|^2 d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}\left(|Y_{MO_w}(mS,\omega)|\cdot N_{MO}-|\tilde{X}_w(mS,\omega)|\right)^2 d\omega}\right\} \tag{3.6}$$

where, again, minimization of the error portion of the SER gives

$$N_{MO} = \frac{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|\tilde{X}_w(mS,\omega)||Y_{MO_w}(mS,\omega)|d\omega}{\sum_{m=-\infty}^{\infty}\frac{1}{2\pi}\int_{\omega=-\pi}^{\pi}|Y_{MO_w}(mS,\omega)|^2 d\omega}. \tag{3.7}$$

The subscript MO, meaning magnitude-only, is used in the above two formulas to indicate that those quantities are computed in reference to the LSEE-MSTFTM signal estimate which is a magnitude-only method. Accordingly, $Y_{MO_w}(mS,\omega)$ is the STFT of the speech estimated from only the modified magnitude using the LSEE-MSTFTM algorithm.

Some important points should be considered when interpreting SER results. First, SER_{MO} is related to how close the STFT magnitude of the signal estimated using the LSEE-MSTFTM algorithm can get to the desired modified STFT magnitude. In this sense, SER_{MO} represents a fundamental limit to the accuracy of an obtainable signal estimate. In addition, the difference between SER_{MO} and SER_{OA} or SER_{WO}

indicates the amount of improvement which can be expected from using the LSEE-MSTFTM algorithm instead of the overlap-add method or the weighted overlap-add method respectively. Therefore, SER_{MO} will always be greater than or equal to SER_{OA} and SER_{WO}. However, the closer SER_{OA} and SER_{WO} are to SER_{MO}, the more useful one of those overlap-add methods become for estimation of the signal. The difference between SER_{MO} and SER_{WO} will illustrate the extent to which the unmodified phase information has helped or hindered the performance of the weighted overlap-add method.

Finally, informal listening tests will be conducted between the speech estimated using overlap-add, $y_{OA}(n)$, the speech estimated using weighted overlap-add, $y_{WO}(n)$, and the speech estimated from only the magnitude, $y_{MO}(n)$, using the LSEE-MSTFTM algorithm. The informal listening tests will serve as a qualitative judgement of algorithm performance.

3.3 Experimental Procedure

The following general procedure is used in the experiments. A STFT, $X_w(mS,\omega)$, is computed from the given speech sentence. The magnitude of this STFT is modified giving the desired magnitude $|\tilde{X}_w(mS,\omega)|$. Then the speech is estimated from the desired STFT magnitude in three ways. In the first method, the desired magnitude is combined with the unmodified phase as follows:

$$\tilde{X}_w(mS,\omega) = |\tilde{X}_w(mS,\omega)| \frac{X_w(mS,\omega)}{|X_w(mS,\omega)|}. \tag{3.8}$$

Speech will be estimated from the modified STFT of Equation (3.8) using the overlap-add method, and SER_{OA} will be computed. In the second method, speech will be estimated from the modified STFT of Equation (3.8) using the weighted overlap-add method, and SER_{WO} will be computed. In the third method, speech will be estimated from the modified STFT magnitude, $|\tilde{X}_w(mS, \omega)|$, using the LSEE-MSTFTM algorithm, and SER_{MO} will be computed. Finally, an informal listening test is conducted between the three estimated speech sentences. This procedure is summarized in Figure 3.1.

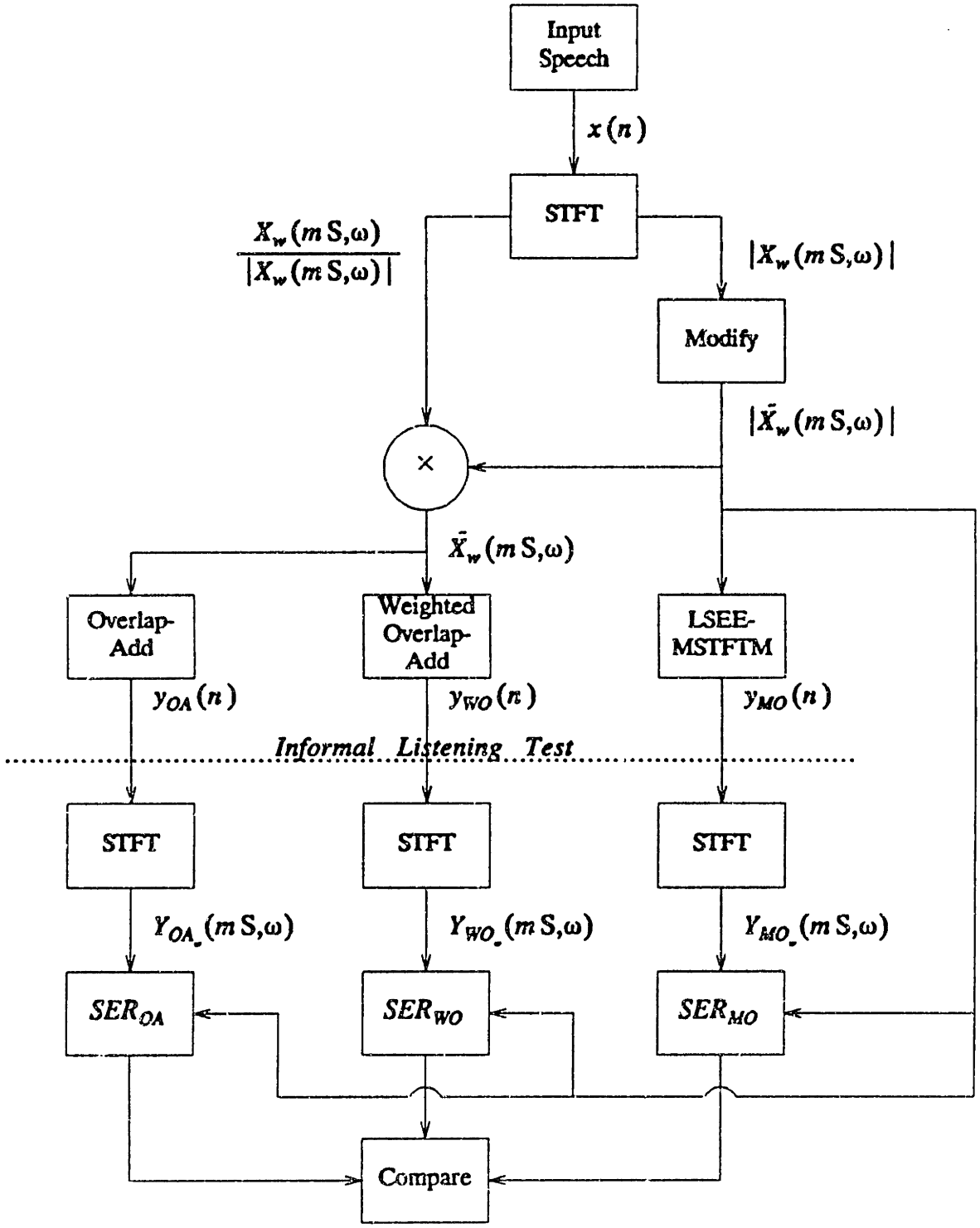


Figure 3.1. Procedure for evaluating performance of algorithms which estimate signals from a modified STFT.

CHAPTER 4

APPLICATION TO ARBITRARY MAGNITUDE MODIFICATIONS

4.1 Introduction

In this chapter, we shall implement some ad-hoc modifications on some STFT magnitudes to test the performance of the three signal estimation algorithms. The modifications are made to either the smooth component of the STFT magnitude, called the spectral envelope, or they are made to the harmonic structure of the STFT magnitude, which corresponds to the pitch of the voiced sections of the speech. The results are useful for predicting the performance of the three estimation methods in applications not covered in this thesis. In addition, the results lend support to some theoretical results on estimation from modified STFTs made in Appendix B.

4.2 Experimental Procedure and Results

For the experiments in this chapter, the speech signal, $x(n)$, is considered to be the result of a convolution of an excitation signal, $\psi(n)$, with a vocal tract impulse response, $t(n)$. The excitation signal corresponds to a quasi-periodic unit-sample train, v(n), or stationary white noise, $u(n)$, depending respectively upon whether the particular portion of the utterance is voiced or unvoiced. The STFT of the vocal tract impulse response is the spectral envelope $T(mS, \omega)$. The STFT of the excitation signal is

$\Psi(mS,\omega)$, which is the harmonic spectrum $V(mS,\omega)$ for voiced speech. Thus, with these assumptions,

$$X_w(mS,\omega) = T(mS,\omega) \cdot \Psi(mS,\omega). \tag{4.1}$$

4.2.1 Various envelope modifications

Five different multiplicative envelope modifications have been chosen to test the performance of the three estimation algorithms presented in Chapter 2. These modifications were not chosen with any specific application in mind, but merely as functions for achieving a wide range of examples.

Eight test sentences were used for the experiments performed here. Four were spoken by males and four were spoken by females. Each sentence was sampled at a rate of 10KHz. For each, the STFT was computed with a 256 point Hamming window, a window shift of 64 points, and 512 point DFTs.

For each STFT magnitude, the spectral envelope, $|T(mS,\omega)|$, was separated from the excitation spectrum, $|\Psi(mS,\omega)|$, using a very high quality pitch detection system [8]. The envelope frames were all multiplied by the given function, $M(\omega)$, and the modified envelope, $|\tilde{T}(mS,\omega)|$, was combined with the original harmonic spectrum, giving

$$|\tilde{X}_w(mS,\omega)| = |\tilde{T}(mS,\omega)| \cdot |\Psi(mS,\omega)|. \tag{4.2}$$

The functions by which the envelopes were multiplied are shown in Figure 4.1. The estimation experiments followed the same general procedure as given in Figure

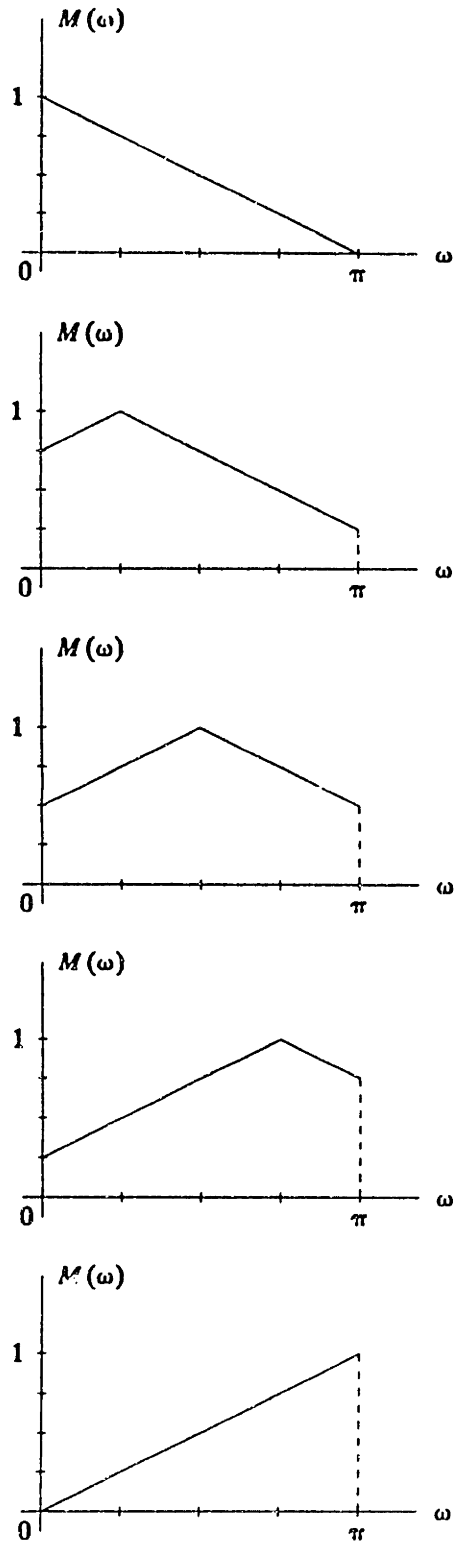


Figure 4.1. Functions by which the STFT magnitude envelopes were multiplied in the envelope modification experiments.

3.1. The results of the estimation experiments, are presented in Tables 4.2 through 4.5 for the male speech, and Tables 4.6 through 4.9 for the female speech. Table 4.1 contains the average SER values computed from Tables 4.2—4.9.

It can be seen from Tables 4.1—4.9 that signal estimation using the LSEE-MSTFTM algorithm has resulted in very little improvement over the other two methods. The differences in SER values range from a 2 dB to a 3 dB increase, which has been found to be insignificant for improvement of quality or intelligibility. Accordingly, in the informal listening tests, no audible difference could be detected among the speech sentences estimated using the three methods.

Also noteworthy are the specific values which were computed for SER_{MO} in these examples. In general, the numbers are high, indicating that the LSEE-MSTFTM algorithm produced estimates of speech signals whose STFT magnitudes come very close in the mean-square sense to the modified STFT magnitudes. The lowest SER_{MO} value is for magnitude modification example 5. This example performs a high-pass filtering operation on speech, and speech energy is typically concentrated in the low frequency bands. Therefore, this example has the lowest SER_{MO} value since it represents the most severe modification to the speech STFT magnitude.

4.2.2 Various pitch modifications

Seven different pitch modifications have been chosen to further investigate the estimation methods of Chapter 2. Again, there were no specific applications in mind with these modifications. The modifications were simply chosen to achieve a wide range of

Table 4.1 - Envelope Modification Average Results				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.3 dB	45.2 dB	47.3 dB	Estimated signals are indistinguishable from each other. All sound good.
2	46.1 dB	46.0 dB	48.6 dB	Same as above
3	43.4 dB	43.4 dB	46.1 dB	Same as above
4	38.7 dB	38.7 dB	41.3 dB	Same as above
5	30.1 dB	30.1 dB	32.7 dB	Same as above

Table 4.2 - Envelope Modification Results for Male Speech "Line up at the screen door."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.6 dB	45.5 dB	47.4 dB	Estimated signals are indistinguishable from each other. All sound good.
2	46.1 dB	46.0 dB	48.3 dB	Same as above
3	43.1 dB	43.2 dB	45.5 dB	Same as above
4	38.1 dB	38.1 dB	41.2 dB	Same as above
5	28.7 dB	28.7 dB	31.5 dB	Same as above

Table 4.3 - Envelope Modification Results for Male Speech "He has the bluest eyes."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.4 dB	45.4 dB	47.0 dB	Estimated signals are indistinguishable from each other. All sound good.
2	46.0 dB	45.9 dB	47.9 dB	Same as above
3	43.1 dB	43.1 dB	45.2 dB	Same as above
4	38.1 dB	38.1 dB	40.1 dB	Same as above
5	29.5 dB	29.5 dB	31.6 dB	Same as above

Table 4.4 - Envelope Modification Results for Male Speech "The chef made lots of stew."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.8 dB	45.7 dB	47.8 dB	Estimated signals are indistinguishable from each other. All sound good.
2	46.1 dB	46.0 dB	48.8 dB	Same as above
3	43.2 dB	43.1 dB	45.9 dB	Same as above
4	37.9 dB	37.9 dB	40.5 dB	Same as above
5	28.4 dB	28.4 dB	31.1 dB	Same as above

Table 4.5 - Envelope Modification Results for Male Speech "You're the biggest man."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.8 dB	45.6 dB	47.5 dB	Estimated signals are indistinguishable from each other. All sound good.
2	45.9 dB	45.8 dB	48.2 dB	Same as above
3	42.7 dB	42.7 dB	45.0 dB	Same as above
4	37.3 dB	37.3 dB	39.5 dB	Same as above
5	27.3 dB	27.3 dB	29.6 dB	Same as above

Table 4.6 - Envelope Modification Results for Female Speech "The bowl dropped from his hand."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.5 dB	45.4 dB	47.9 dB	Estimated signals are indistinguishable from each other. All sound good.
2	46.3 dB	46.2 dB	49.3 dB	Same as above
3	43.5 dB	43.5 dB	46.7 dB	Same as above
4	38.6 dB	38.6 dB	41.6 dB	Same as above
5	29.6 dB	29.6 dB	32.6 dB	Same as above

Table 4.7 - Envelope Modification Results for Female Speech "That shirt seems much to long."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.1 dB	45.1 dB	47.6 dB	Estimated signals are indistinguishable from each other. All sound good.
2	46.1 dB	46.0 dB	49.4 dB	Same as above
3	43.4 dB	43.4 dB	46.9 dB	Same as above
4	38.7 dB	38.7 dB	41.8 dB	Same as above
5	30.5 dB	30.5 dB	33.7 dB	Same as above

Table 4.8 - Envelope Modification Results for Female Speech "We made some fine brownies."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	45.2 dB	45.1 dB	47.1 dB	Estimated signals are indistinguishable from each other. All sound good.
2	46.0 dB	45.9 dB	48.4 dB	Same as above
3	43.2 dB	43.1 dB	45.7 dB	Same as above
4	38.2 dB	38.2 dB	40.5 dB	Same as above
5	29.0 dB	29.0 dB	31.4 dB	Same as above

Table 4.9 - Envelope Modification Results for Female Speech "The yellow lion roared."				
Envelope Modification	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	43.7 dB	43.8 dB	46.1 dB	Estimated signals are indistinguishable from each other. All sound good.
2	45.9 dB	46.0 dB	48.7 dB	Same as above
3	45.2 dB	45.3 dB	48.2 dB	Same as above
4	42.3 dB	42.4 dB	45.0 dB	Same as above
5	37.8 dB	37.8 dB	40.4 dB	Same as above

examples. These experiments used the same test sentences, STFT computation details, and pitch detection system as in Section 4.2.1.

The modifications involved increasing the fundamental frequency of the harmonic spectrum of the voiced portions, $|V(mS,\omega)|$, by some specific amount. This modified harmonic spectrum, $|\tilde{V}(mS,\omega)|$, was then combined with the original STFT spectral envelope, $|T(mS,\omega)|$, as follows:

$$|\tilde{X}_w(mS,\omega)| = |\tilde{V}(mS,\omega)| \cdot |T(mS,\omega)|. \tag{4.3}$$

These estimation experiments followed the same general procedure as given in Figure 3.1. The amounts of pitch increase, along with the results of these estimation experiments, are presented in Tables 4.11 through 4.18. Table 4.10 contains the average SER values computed from Tables 4.11—4.18.

It can be seen from Tables 4.10—4.18 that estimation using the LSEE-MSTFTM algorithm has, in certain cases, resulted in quite a dramatic improvement over the other two methods. The differences in SER values range from a 3 dB to a 15 dB increase, the latter of which corresponds to an overwhelming improvement of quality and intelligibility. There was not generally a significant difference between the SER values of the two overlap-add methods. In the informal listening tests, the audible difference between the speech sentences estimated using the LSEE-MSTFTM algorithm and the speech sentences estimated using the two overlap-add methods became very distinct as the factor by which the pitch was increased approached 50%. The audible difference was a "multiple pitch" quality in the speech estimated using the overlap-add

Table 4.10 - Pitch Modification Average SER Values				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	14.0 dB	14.3 dB	16.2 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	6.2 dB	6.3 dB	15.5 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	4.8 dB	4.6 dB	15.7 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	5.4 dB	5.2 dB	14.6 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	6.8 dB	6.5 dB	17.6 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	6.2 dB	6.1 dB	15.1 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	14.3 dB	14.1 dB	16.8 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.11 - Pitch Modification Results for Male Speech "Line up at the screen door."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	13.8 dB	14.1 dB	16.5 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	6.3 dB	6.6 dB	14.8 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	4.5 dB	4.6 dB	14.8 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	3.8 dB	3.7 dB	16.3 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	3.4 dB	3.3 dB	18.0 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	5.9 dB	5.9 dB	14.1 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	14.8 dB	15.3 dB	17.7 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.12 - Pitch Modification Results for Male Speech "He has the bluest eyes."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	13.7 dB	14.1 dB	16.0 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	6.3 dB	6.5 dB	15.0 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	4.6 dB	4.6 dB	14.5 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	3.6 dB	3.4 dB	16.9 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	4.6 dB	4.3 dB	18.4 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	5.7 dB	5.6 dB	15.0 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	14.6 dB	15.2 dB	17.7 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.13 - Pitch Modification Results for Male Speech "The chef made lot's of stew."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	14.2 dB	14.8 dB	16.2 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	7.8 dB	8.2 dB	15.1 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	5.2 dB	5.7 dB	14.2 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	5.2 dB	5.4 dB	16.6 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	4.7 dB	4.7 dB	16.6 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	7.2 dB	7.2 dB	14.2 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	13.6 dB	14.2 dB	16.0 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.14 - Pitch Modification Results for Male Speech				
"You're the biggest man."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	14.2 dB	14.9 dB	16.3 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	8.7 dB	8.9 dB	14.4 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	5.5 dB	5.8 dB	16.8 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	4.5 dB	4.3 dB	17.2 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	4.5 dB	4.3 dB	17.2 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	8.5 dB	8.5 dB	15.5 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	14.7 dB	15.7 dB	17.2 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.15 - Pitch Modification Results for Female Speech				
"The bowl dropped from his hand."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	14.9 dB	15.0 dB	18.5 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	5.2 dB	5.1 dB	17.3 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	4.2 dB	3.7 dB	17.3 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	5.7 dB	6.5 dB	17.9 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	9.1 dB	8.6 dB	18.3 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	5.7 dB	5.5 dB	16.8 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	16.1 dB	16.4 dB	16.8 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.16 - Pitch Modification Results for Female Speech				
"That shirt seems much too long."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	14.9 dB	15.0 dB	18.4 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	5.0 dB	5.0 dB	15.9 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	3.7 dB	3.3 dB	18.0 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	5.7 dB	5.2 dB	14.9 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	9.0 dB	8.9 dB	16.2 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	5.5 dB	5.4 dB	15.4 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	14.7 dB	15.1 dB	17.0 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.17 - Pitch Modification Results for Female Speech				
"We made some fine brownies."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	13.7 dB	13.9 dB	14.2 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	5.5 dB	5.4 dB	17.1 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	4.8 dB	4.2 dB	17.6 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	6.6 dB	5.8 dB	19.1 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	9.1 dB	8.6 dB	19.6 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	5.2 dB	5.0 dB	17.0 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	15.6 dB	10.0 dB	18.5 dB	Estimated signals are indistinguishable from each other. All sound good.

Table 4.18 - Pitch Modification Results for Female Speech				
"The yellow lion roared."				
Pitch Increase	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1%	12.8 dB	12.7 dB	13.1 dB	Estimated signals are indistinguishable from each other. All sound good.
5%	4.9 dB	4.7 dB	14.4 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
10%	5.6 dB	5.0 dB	15.7 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
33%	6.6 dB	5.9 dB	14.3 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
50%	9.7 dB	9.5 dB	16.2 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
90%	5.8 dB	5.3 dB	12.8 dB	Slight multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
99%	10.3 dB	10.6 dB	12.5 dB	Estimated signals are indistinguishable from each other. All sound good.

methods. The speech produced by the LSEE-MSTFTM algorithm was identical to the unmodified speech except that it sounded as though it were spoken by higher pitched individuals. There was no multiple pitch quality or any other degradation audible in the speech estimated using the LSEE-MSTFTM algorithm. There was no audible difference between the speech estimated using overlap-add and the speech estimated using weighted overlap-add.

On the other hand, the audible difference between the speech sentences estimated using the LSEE-MSTFTM algorithm and the speech estimated using the two overlap-add methods became less distinct as the factor by which the pitch was increased approached 100%. In general, there was no audible difference between the speech estimated using overlap-add and the speech estimated using weighted overlap-add.

Also notice the specific values which were computed for SER_{MO} in these examples. The numbers are not as high as the ones computed in Tables 4.1—4.8. This indicates that pitch modifications are more severe than envelope modifications. This means that the modified STFT magnitudes for these cases are somewhat more distant from the STFT magnitude of any actual signal. Hence, for pitch modifications, the LSEE-MSTFTM algorithm converges upon signals whose STFT magnitudes are more distant from the modified STFT magnitudes for the given window length and DFT size.

4.3 Summary

In this chapter we have tested the performance of the three methods of signal estimation from the modified STFT. For spectral envelope modifications, it was found that the

speech estimated using the LSEE-MSTFTM algorithm was indistinguishable from the speech estimated using the two overlap-add methods. For pitch modifications, it was determined that the performances of the two overlap-add methods became poor for pitch modifications further from an integer multiple of the original pitch. A "multiple pitch" quality exists for each example of pitch modified speech which was estimated using the two overlap-add methods. However, the listening test results indicate that the multiple pitches become discernible when the desired pitch is within approximately 10% above or below an integer multiple of the original pitch, and they are strongest when the desired pitch is midway between integer multiples of the original pitch. No such multiple pitch quality was present in the speech estimated using the LSEE-MSTFTM algorithm, and these estimates indeed sounded as though they were spoken by higher pitched individuals.

The experimental results given here indicate some circumstances under which the overlap-add methods are useful for signal estimation from the modified STFT. Specifically, if the desired STFT magnitude is an envelope modification of the original STFT magnitude, then the desired signal can be estimated quite well by combining the modified magnitude with the unmodified phase and using one of the overlap-add methods. Alternatively, if the desired STFT magnitude is a pitch modification of the original STFT magnitude, then the desired signal cannot be adequately estimated by one of the overlap-add methods. In such a case, the only useful signal estimation method is the LSEE-MSTFTM algorithm since it employs only the desired STFT magnitude.

These experimental results agree with the theoretical results given in Appendix B on modified STFT magnitude consistency with the unmodified phase. The theoretical results state that an envelope modified STFT magnitude remains consistent with the unmodified STFT phase, making either of the overlap-add methods just as suitable for estimating the signal as the LSEE-MSTFTM method. On the other hand, the theoretical results state that a pitch modified STFT magnitude does not generally remain consistent with the unmodified STFT phase, making the LSEE-MSTFTM algorithm the only useful method for estimating the signal. The theoretical results of Appendix B can also be compared to the experimental results of Chapters 5, 6, and 7, and this comparison is discussed in section B.3.

CHAPTER 5

APPLICATION TO SPECTRAL SUBTRACTION

5.1 Introduction

The estimation methods of Chapter 2 are now applied to some typical problems of speech processing. In this chapter, we consider the enhancement of single speaker speech degraded by additive noise. The technique of spectral subtraction [8] is used to estimate an enhanced STFT magnitude from the STFT magnitude of noisy speech by subtracting an estimate of the noise power spectrum from the noisy speech power spectrum.

The following sections illustrate how the three signal estimation algorithms can be used in this enhancement problem. We begin with the spectral subtraction approach to modifying a noisy STFT magnitude. Then, experiments are performed to evaluate the performance of each of the algorithms for the case of spectral subtraction.

5.2 STFT Modifications for Spectral Subtraction

A speech signal, $x(n)$, when combined with additive white noise, $d(n)$, has the following form:

$$r(n) = x(n) + d(n) \tag{5.1}$$

where $r(n)$ is the resulting noisy speech. In this application, we define the signal-to-noise ratio (SNR) as

$$SNR = 10\log\left\{\frac{\sum_{n=-\infty}^{\infty} x^2(n)}{\sum_{n=-\infty}^{\infty} d^2(n)}\right\}. \tag{5.2}$$

From Equation (5.1), the STFT of the noisy speech is

$$R_w(mS,\omega) = X_w(mS,\omega) + D_w(mS,\omega). \tag{5.3}$$

The resulting power spectrum is

$$|R_w(mS,\omega)|^2 = |X_w(mS,\omega)|^2 + |D_w(mS,\omega)|^2 + 2Re\left[X_w(mS,\omega)D_w^*(mS,\omega)\right] \tag{5.4}$$

which means that for the clean speech we have

$$|X_w(mS,\omega)|^2 = |R_w(mS,\omega)|^2 - |D_w(mS,\omega)|^2 - 2Re\left[X_w(mS,\omega)D_w^*(mS,\omega)\right]. \tag{5.5}$$

Spectral subtraction obtains its estimate of the enhanced STFT magnitude, $|\tilde{X}_w(mS,\omega)|$, by replacing the unknown quantities in Equation (5.5) with their expected values. This gives us the following estimation rule,

$$|\tilde{X}_w(mS,\omega)|^2 = |R_w(mS,\omega)|^2 - E\left[|D_w(mS,\omega)|^2\right] \tag{5.6}$$

where $E[\cdot]$ is the expectation operator.

However, this estimate has a large variance [8] and can also be negative. Thus, the actual form of our magnitude modification becomes [8]:

$$|\tilde{X}_w(mS,\omega)| = \begin{cases} \sqrt{|R_w(mS,\omega)|^2 - kE\left[|D_w(mS,\omega)|^2\right]} \\ \qquad \text{for} \quad kE\left[|D_w(mS,\omega)|^2\right] \le |R_w(mS,\omega)|^2 \\ 0 \qquad \text{otherwise.} \end{cases} \tag{5.7}$$

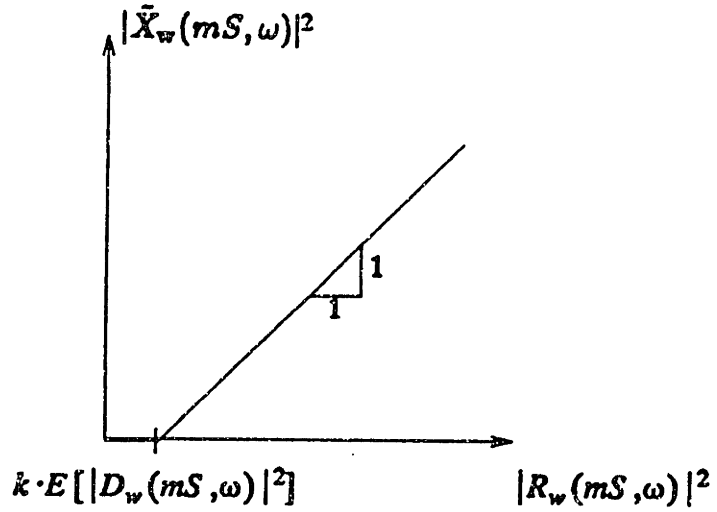


Figure 5.1. A piecewise-linear representation of spectral subtraction.

The parameter k is chosen to be greater than one to help reduce the variance of the estimate. A piecewise-linear representation of the input-output relationship in Equation (5.7) is shown in Figure 5.1.

5.3 Experimental Procedure and Results

Eight test sentences were used for the experiments performed here. Four were spoken by males, and four were spoken by females. For each sentence, enough stationary, white Gaussian noise was added to produce SNRs of 100, 10, 5, 0, and -5 dB. All sentences were sampled at a 10 KHz rate. The STFTs were computed using 256 point Hamming windows with a shift of 64 points, and the DFT size was 512 points.

The STFT magnitudes were modified according to Equation (5.7) with $k = 4$. The algorithm performance was evaluated for each case by comparing the results of the overlap-add, weighted overlap-add, and the LSEE-MSTFTM estimation methods. This

is the same procedure as given in Figure 3.1. The results of listening comparisons and numeric comparisons (through Equations (3.2), (3.4), and (3.6)) are given in Tables 5.2 through 5.9. Table 5.1 contains the average SER values computed from Tables 5.2–5.9.

It can be seen from Tables 5.1–5.9 that signal estimation using the LSEE-MSTFTM algorithm has resulted in very little improvement over estimation using the two overlap-add methods. The differences in SER values range from no difference to a 2 dB increase, which has been found to be insignificant for improvement of quality or intelligibility. Accordingly, in the informal listening tests, no audible difference could be detected between the speech sentences estimated using the three methods.

In addition, notice that the lowest SER_{MO} value is for the case of -5 dB SNR. This example has the lowest value since it corresponds to the most severe modification to the speech STFT magnitude. This means that the modified STFT magnitude for this case is very distant from the STFT magnitude of any actual signal. Hence, for the -5 dB example, the LSEE-MSTFTM algorithm converges upon signals whose STFT magnitudes are more distant from the modified STFT magnitudes for the given window length and DFT size.

5.4 Summary

In this chapter we considered the STFT magnitude modifications required for spectral subtraction enhancement of noisy speech. For all of the examples, there was no noticeable difference detected in the informal listening tests among the enhanced speech sentences estimated using the three different methods.

Table 5.1 - Spectral Subtraction Average SER Values				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	101.2 dB	101.8 dB	102.1 dB	Estimated signals are indistinguishable from each other. All sound good.
10	21.4 dB	22.1 dB	22.1 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	17.3 dB	17.9 dB	17.9 dB	Same as above
0	12.9 dB	13.5 dB	13.6 dB	Same as above
-5	8.7 dB	9.2 dB	9.2 dB	Same as above

Table 5.2 - Spectral Subtraction Results for Male Speech "Line up at the screen door."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	100.4 dB	100.5 dB	100.6 dB	Estimated signals are indistinguishable from each other. All sound good.
10	21.4 dB	22.0 dB	22.1 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	16.8 dB	17.4 dB	17.5 dB	Same as above
0	12.2 dB	12.7 dB	12.8 dB	Same as above
-5	7.8 dB	8.3 dB	8.3 dB	Same as above

Table 5.3 - Spectral Subtraction Results for Male Speech "He has the bluest eyes."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	99.2 dB	99.5 dB	99.6 dB	Estimated signals are indistinguishable from each other. All sound good.
10	21.1 dB	21.8 dB	21.9 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	16.8 dB	17.4 dB	17.5 dB	Same as above
0	12.4 dB	13.0 dB	13.0 dB	Same as above
-5	7.9 dB	8.5 dB	8.5 dB	Same as above

Table 5.4 - Spectral Subtraction Results for Male Speech "The chef made lots of stew."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	98.1 dB	98.8 dB	98.9 dB	Estimated signals are indistinguishable from each other. All sound good.
10	22.6 dB	23.3 dB	23.4 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	18.4 dB	19.0 dB	19.1 dB	Same as above
0	13.6 dB	14.2 dB	14.3 dB	Same as above
-5	9.0 dB	9.5 dB	9.6 dB	Same as above

Table 5.5 - Spectral Subtraction Results for Male Speech "You're the biggest man."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	93.1 dB	93.8 dB	93.9 dB	Estimated signals are indistinguishable from each other. All sound good.
10	22.2 dB	22.9 dB	23.0 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	17.6 dB	18.3 dB	18.4 dB	Same as above
0	12.8 dB	13.4 dB	13.5 dB	Same as above
-5	8.1 dB	8.7 dB	8.7 dB	Same as above

Table 5.6 - Spectral Subtraction Results for Female Speech "The bowl dropped from his hand."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	95.0 dB	95.7 dB	95.9 dB	Estimated signals are indistinguishable from each other. All sound good.
10	21.6 dB	22.2 B	22.3 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	17.7 dB	18.2 dB	18.3 dB	Same as above
0	13.5 dB	14.1 dB	14.1 dB	Same as above
-5	9.5 dB	10.1 dB	10.1 dB	Same as above

Table 5.7 - Spectral Subtraction Results for Female Speech "That shirt seems much too long."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	93.7 dB	94.4 dB	94.5 dB	Estimated signals are indistinguishable from each other. All sound good.
10	21.4 dB	21.9 dB	22.0 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	17.8 dB	18.4 dB	18.4 dB	Same as above
0	13.8 dB	14.5 dB	14.5 dB	Same as above
-5	9.9 dB	10.5 dB	10.5 dB	Same as above

Table 5.8 - Spectral SubtracD ..... "We made some fine brownies."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	91.8 dB	92.8 dB	92.9 dB	Estimated signals are indistinguishable from each other. All sound good.
10	21.6 dB	22.2 dB	22.3 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	17.4 dB	18.0 dB	18.1 dB	Same as above
0	13.1 dB	13.6 dB	13.7 dB	Same as above
-5	8.8 dB	9.3 dB	9.3 dB	Same as above

Table 5.9 - Spectral Subtraction Results for Female Speech "The yellow lion roared."				
SNR	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
100	138.5 dB	139.1 dB	140.1 dB	Estimated signals are indistinguishable from each other. All sound good.
10	19.6 dB	20.2 dB	20.3 dB	Estimated signals are indistinguishable from each other. All have artifacts.
5	15.5 dB	16.2 dB	16.2 dB	Same as above
0	12.0 dB	12.6 dB	12.6 dB	Same as above
-5	8.3 dB	8.9 dB	8.9 dB	Same as above

CHAPTER 6

APPLICATION TO HELIUM SPEECH ENHANCEMENT

6.1 Introduction

Since sound travels faster in a hyperbaric helium-oxygen atmosphere than in air at normal pressure, speech uttered in this type of environment suffers from certain severe degradations. This effect handicaps communication systems for deep-sea divers and others who must work in such an atmosphere. A STFT-based enhancement method exists [3], but the required modifications to the STFT apply only to its magnitude.

The following sections illustrate how the three signal estimation algorithms can be used in this enhancement problem. We begin with a short discussion of the effects of helium on the speech STFT. Then the STFT-based helium speech enhancement method is presented. Finally, experiments are performed to evaluate the performance of each of the algorithms for the case of helium speech enhancement.

6.2 Helium Effects upon Speech

The effects of the helium can be easily identified using short-time Fourier analysis. Specifically, the frequencies of the spectral envelope are increased non-linearly and the formant bandwidths are increased. These phenomena take place while the pitch information is left relatively undisturbed. A model exists for translating the spectral

envelope frequencies of helium speech back to their normal frequencies, and this model is suitable for use with the STFT.

Most models of helium speech assume that the speaker's pitch is unchanged in a helium atmosphere. This assumption is acceptable for several reasons. First, elementary acoustic theory does not predict a change in vocal pitch as it does for envelope frequencies. In addition, the indicated changes of glottal loading have an unknown effect upon the pitch. Another reason for assuming no pitch change is the fact that a speaker can very easily alter his own pitch. Thus attempts to model any kind of change would probably be futile if the speaker tried to compensate for changes himself. Finally, it is known that for a wide range of pitch variations the intelligibility of the associated speech is not significantly affected.

6.3 STFT Modifications for Helium Speech Enhancement

Based on the above cited differences between speech in helium atmosphere and speech in air, Quick [9] has given a mapping function useful for translating between helium speech envelope frequencies and normal envelope frequencies. The simplified version of this function is

$$f_h^2 = \alpha^2 f_a^2 + f_c^2. \tag{6.1}$$

Here f_a is the speech envelope frequency in air, α is the ratio of the speed of sound in pressurized helium-oxygen to that in air, and f_h is the corresponding speech envelope frequency spoken in the helium environment. In addition, f_c is the closed lip resonant

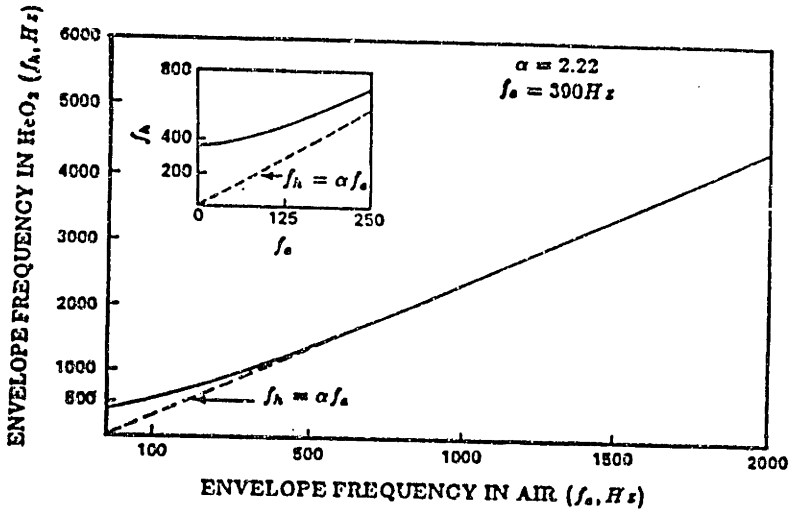
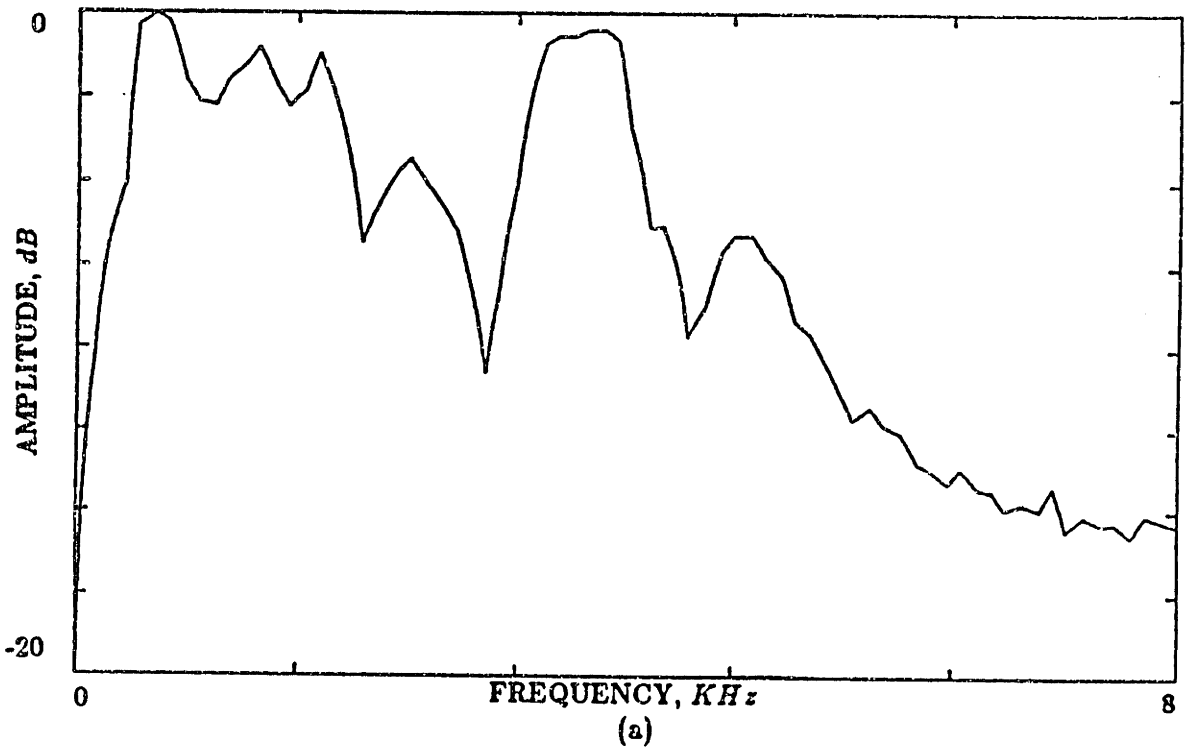


Figure 6.1. A typical envelope mapping function (After Richards [3]).

frequency of the vocal tract in a helium atmosphere. This number is highly dependent upon vocal tract length, thus it differs from individual to individual. It should also be noted that f_c is important because it helps to model the non-linear effect which the vocal tract wall vibration gives to the frequency translation.

In Figure 6.1 we see the nearly linear relationship which this mapping function takes on for a typical situation of 90% helium, 10% oxygen at an ambient pressure of 10 atmospheres. The subtle nonlinearity becomes significant at low frequencies. An example of the required change in frequency characteristics for one frame of speech is shown in Figures 6.2a and 6.2b. In Figure 6.2a, the STFT magnitude envelope for a voiced utterance of helium speech is shown. Then the frequency translation formula of Equation (6.1) was applied to the envelope with $\alpha = 2.22$ and $f_0 = 390Hz$. The result of this translation procedure is shown in Figure 6.2b for the same frame.



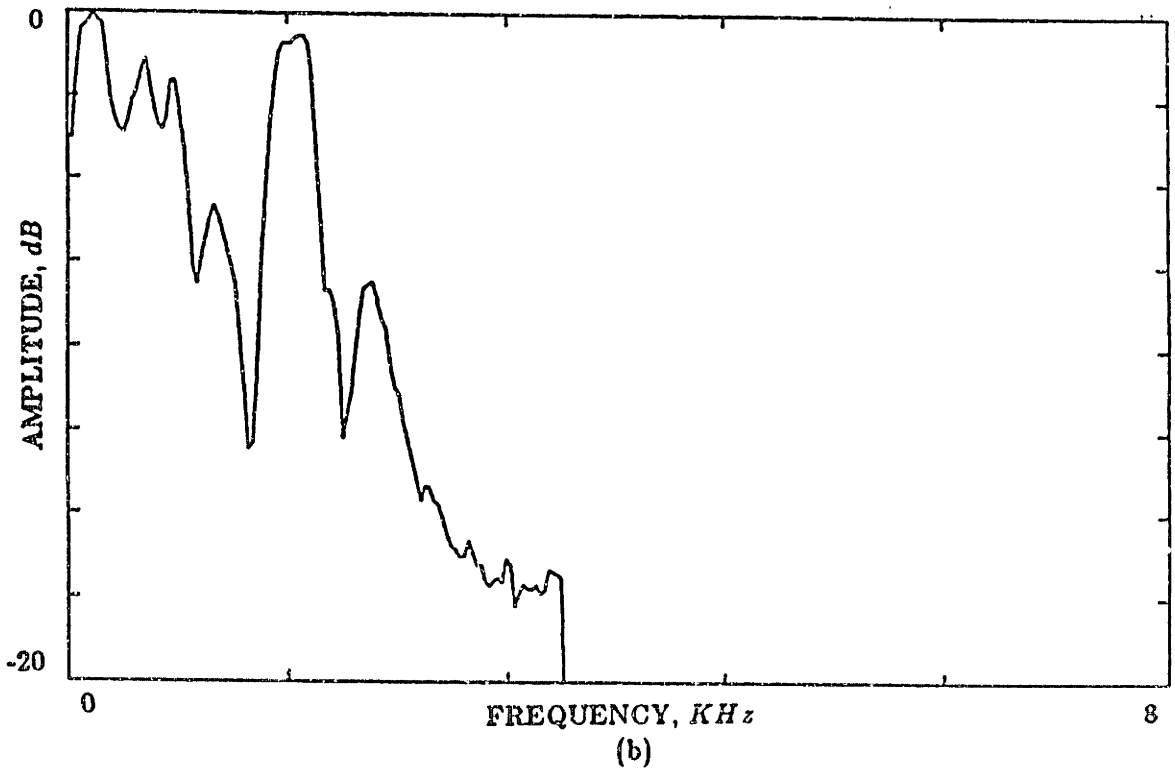


Figure 6.2. (a) A typical STFT envelope frame of voiced helium speech. (b) The same frame after appropriate enhancement using Equation (6.1).

6.4 Experimental Procedure and Results

Five helium speech sentences, spoken by a diver at depths ranging from 350 feet to 1000 feet, were used for the experiments performed here. All sentences were sampled at a 16KHz rate. The STFTs were computed using 400 point Hamming windows with a shift of 100 points, and the DFTs were 512 points long.

The envelope, $|T(mS,\omega)|$, and excitation components of the STFT magnitudes were separated using the same simple peak-picking scheme as performed by Richards [3]. The envelope frequencies were modified using Equation (6.1). The values of α and f_c used for each depth are given in Table 6.1. In each case, the modified STFT magnitude, $|\tilde{X}_w(mS,\omega)|$, is obtained from the modified envelope, $|\tilde{T}(mS,\omega)|$, and the original excitation just as in Equation (4.1).

In order to evaluate the performance of the three algorithms, the enhanced helium speech was estimated using the same procedure outlined in Figure 3.1. The SER values and the results of the informal listening tests are given in Table 6.2.

It can be seen from Table 6.2 that signal estimation using the LSEE-MSTFTM algorithm has resulted in very little improvement over estimation using the two overlap-add methods. The differences in SER values range from a 3 dB to 4 dB increase, which has been found to be insignificant for improvement of quality or intelligibility. Accordingly, in the informal listening tests, no audible difference could be detected among the speech sentences estimated using the three methods.

Also notice the specific values of SER_{MO} in these examples. The numbers are

Table 6.1 - Helium Speech Enhancement Parameters				
Male Speech				
Example	Depth	α	f_c	Sentence
1	1000 ft.	3.0	390 Hz	"It was the first of a three part series."
2	560 ft.	2.7	390 Hz	"I've had it all my life."
3	560 ft.	2.5	390 Hz	"I have a temper, said Hayes."
4	266 ft.	2.2	390 Hz	"Do you expect me to go on crying over spilt milk?"
5	266 ft.	2.0	390 Hz	"I have a lot of regrets."

Table 6.2 - Helium Speech Enhancement Results				
Male Speech				
Example	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
1	9.1 dB	9.2 dB	12.1 dB	Estimated signals are indistinguishable from each other. All sound good.
2	10.2 dB	10.5 dB	13.6 dB	Same as above
3	9.3 dB	9.4 dB	12.7 dB	Same as above
4	9.7 dB	9.9 dB	13.4 dB	Same as above
5	10.4 dB	10.6 dB	14.5 dB	Same as above

not as high as the ones computed in Tables 4.1—4.8. This indicates that these modifications are rather severe envelope modifications. This means that the modified STFT magnitudes for these cases are somewhat more distant from the STFT magnitude of any actual signal. Hence, the LSEE-MSTFTM algorithm converges upon signals whose STFT magnitudes are more distant from these modified STFT magnitudes for the given window length and DFT size.

6.5 Summary

In this chapter we considered the STFT magnitude modifications required for helium speech enhancement. For all of the examples, there was no noticeable difference detected in the informal listening tests among the enhanced speech sentences estimated using the three different methods.

CHAPTER 7

APPLICATION TO TIME-SCALE MODIFICATION

7.1 Introduction

The goal of time-scale modification is to increase or decrease the apparent rate of articulation of the speech phonemes contained in the original signal. Methods for accomplishing this task include time domain processing [10] and STFT-based approaches [11].

In the time domain approach, speech is windowed into sections long enough to include several pitch periods, while short enough that the speech signal may be considered time-invariant over each section. The usual procedure is to then discard sections or replicate sections corresponding to time compression or expansion. However, due to the resulting discontinuities and abrupt pitch changes between sections of the modified speech, the frequency content of the modified speech signal is altered in an undesirable way from the original speech. Thus, modification of a signal's STFT is a more appropriate method of time-scale alteration. The same general idea of discarding and replicating sections can be applied to the set of STFT frames. However, since the frames are Fourier transforms of sections of speech, more control over the frequency content of the modified speech is available.

The following sections illustrate how the three signal estimation algorithms can be used in this speech processing problem. A STFT-based time-scale modification method is presented. Then, experiments are performed to evaluate the performance of each of the algorithms for the case of time scale modification.

7.2 STFT Modifications for Time-Scale Modification

The discussion and experimental results presented in this chapter shall be limited to time-scale compression of speech at an integer ratio. Time-scale expansion and non-integer ratio modifications are possible with the STFT by changing the window shift parameter, S, during synthesis to a value different from the original window shift used for the STFT analysis [4]. However, consideration of these modifications would unnecessarily complicate the models used here. Therefore, only basic analysis and experimentation is presented with the understanding that the results have application to many other types of time-scale modification.

The time-scale compressed STFT magnitude can be obtained directly from Equation (2.1):

$$|\tilde{X}(mS, \omega)| = |X_w(\beta mS, \omega)| \tag{7.1}$$

where β is the compression ratio. The effect of Equation (7.1) is to create a modified STFT magnitude by using only every β frames, and discarding the others.

However, this modification method is not valid for estimating the appropriate STFT phase. Portnoff [11] shows that the phase component due to the pitch informa-

tion would be incorrect. In keeping with the modeling techniques used up to this point, the alternative estimate of $\arg[\tilde{X}(m_0 S, \omega_0)]$ is $\arg[X(m_0 S, \omega_0)]$. Using this approach, our STFT-based estimate of time compressed speech becomes:

$$\tilde{X}_w(mS, \omega) = |X_w(\beta mS, \omega)| \frac{X_w(mS, \omega)}{|X_w(mS, \omega)|}. \tag{7.2}$$

7.3 Experimental Procedure and Results

Eight test sentences are used for the experiments performed here. Four are spoken by males, and four are spoken by females. The sentences were sampled at a 10 KHz rate. The STFTs were computed using 256 point Hamming windows with a shift of 64 points, and a DFT size of 512 points. The time-scale modified STFT magnitude for each sentence was generated using Equation (7.1) with β = 2,3, and 4.

In order to evaluate the performance of the three algorithms, the time-scale modified speech was estimated using the same procedure outlined in Figure 3.1. The SER values and the results of the informal listening tests are given in Tables 7.2 through 7.9. Table 7.1 contains the average SER values computed from Tables 7.2—7.9.

It can be seen from Tables 7.1—7.9 that estimation using the LSEE-MSTFTM algorithm has resulted in quite a dramatic improvement over estimation using the two overlap-add methods. The differences in SER values range from a 6 dB to a 12 dB increase, which corresponds to a substantial improvement of quality and intelligibility. In the informal listening tests, the audible difference between the speech sentences estimated using the LSEE-MSTFTM algorithm and the sentences estimated using the

Table 7.1 - Time-Scale Modification Average SER Values				
Compression Ratio	SER_{OA}	SER_{WO}	SEP_{MO}	Listening Test
2:1	6.5 dB	6.5 dB	17.3 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	6.1 dB	6.2 dB	15.1 dB	Same as above
4:1	5.0 dB	5.9 dB	13.0 dB	Same as above

Table 7.2 - Time-Scale Modification Results for Male Speech				
"Line up at the screen door."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	6.4 dB	5.9 dB	16.8 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	6.2 dB	6.6 dB	14.0 dB	Same as above
4:1	5.3 dB	5.3 dB	12.2 dB	Same as above

Table 7.3 - Time-Scale Modification Results for Male Speech				
"He has the bluest eyes."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	5.9 dB	6.1 dB	17.6 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	6.1 dB	5.8 dB	15.0 dB	Same as above
4:1	5.0 dB	5.1 dB	14.2 dB	Same as above

Table 7.4 - Time-Scale Modification Results for Male Speech "The chef made lot's of stew."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	7.3 dB	7.5 dB	16.5 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	6.8 dB	7.0 dB	15.1 dB	Same as above
4:1	6.2 dB	6.5 dB	13.7 dB	Same as above

Table 7.5 - Time-Scale Modification Results for Male Speech "You're the biggest man."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	7.3 dB	7.4 dB	16.9 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	6.3 dB	6.3 dB	15.8 dB	Same as above
4:1	6.4 dB	6.4 dB	13.6 dB	Same as above

Table 7.6 - Time-Scale Modification Results for Female Speech "The bowl dropped from his hand."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	5.9 dB	5.9 dB	18.2 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	4.9 dB	5.0 dB	13.8 dB	Same as above
4:1	4.9 dB	5.3 dB	11.5 dB	Same as above

Table 7.7 - Time-Scale Modification Results for Female Speech				
"That shirt seems much too long."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	5.9 dB	6.2 dB	17.4 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	5.7 dB	6.1 dB	16.6 dB	Same as above
4:1	5.2 dB	6.3 dB	12.6 dB	Same as above

Table 7.8 - Time-Scale Modification Results for Female Speech				
"We made some fine brownies."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	6.8 dB	6.5 dB	18.3 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	6.3 dB	6.1 dB	16.2 dB	Same as above
4:1	5.4 dB	5.5 dB	13.8 dB	Same as above

Table 7.9 - Time-Scale Modification Results for Female Speech				
"The yellow lion roared."				
Compression Ratio	SER_{OA}	SER_{WO}	SER_{MO}	Listening Test
2:1	6.8 dB	6.7 dB	16.5 dB	Pronounced multiple pitch quality in both overlap-add estimates. Only LSEE-MSTFTM estimate sounds good.
3:1	6.5 dB	6.6 dB	13.9 dB	Same as above
4:1	6.2 dB	6.6 dB	12.0 dB	Same as above

two overlap-add methods was very distinct. The audible difference was the "multiple pitch" quality in the speech estimated using the overlap-adds, similar to that obtained in Section 4.2.2. There was no multiple pitch quality or any other degradation audible in the speech estimated using the LSEE-MSTFTM algorithm. There was no audible difference between the speech estimated using the overlap-add method and the speech estimated using the weighted overlap-add method.

7.4 Summary

In this chapter we considered the STFT magnitude modifications required for time-scale modification of speech. The experimental results presented here show that the performances of the two overlap-add methods are poor compared to the LSEE-MSTFTM performance. A "multiple pitch" quality exists for each example of time-scale modified speech which was estimated using the two overlap-add methods. The interfering pitch is related to the pitch of the original speech, which, in this application, is simply the non- time-scale modified pitch track.

CHAPTER 8

CONCLUSIONS

8.1 Summary

Several applications of signal estimation from the modified STFT have been explored here. The conventional overlap-add method has been compared with the more correct approach of a weighted overlap-add approach and the more accurate approach of the iterative, magnitude-only, LSEE-MSTFTM algorithm. The performance of these algorithms has been evaluated based on the mean square error between the STFT magnitudes of the estimated signals and the desired STFT magnitude for each example. The overall goal has been to determine those applications for which the STFT magnitude of the signal estimated using the LSEE-MSTFTM algorithm is significantly closer to the desired STFT magnitude than those estimated using either of the overlap-add methods. In the cases where the LSEE-MSTFTM algorithm does not produce a significantly better estimate, the computational speed and convenience of the overlap-add methods make them better choices as signal estimators.

The three methods of signal estimation were first applied to some arbitrary magnitude modifications. For the envelope modifications considered here, the LSEE-MSTFTM algorithm did not produce significantly higher signal-to-error ratio values than the overlap-add methods. In addition, there were no audible differences among

the three signal estimates. For the pitch modifications considered here, the LSEE-MSTFTM algorithm did, in general, produce significantly higher SER values than the overlap-add methods. Substantial audible differences were detectable between the LSEE-MSTFTM estimates and the overlap-add type estimates. The LSEE-MSTFTM estimates were clearly superior. Therefore, the envelope modified STFT magnitudes can be adequately estimated using one of the overlap-add methods, while the pitch modified STFT magnitudes can only be adequately estimated using the LSEE-MSTFTM algorithm.

There are some important implications of these results. Specifically, we can conclude that when the modification in question is made only to the envelope component of the STFT magnitude, any of the three signal estimation methods considered here are useful. On the other hand, we can conclude that when the modification in question affects the harmonic component of the STFT magnitude, only the LSEE-MSTFTM algorithm is a useful signal estimation method.

Next, the three methods of signal estimation were applied to the case of noisy speech enhancement by spectral subtraction. For the examples considered here, the LSEE-MSTFTM algorithm did not produce significantly higher SER values than the two overlap-add methods. In addition, there were no audible differences among the three signal estimates. Therefore, we can conclude for this application that there is no advantage to estimation using the LSEE-MSTFTM, and the computational savings of estimation using one of the overlap-add methods should always be exploited.

In the third case, the three signal estimation algorithms were applied to the case of helium speech enhancement. Again, in these examples, the LSEE-MSTFTM algorithm did not produce significantly higher SER values than the two overlap-add methods. The informal listening tests found no audible differences among the three signal estimates. Therefore, we can conclude for this application that one of the overlap-add methods should always be used.

Finally, the case of time-scale modification of speech was considered. For the examples considered here, the LSEE-MSTFTM algorithm always produced significantly higher SER values than the two overlap-add methods. In addition, there were substantial audible differences between the speech examples estimated using the LSEE-MSTFTM algorithm and the speech examples estimated using the two overlap-add methods. The LSEE-MSTFTM estimates were clearly superior. Therefore, we can conclude that time-scale modified speech should always be estimated using the LSEE-MSTFTM algorithm.

The STFT magnitude modifications for spectral subtraction and helium speech enhancement loosely correspond to envelope modifications, while the STFT based time-scale modifications most significantly correspond to pitch modifications. Therefore, the results of Chapters 5,6,and 7 are in agreement with the results of the arbitrary magnitude modifications in Chapter 4, and perhaps could have been predicted from the Chapter 4 results. Using this reasoning, these results can be applied to some extent to speech processing applications not covered here. That is, if a desired STFT

magnitude modification can be classified as one which affects the harmonic structure of the magnitude, then the LSEE-MSTFTM algorithm is the only useful signal estimation algorithm. On the other hand, if the desired modification is known to only affect the magnitude envelope, then any of the signal estimation methods are useful. If the above classification cannot be applied to the desired modification, then the same type of testing procedure as employed in these experiments will have to be used to reach a conclusion.

8.2 Suggested Further Work

In view of these results and conclusions, it is apparent that some guidelines of a more general nature are warranted for the prediction of an estimation algorithm's usefulness in a particular speech processing application. This is because we may not be able to classify some modification processes as simply envelope or pitch modifications. In addition, some signal estimation methods may not be classified as estimation from the modified STFT magnitude alone or estimation from the modified magnitude and unmodified phase. Therefore, it could be useful to consider in a broader sense the problem of signal estimation from the modified STFT.

All of the speech examples which were processed in this study were signals which fit the traditional terminal-analog model of speech production. However, the signal estimation algorithms which have been tested here were not developed with any speech production model in mind. Thus, further testing should be done using non-traditional signals. For example, the experiments should be extended to multiple speaker speech

processing applications. Such applications include time-scale modification of multiple speaker speech and enhancement of speech which is degraded by the presence of other speech. Other non-traditional signals include singing and instrumental music.

It is conceivable that new signal estimation techniques could be developed with the specifics of human perception in mind. For example consider the LSEE-MSTFTM algorithm. This method was developed with the specific intent of estimating a signal whose STFT magnitude comes close in the mean square error sense to the desired STFT magnitude. However, human perception may not be sensitive to the mean square error only. Or, the ear may only be sensitive to the mean square error in certain frequency and time bands. Development of a perception specific distance measure may lead to a high quality signal estimation method which would perform even better than the LSEE-MSTFTM algorithm.

APPENDIX A

SHORT-TIME FOURIER ANALYSIS OF SPEECH

A.1 Introduction

In order to gain more understanding of the results obtained in the preceding chapters, we can establish certain properties of the signals with which we are concerned. Speech production can be adequately modeled as the output of the terminal-analog system shown in Figure A.1. Using this model, the speech signal can be categorized as voiced or unvoiced. Since the unvoiced speech is generated from a filtered random signal, the Fourier transforms of unvoiced frames are also random sequences. For this reason, analysis of magnitude and phase consistency for unvoiced speech is rather complicated and will not be considered here.

The voiced speech is generated from a filtered periodic impulse train, the Fourier transforms of voiced frames have pronounced harmonic characteristics. This work focuses only on the issue of magnitude and phase consistency for voiced speech. The framework for analyzing this consistency shall be established in the next two sections.

A.2 Harmonic Representation of Voiced Speech

The mathematical model of voiced speech which is presented next is summarized from Portnoff's work [6]. A harmonic representation for voiced speech signals will be formulated by expressing the vocal excitation signal as a sum of harmonically related complex

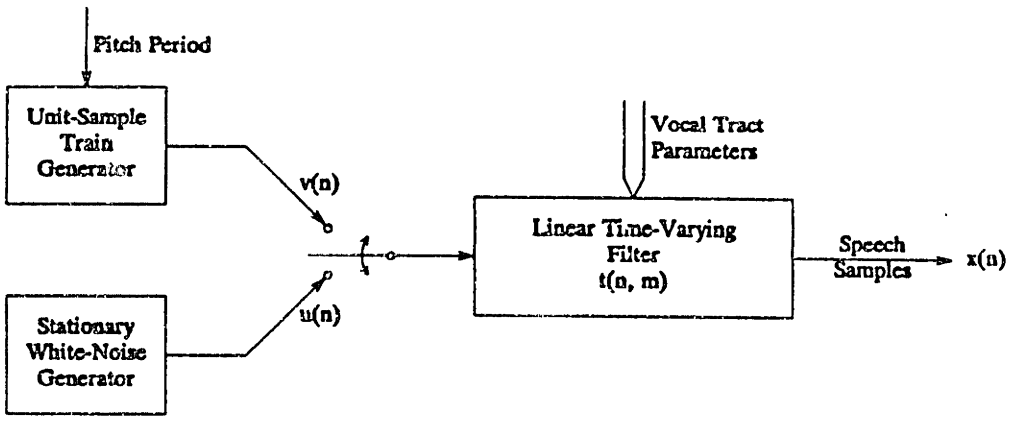


Figure A.1. Terminal-analog model of the vocal system (after Rabiner and Schafer [12]).

exponentials and using this representation in a convolution sum with the vocal tract impulse response.

Consider again the terminal-analog representation of speech. Voiced speech corresponds to the model of Figure A.1 with the unit-sample train input. Since the period, $P(n)$, of this unit-sample train, $v(n)$, slowly changes with time, it can be considered as periodic only for points that are local to the point of interest, as in Figure A.2. In this figure let $P(n_0)$ denote the local pitch period of $v(n)$ in the neighborhood of n_0, let $D(n_0)$ denote the number of samples to the sample arriving most recently before, or at the sample n_0, and let $\delta(n)$ denote the unit-sample function. Then, the local representation of $v(n)$ is

$$v(n_0 + \tau) \approx \sum_{i=-\infty}^{\infty} \delta(\tau + D(n_0) + iP(n_0)) \qquad (A.1)$$

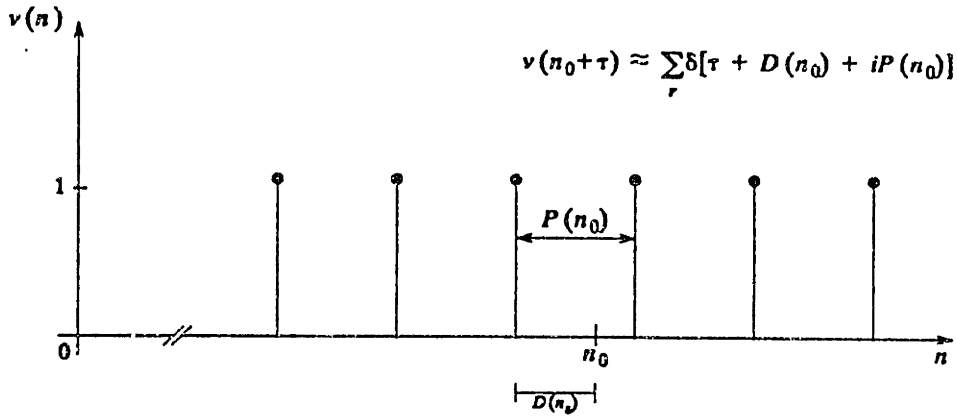


Figure A.2. Quasi-periodic unit-sample train (after Portnoff [6]).

for small $|\tau|$. This approximation has an equivalent expression as a sum of harmonically related complex exponentials:

$$v(n_0+\tau) \approx \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} e^{jk(\phi(n_0)+\Omega(n_0)\tau+\phi_0)} \qquad (A.2)$$

where

$$\Omega(n) = \frac{2\pi}{P(n)} \qquad (A.3a)$$

$$\phi_0 = \Omega(0)D(0) \qquad (A.3b)$$

$$\phi(n) = \Omega(n)D(n) + 2\pi I(n) - \phi_0. \qquad (A.3c)$$

The slowly varying quantity $\Omega(n)$ is called the "instantaneous frequency" of the fundamental, and $\phi(n)$ is referred to as the "instantaneous phase" of the fundamental. The integer quantity $I(n)$ specifies the additive multiple of 2π in Equation (A.3c) by requiring that the value of the exponent in Equation (A.2) be uniquely defined for each

value of $n = n_0 + \tau$. In order to satisfy this uniqueness condition, Portnoff [6] has explicitly *defined* $\phi(n)$ as

$$\phi(n) = \begin{cases} \sum_{r=1}^{n} \Omega(r) & \text{for} \quad n > 0, \\ 0 & \text{for} \quad n = 0, \\ \sum_{r=0}^{n+1} -\Omega(r) & \text{for} \quad n < 0. \end{cases} \tag{A.4}$$

Therefore, the excitation for voiced speech, $v(n)$, will be modeled as the sum of harmonically related complex exponentials

$$v(n) = \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} e^{jk(\phi(n)+\phi_0)}. \tag{A.5}$$

The resulting voiced speech signal, $x(n)$, modeled as the output of the time varying vocal tract impulse response $t(n, m)$ driven by $v(n)$, is given as the following convolution sum

$$x(n) = \sum_{m=-\infty}^{\infty} t(n, m)v(n - m). \tag{A.6}$$

For voiced speech, the pitch, $\Omega(n)$, is assumed to be constant for the duration of the memory of $t(n, m)$. Therefore, performing the convolution indicated in Equation (A.6) gives

$$x(n) = \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} T(n, k\Omega(n))e^{jk(\phi(n)+\phi_0)} \tag{A.7}$$

where $T(n, k\Omega(n))$ is the STFT of $t(n, m)$ evaluated at the pitch harmonic $k\Omega(n)$, and $T(n, k\Omega(n))$ is also slowly varying with time n. If we define these samples of $T(n, \omega)$ as a set of complex harmonic amplitudes

$$c_k(n) = \frac{1}{P(n)} T(n, k\Omega(n))e^{jk\phi_0}, \tag{A.8}$$

then we finally see the output voiced speech as a linear combination of harmonically related complex exponentials, i.e.

$$x(n) = \sum_{k=0}^{P(n)-1} c_k(n) e^{jk\phi(n)}. \qquad (A.9)$$

A.3 The Narrowband STFT of Voiced Speech

The model for voiced speech was summarized in section A.2. If we use that result we can determine the STFT of voiced speech by substituting Equation (A.9) into Equation (2.1), giving:

$$X_w(mS, \omega) = \sum_{n=-\infty}^{\infty} \sum_{k=0}^{P(n)-1} w(mS - n) c_k(n) e^{jk\phi(n)} e^{-j\omega n}. \qquad (A.10)$$

It was stated in section A.2 that $P(n)$ is slowly varying relative to the length of the analysis window $w(mS - n)$. Thus, we can assume the following:

$$P(n) \approx P(mS)$$

$$\phi(n) \approx \phi(mS) + \Omega(mS) \cdot (n - mS) \qquad \text{for} \quad w(mS - n) \neq 0. \qquad (A.11)$$

Making the indicated substitutions into Equation (A.10) gives:

$$\begin{aligned} X_w(mS, \omega) &= \sum_{k=0}^{P(mS)-1} \sum_{n=-\infty}^{\infty} w(mS - n) c_k(n) e^{-j(\omega - k\Omega(mS))n} e^{jk(\phi(mS) - \Omega(mS)mS)} \\ &= \sum_{k=0}^{P(mS)-1} c_k(mS) W(k\Omega(mS) - \omega) e^{j(k\phi(mS) - \omega mS)} \qquad (A.12) \end{aligned}$$

where $W(\omega)$ is the Fourier transform of $w(n)$. Equation (A.12) results from the slowly varying nature of the complex harmonic amplitudes, thus giving $w(mS - n)c_k(n) \approx w(mS - n)c_k(mS)$.

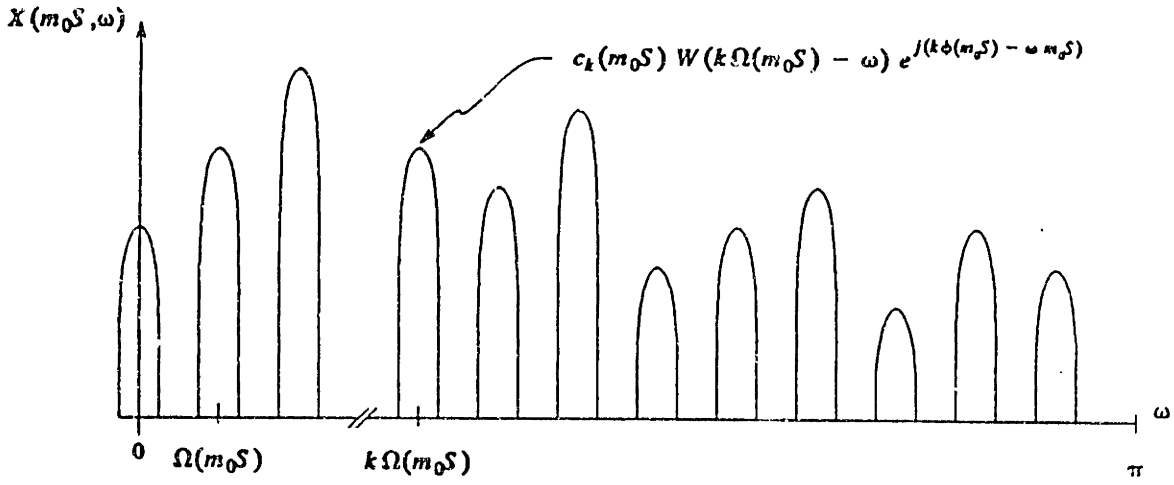


Figure A.3. Short-time Fourier transform of an idealized speech signal for a particular value of $mS = m_0S$ (after Portnoff [6]).

For narrowband analysis of voiced speech we employ an analysis window, w(n), whose length is a few times greater than the instantaneous pitch period, $P(n)$. Thus, the bandwidth of $W(\omega)$ is less than the instantaneous fundamental frequency, $\Omega(n)$. This means that the shifted and weighted images of $W(\omega)$ which make up Equation (A.12) are nonoverlapping, reducing Equation (A.12) to

$$X_w(mS,\omega) = \begin{cases} c_k(mS)W(k\Omega(mS)-\omega)e^{j(k\phi(mS)-\omega mS)} \\ \qquad\qquad \text{for} \quad |\omega - k\Omega(mS)| < \omega_h, \\ 0 \qquad\qquad \text{otherwise} \end{cases} \tag{A.13}$$

where ω_h is the cutoff frequency of $W(\omega)$. The idealized narrowband STFT is illustrated in Figure A.3.

A.4 Summary

In this appendix, Portnoff's work which models the narrowband STFT of voiced speech

was summarized. The next appendix will use this voiced speech model to discuss the effects of STFT magnitude modifications upon the speech estimated using the overlap-add method.

APPENDIX B

SOME OBSERVATIONS ON MODIFIED STFT MAGNITUDE CONSISTENCY WITH THE UNMODIFIED PHASE

B.1 Introduction

The analysis of voiced speech which was summarized in Appendix 1 will now enable us to investigate some issues related to modified STFT magnitude consistency with the original phase. Consistency is of concern when speech is to be estimated from the modified magnitude and original phase using a method such as overlap-add. For simplicity, we will say that inconsistency exists when the speech estimated using overlap-add is significantly different from the speech estimated from only the modified magnitude using a method such as the LSEE-MSTFTM algorithm. That is we are not concerned with how accurately the LSEE-MSTFTM algorithm estimates a signal with STFT magnitude close to the desired magnitude, rather we are only interested in the problems associated with combining a given magnitude with a phase function which is known to be incorrect for that magnitude. For this reason, this chapter will be devoted to determining the effects of magnitude modifications upon the speech estimated using the overlap-add method.

B.2 Signal Estimation from a Modified STFT by Overlap-Add

A single frame of the STFT magnitude of voiced speech has two components. One is the comb-like harmonic structure which comes from the quasi-periodic nature of the excitation signal. This component is called the pitch structure. Another part of the spectrum is the weighting of each harmonic, which comes from the frequency response of the vocal tract and has a smooth spectral shape. This component is called the envelope. We shall now consider separately the outcome of modifying each of these magnitude components and estimating speech from the resulting modified STFT.

B.2.1 Envelope modifications

Consider the narrowband STFT of voiced speech given by Equation (A.12). Modifying the envelope component of the magnitude gives

$$\tilde{X}_w(mS,\omega) = \sum_{k=0}^{P(mS)-1} \tilde{c}_k(mS) W(k\Omega(mS) - \omega) e^{j(k\phi(mS) - \omega mS)} \qquad (B.1)$$

where

$$\tilde{c}_k(n) = \frac{1}{P(n)} \tilde{T}(n, k\Omega(n)) e^{j[\theta(n,k\Omega(n)) - \tilde{\theta}(n,k\Omega(n))]} e^{jk\phi_0} \qquad (B.2)$$

and

$$\tilde{T}(n,\omega) = |\tilde{T}(n,\omega)| e^{j\tilde{\theta}(n,\omega)}.$$

That is, modifying the envelope is merely modifying the magnitude of the complex harmonic weights to correspond to a new vocal tract model, $\tilde{T}(n,\omega)$. The $\theta(n,k\Omega(n)) - \tilde{\theta}(n,k\Omega(n))$ phase correction term is included in Equation (B.2) since the complex

harmonic amplitudes are intended to maintain the STFT phase characteristic of the *original* vocal tract model.

Now, in order to establish the effects of this modification upon the estimated time-domain signal, we will synthesize the speech using overlap-add. Substituting the modified STFT of Equation (B.1) for the unmodified STFT in Equation (2.4) we get the following inverse Fourier transform of each STFT frame:

$$y_{OA_w}(mS,n) = \sum_{k=0}^{P(mS)-1} \tilde{c}_k(mS)e^{jk\phi(mS)} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} W(k\Omega(mS)-\omega)e^{j\omega(n-mS)}d\omega. \tag{B.3}$$

Using a substitution of variables in Equation (B.3) we get:

$$y_{OA_w}(mS,n) = \sum_{k=0}^{P(n)-1} \mathrm{w}(mS-n)\tilde{c}_k(n)e^{jk\phi(n)} \tag{B.4}$$

which by definition implies that

$$y_{OA}(n) = \sum_{k=0}^{P(n)-1} \tilde{c}_k(n)e^{jk\phi(n)}. \tag{B.5}$$

Now it is desirable to separate the estimated speech signal, $y_{OA}(n)$, into recognizable components. Substituting Equation (B.2) into Equation (B.5) gives us

$$\begin{aligned} y_{OA}(n) &= \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \tilde{T}(n,k\Omega(n))e^{j[\theta(n,k\Omega(n))-\tilde{\theta}(n,k\Omega(n))]}e^{jk(\phi(n)+\phi_0)} \\ &= \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} \left(\sum_{m=-\infty}^{\infty} \tilde{t}(n,m)e^{-jk\Omega(n)m} \right) \\ &\quad \cdot e^{j[\theta(n,k\Omega(n))-\tilde{\theta}(n,k\Omega(n))]}e^{jk(\phi(n)+\phi_0)} \end{aligned}$$

$$= \sum_{m=-\infty}^{\infty} \tilde{t}(n,m) \cdot \left\{ \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} e^{jk[\phi(n)-\Omega(n)m+\phi_0]} e^{j[\theta(n,k\Omega(n))-\tilde{\theta}(n,k\Omega(n))]} \right\}. \quad (B.6)$$

Assuming that the original and the new (modified) vocal tract functions are slowly varying with respect to the analysis window allows us to substitute $\theta(n-m,\omega) \approx \theta(n,\omega)$ and $\tilde{\theta}(n-m,\omega) \approx \tilde{\theta}(n,\omega)$ into Equation (B.6), giving

$$y_{OA}(n) = \sum_{m=-\infty}^{\infty} \tilde{t}(n,m) \cdot \left\{ \frac{1}{P(n-m)} \sum_{k=0}^{P(n-m)-1} e^{jk(\phi(n-m)+\phi_0)} e^{j[\theta(n-m,k\Omega(n-m))-\tilde{\theta}(n-m,k\Omega(n-m))} \right\}. \quad (B.7)$$

Equation (B.7) is a convolution sum which indicates that the new vocal tract model is being driven by the following excitation signal:

$$\tilde{v}(n) = \frac{1}{P(n)} \sum_{k=0}^{P(n)-1} e^{j[\theta(n,k\Omega(n))-\tilde{\theta}(n,k\Omega(n))]} e^{jk(\phi(n)+\phi_0)}. \quad (B.8)$$

Equation (B.8) shows us that an envelope modification can be modeled as a modification to the excitation signal. The usefulness of this approach can soon be appreciated. As stated earlier, for a small neighborhood surrounding $n = n_0$, we can approximate $\tilde{v}(n)$ as a strictly periodic sequence. That is

$$\tilde{v}(n_0+\tau) = \frac{1}{P(n_0)} \sum_{k=0}^{P(n_0)-1} e^{j[\theta(n_0,k\Omega(n_0))-\tilde{\theta}(n_0,k\Omega(n_0))]} e^{jk(\Omega(n_0)(n_0+\tau)+\phi_0)} \quad (B.9)$$

for small $|\tau|$. Since $\theta(n_0,\omega)$ and $\tilde{\theta}(n_0,\omega)$ are both odd functions of ω, $\tilde{v}(n_0+\tau)$ is real. Now we can view $\theta(n_0,k\Omega(n_0)) - \tilde{\theta}(n_0,k\Omega(n_0))$ as simply time shifting each harmonic

component. Moreover, since $\theta(n_0, \omega)$ and $\tilde{\theta}(n_0, \omega)$ are both STFT phase functions of real sequences, we can make the following statements:

$$\theta(n_0, 0) - \tilde{\theta}(n_0, 0) = 0 \qquad (B.10a)$$

$$\theta(n_0, k\Omega(n_0)) - \tilde{\theta}(n_0, k\Omega(n_0)) \approx 0 \qquad \text{for } k \text{ close to } \quad 0, P(n_0)/2, P(n_0). \qquad (B.10b)$$

Thus, some of the harmonic components of $\tilde{v}(n)$ are not significantly time shifted from the corresponding components of $v(n)$. Since envelope phase differences manifest themselves as a time shifting of the harmonic components of the excitation signal, the over-all periodicity of the excitation signal is preserved. In addition, some work [13] suggests that this time shifting of harmonic components is a condition to which humans are not very sensitive. Thus we conclude that for modifications to the spectral envelope, consistency with the original STFT phase is not much of an issue, and the signal estimated from the modified magnitude and unmodified phase using overlap-add will not be significantly poorer than the LSEE-MSTFTM signal estimate.

B.2.2 Pitch modifications

Now consider a modification to the STFT magnitude which effectively alters the pitch harmonic structure of the voiced speech while maintaining the same spectral envelope shape. Here we want to know what time domain signal results when the STFT magnitude is modified in this manner, then combined with the unmodified STFT phase, and

the resulting modified STFT is then substituted in the overlap-add procedure of Equation (2.4). We start with the narrowband STFT magnitude which is a modification of the magnitude of the STFT in Equation (A.12):

$$|\tilde{X}_w(mS,\omega)| = \sum_{i=0}^{\gamma P(mS)-1} |c_i(mS)| W\left(i\frac{1}{\gamma}\Omega(mS) - \omega\right) \qquad (B.11)$$

The indicated modifications are: 1) the pitch period changed by a factor γ; 2) the corresponding change in pitch fundamental frequency by the factor $\frac{1}{\gamma}$; and 3) the new set of complex harmonic amplitudes, $c_i(mS)$, which correspond to samples of the vocal tract spectrum, $T(mS,\omega)$, taken at harmonics of the new pitch frequency.

Consider the weighted images of $W(\omega)$ which make up the harmonic structure of the STFT magnitudes in Equation (A.12) and Equation (B.11). Typically, for rational γ, there are a few cases in each STFT frame where a particular harmonic, say i_0,of the pitch modified magnitude will be positioned exactly at a point in frequency where a harmonic, say k_0, of the original magnitude is located. In such a case we have $k_0 = \frac{1}{\gamma}i_0$. The result in terms of magnitude is

$$W\left(k_0\Omega(mS) - \omega\right) = W\left(\frac{i_0}{\gamma}\Omega(mS) - \omega\right) \qquad \text{for} \quad -\pi < \omega < \pi.$$

From Equation (A.13), in the region $|\omega - k_0\Omega(mS)| < \omega_h$, the phase of the original STFT is $(k_0\phi(mS) - \omega mS)$, and the phase which is consistent with the modified magnitude is $\left(i_0\frac{1}{\gamma}\phi(mS) - \omega mS\right)$, But, we have already said that $k_0 = \frac{1}{\gamma}i_0$ in this example. Therefore, the original STFT phase is consistent with the modified magnitude for this frequency band.

Using the above result, if we create a modified STFT as a combination of the pitch modified STFT magnitude of Equation (B.11) and the unmodified STFT phase of Equation (A.12), we can express our modified STFT as a combination of harmonics with consistent magnitude and phase, and harmonics with inconsistent magnitude and phase, i.e.

$$\tilde{X}_w(mS,\omega) = \sum_{\substack{i=0 \\ \frac{1}{\gamma}i=\text{integer}}}^{\gamma P(mS)-1} c_i(mS) W\left(i\frac{1}{\gamma}\Omega(mS)-\omega\right) e^{j\left(i\frac{1}{\gamma}\phi(mS)-\omega mS\right)} + \sum_{\substack{i=0 \\ \frac{1}{\gamma}i\neq\text{integer}}}^{\gamma P(mS)-1} |c_i(mS)| W\left(i\frac{1}{\gamma}\Omega(mS)-\omega\right) \frac{X_w(mS,\omega)}{|X_w(mS,\omega)|} \quad . \tag{B.12}$$

Incidentally, a special case exists when $\frac{1}{\gamma}$ is itself an integer. In such a case, the second component of Equation (B.12) disappears and our original phase becomes completely consistent with the modified magnitude. Thus, for pitch frequency increases of an integer factor, the desired speech can be estimated extremely well using overlap-add.

At this point we will not attempt to determine the general result of overlap-add estimation from Equation (B.12). However certain properties of such an estimated signal can be deduced from Equation (B.12). First, we see that the resulting speech will contain a signal corresponding to some vocal tract function driven by an impulse train of the desired pitch frequency, $\frac{1}{\gamma}\Omega(mS)$. Second, the resulting speech will contain a signal corresponding to some other vocal tract function driven by an excitation whose frequency is some combination of the original and the modified pitch. Therefore, the speech synthesized using overlap-add will take on a "multiple pitch" quality.

Experimentation supports this conclusion.

B.3 Discussion of Experimental Results

In Chapter 4, the types of magnitude modifications that were implemented are exactly the same as those considered in Section B.2. The experimental results support the observations made in that section. That is, it was found that for envelope modifications, the overlap-add method is a useful signal estimation method, while for pitch modifications, the overlap-add method is not a useful signal estimation method.

In Chapter 5, the spectral subtraction modifications were made according to Equation (5.7). The effect of Equation (5.7) is to reduce the spectral magnitude in areas of low SNR, such as at high frequencies and between harmonic peaks. It should be noted that the model of voiced speech which was summarized in Appendix A applies to clean speech. In addition, only multiplicative envelope modifications were considered in Section B.2.1. However, the spectral subtraction STFT magnitude modification in no way affects the positioning of the low frequency harmonic peaks. Therefore, although spectral subtraction is not strictly an envelope modification process, it can be loosely approximated as one. Since it was found that the overlap-add method is a useful signal estimation method for spectral subtraction, the experimental results of Chapter 5 support the prediction that this type of STFT magnitude modification does not cause a consistency problem with the phase function.

In Chapter 6, the envelope translation technique of Equation (6.1) for enhancement of helium speech belongs in the category of envelope modifications of the STFT magni-

tude. As such, the experimental results support the prediction that this type of STFT magnitude modification does not cause a consistency problem with the phase function.

In Chapter 7, we considered one method of time-scale compression based on the STFT which used Equation (7.2) for its modification procedure. The modified magnitude and original phase of Equation (7.2) are clearly unrelated. Therefore, we can view this procedure as a modification of both components of the original STFT magnitude. That is, we have effectively modified the spectral envelope but more importantly we have effectively modified the pitch harmonic structure. Since it was found that the overlap-add method is not a useful method for time-scale modification of speech, the experimental results of Chapter 7 support the prediction that this type of STFT magnitude modification causes a consistency problem with the phase function.

B.4 Summary

In this chapter we have explored the issues which are most important for STFT magnitude and phase consistency. In addition, we have determined the effects of estimating speech from a modified STFT magnitude and unmodified phase using the overlap-add process. Specifically, we can conclude that when the spectral envelope of a STFT magnitude is modified, it remains highly consistent with the unmodified STFT phase. In addition, we have determined that, in general, when the harmonic structure of a STFT magnitude is modified, the result becomes highly inconsistent with the unmodified STFT phase. In this case, the inconsistency manifests itself as a "multiple pitch" phenomenon upon the signal estimated using overlap-add.

REFERENCES

[1] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–32, pp. 236—243, Apr. 1984.

[2] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, pp. 1558—1564, Nov. 1977.

[3] M. Richards, "Helium speech enhancement using the short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–30, pp. 841—853, Dec. 1982.

[4] D. Griffin, "Signal estimation from modified short-time Fourier transform magnitude," M. S. Thesis, M. I. T., Dept. of EECS, Dec. 1983.

[5] D. Griffin, D. Deadrick, and J. Lim, "Speech synthesis from short-time Fourier transform magnitude and its application to speech processing," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 2.4.1—2.4.4 Mar. 1984.

[6] M. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, pp. 364—373, Jun. 1981.

[7] D. Griffin, and J. Lim, "A new pitch detection algorithm," *Proc. Intl. Conf. on Digital Signal Processing*, Sept. 1984, to be published.

[8] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586—1604, Dec. 1979.

[9] R. F. Quick, Jr., "Helium speech translation using homomorphic techniques," Air Force Cambridge Research Laboratories, Rep. AFCRL-70-0424, Jul. 1970.

[10] G. Fairbanks, W. Everitt, and R. Jaeger, "Method for time or frequency compression-expansion of speech," *IRE Trans. Professional Group on Audio*, vol. AU–2, pp. 7—12, Jan. 1954.

[11] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, pp. 374—390, Jun. 1981.

[12] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall Inc., Englewood Cliffs, NJ, 1978.

[13] J. Flanagan, *Speech Analysis, Synthesis, and Perception, 2nd ed.*, Springer-Verlag, New York, NY, 1972.

REFERENCES

[1] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–32, pp. 236—243, Apr. 1984.

[2] J. Allen and L. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE*, vol. 65, pp. 1558—1564, Nov. 1977.

[3] M. Richards, "Helium speech enhancement using the short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–30, pp. 841—853, Dec. 1982.

[4] D. Griffin, "Signal estimation from modified short-time Fourier transform magnitude," M. S. Thesis, M. I. T., Dept. of EECS, Dec. 1983.

[5] D. Griffin, D. Deadrick, and J. Lim, "Speech synthesis from short-time Fourier transform magnitude and its application to speech processing," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 2.4.1—2.4.4 Mar. 1984.

[6] M. Portnoff, "Short-time Fourier analysis of sampled speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, pp. 364—373, Jun. 1981.

[7] D. Griffin, and J. Lim, "A new pitch detection algorithm," *Proc. Intl. Conf. on Digital Signal Processing*, Sept. 1984, to be published.

[8] J. Lim and A. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, pp. 1586—1604, Dec. 1979.

[9] R. F. Quick, Jr., "Helium speech translation using homomorphic techniques," Air Force Cambridge Research Laboratories, Rep. AFCRL-70-0424, Jul. 1970.

[10] G. Fairbanks, W. Everitt, and R. Jaeger, "Method for time or frequency compression-expansion of speech," *IRE Trans. Professional Group on Audio*, vol. AU–2, pp. 7—12, Jan. 1954.

[11] M. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–29, pp. 374—390, Jun. 1981.

[12] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, Prentice Hall Inc., Englewood Cliffs, NJ, 1978.

[13] J. Flanagan, *Speech Analysis, Synthesis, and Perception, 2nd ed.*, Springer-Verlag, New York, NY, 1972.