SIGNAL ESTIMATION FROM MODIFIED SHORT-TIME FOURIER TRANSFORM Magnitude

by

DANIEL W. GRIFFIN

B.S., University of Michigan (1981)

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

December 1983

© Massachusetts Institute of Technology, 1983

Signature of Author —

Department of Electrical Engineering
December 7, 1983

Certified by —

Jae S. Lim
Thesis Supervisor

Accepted by —
Chairman, Departmental Committee on Graduate Students

ARCHIVES
 MASSACHUSETTS INSTITUTE OF TECHNOLOGY
 JUN 21 1984
 LIBRARIES
SIGNAL ESTIMATION FROM
MODIFIED SHORT-TIME FOURIER TRANSFORM

Daniel W. Griffin

Submitted to the Department of Electrical Engineering and Computer Science,
December 1983, in partial fulfillment of the requirements for the degree of
Master of Science

ABSTRACT

In a number of applications, the Short-Time Fourier Transform (STFT) of a signal is processed in some manner to produce a Modified Short-Time Fourier Transform (MSTFT). Since, in general, this MSTFT is not the STFT of any signal, it is important to develop algorithms to estimate a signal with STFT close to the desired MSTFT. Two distinct problems are investigated: 1) Estimating a signal with STFT (both magnitude and phase) close to a given MSTFT (both magnitude and phase) and 2) Estimating a signal with STFT Magnitude (STFTM) close to a given MSTFT Magnitude (MSTFTM). For the first problem, a least-squares error criterion is defined between the STFT of the signal estimate and the given MSTFT. Minimization of this error criterion yields a closed form solution which is termed LSEE-MSTFT for Least Squares Error Estimation from the MSTFT.

The second problem is important for applications where an accurate phase estimate is not available. A number of algorithms which minimize error criterions involving only STFTM are investigated. Of these algorithms, Least Squares Error Estimation from MSTFTM (LSEE-MSTFTM) and Least Squares Error Estimation using Nawab's iterative method [10] (LSEE-Nawab) which were both developed from the LSEE-MSTFT algorithm are the best in terms of convergence rate and the signal estimate produced. The LSEE-MSTFTM algorithm has the advantage that it can be implemented in real-time with cascaded processors. The LSEE-Nawab algorithm has the advantage of being more efficient to implement with general purpose computers.

All of these algorithms are compared for signal estimation from unmodified STFT and for signal estimation from the MSTFT given by the spectral subtraction estimate. The magnitude algorithms are also applied to time-scale modification of speech with the result that LSEE-MSTFTM and LSEE-Nawab produce very high quality estimates of time-scale modified speech.

Thesis Supervisor: Jae S. Lim

Title: Associate Professor of Electrical Engineering
Acknowledgments

I would like to express my gratitude to Professor Jae S. Lim for his supervision of this thesis. His guidance and insight were invaluable in providing direction and structure for this work. I would also like to thank Hamid Nawab for introducing me to the topic of reconstruction from the Short-Time Fourier Transform magnitude.

I wish to thank the members of the Digital Signal Processing Group for helping to provide an environment conducive to research. Special thanks to Webster Dove for maintaining the computer system during the majority of this work.

I would also like to thank Valerie Moses who is responsible for my initial selection of MIT and staying at MIT after experiencing Boston. My parents also deserve thanks for encouraging my early studies.
Table of Contents

Abstract 2

Acknowledgements 3

Table of Contents 4

Chapter 1: Introduction 7

Chapter 2: Signal Estimation from MSTFT 11

2.1: Introduction 11

2.2: Least Squares Error Estimation from the MSTFT 11

Chapter 3: Signal Estimation from MSTFTM 16

3.1: Introduction 16

3.2: Least Squares Error Estimation from MSTFTM 17

3.3: Other Simultaneous Estimation Methods 21

3.3.1: Simultaneous Steepest Descent 22

3.3.2: Simultaneous Conjugate Gradient 25

3.4: Sequential Estimation Methods 27

3.4.1: Signal Estimation for a Single Window from MFTM 28

3.4.1.1: Sequential Steepest Descent 29

3.4.1.2: Sequential Conjugate Gradient 31

3.4.1.3: Nawab's Algorithm 33

3.4.2: Signal Estimation using Estimates from MFTM Algorithms 34

Chapter 4: Criteria for Performance Comparison 37

4.1: Introduction 37
4.2: Algorithms
4.3: Comparison Criteria

Chapter 5: Signal Reconstruction from Unmodified STFT
5.1: Introduction
5.2: Signal Reconstruction from Unmodified STFT
5.3: Signal Reconstruction from Unmodified STFT Magnitude
5.3.1: Convergence Rate Examples
5.3.2: Reconstruction Examples
5.4: Summary

Chapter 6: Speech Enhancement by Spectral Subtraction
6.1: Introduction
6.2: Spectral Subtraction Signal Estimation from MSTFT
6.3: Spectral Subtraction Signal Estimation from MSTFTM
6.3.1: Simultaneous Algorithms
6.3.2: Sequential Algorithms
6.4: Conclusions

Chapter 7: Time-Scale Modification of Speech
7.1: Introduction
7.2: Signal Estimation from Time-Scale Modified STFTM
7.3: Conclusions

Chapter 8: Conclusions
8.1: Summary
8.1.1: Summary of Signal Estimation from MSTFT
8.1.2: Summary of Signal Estimation from MSTFTM
8.2: Suggestions for Further Research

Appendix A: Convergence Proof for LSEE-MSTFTM 102

Appendix B: Convergence Proof for Nawab's Algorithm 108

References 112
Chapter 1

Introduction

In a number of practical applications [1]-[5], it is desirable to modify the short-time Fourier transform (STFT) or the short-time Fourier transform magnitude (STFTM) and then estimate the processed signal from the modified STFT (MSTFT) or the modified STFTM (MSTFTM). For example, in speech enhancement by spectral subtraction [2,3], the STFT is modified by combining the STFT phase of the degraded speech with a MSTFTM and then a signal is reconstructed from the MSTFT. As another example, in time-scale modification of speech, one approach is to modify the STFTM and then reconstruct a signal from the MSTFTM. In most applications, including the two cited above, the MSTFT or MSTFTM is not valid in the sense that no signal has the MSTFT or MSTFTM, and therefore it is important to develop algorithms to estimate a signal with STFT or STFTM close in some sense to the desired MSTFT or MSTFTM.

Previous approaches to the problem of estimating a signal with STFT close to a given MSTFT include the overlap-add method [6,7] and the weighted overlap-add method [8,9]. These methods were developed to reconstruct a signal from its exact STFT and have been
applied heuristically to the problem of estimating a signal from a MSTFT. In Chapter 2, this problem is formulated as a minimization of a least squares distance measure between the STFT of the signal estimate and the given MSTFT. The method which results, termed Least Squares Error Estimation from MSTFT (LSEE-MSTFT), estimates a signal with STFT closest to a given MSTFT in terms of this distance measure. In Chapter 4, LSEE-MSTFT is shown to reconstruct the original signal exactly from unmodified STFT. This method is applied to speech enhancement by spectral subtraction in Chapter 5 with results comparable to the overlap-add algorithm. In Chapter 3, the LSEE-MSTFT algorithm is used to develop algorithms for estimating a signal with STFTM close to a given MSTFTM.

Although, signal estimation from MSTFT usually involves only simple non-iterative techniques, signal estimation from the MSTFTM is desirable since distance measures of STFTM conform more closely to the distance measure of the auditory system than distance measures of the STFT which include phase. The problem of reconstructing a signal from its STFTM has been previously considered by Nawab [10]-[12]. Nawab presents several methods for reconstructing a signal from its exact STFTM. These methods start with known points of the signal estimate and then estimate the remaining points covered by the first window of the signal estimate. Then, these points are fixed and the remaining points covered by the next window are estimated. In this manner, the signal estimate is obtained sequentially. It is important to note that when the STFTM has been modified, none of these methods are guaranteed to converge to the critical points of any distance measure between the STFTM of the signal estimate and the given MSTFTM. Of these methods, the iterative algorithm which he developed usually obtains the best signal estimate when the STFTM has been modified. This algorithm can be shown to converge (Chap. 3) to a set consisting of the critical points of a distance measure between the modified Fourier transform magnitude (MFTM) for a single window of the given MSTFTM and the FTM for a single window of the signal estimate. This method tends to
produce glitches in the signal estimate due to convergence to local minima in the distance measure.

The first signal estimation method developed in Chapter 3 is an iterative method based on the LSEE-MSTFT method developed in Chapter 2. In Appendix A, this algorithm is shown to converge to a set consisting of the critical points of a distance measure between the given MSTFTM and the STFTM of the signal estimate and has been termed LSEE-MSTFTM for Least Squares Error Estimation from the MSTFTM.

Several other methods which minimize a distance function over all of the windows of the STFTM simultaneously are also developed in Chapter 3. These include steepest descent, conjugate gradient, and an iterative algorithm based on the overlap-add method (OA-MSTFTM). Although the OA-MSTFTM algorithm does not converge in general, it appears to reduce the distance between the given MSTFTM and the STFTM of the signal estimate enough to produce a good signal estimate for the applications studied. Of simultaneous methods, the LSEE-MSTFTM and OA-MSTFTM have the best performance in terms of convergence rate and quality of the signal estimate obtained. These two algorithms have the additional advantage of being implementable in real-time with cascaded processors, one for each iteration.

Several sequential estimation methods have also been developed in this thesis. These include a combination of Nawab’s algorithm with the LSEE-MSTFT method discussed earlier and more conventional minimization techniques such as steepest descent and conjugate gradient. Of these sequential methods, Nawab’s method combined with LSEE-MSTFT (LSEE-Nawab) appears to have the best performance. Since Nawab’s algorithm and all similar sequential algorithms decrease distance functions over individual windows and do not necessarily decrease distance functions over all the windows of the STFTM, better performance is expected from the algorithms which minimize a distance function over all of the windows of
the STFTM simultaneously.

In Chapter 4, the criteria for performance comparisons of the algorithms presented in Chapters 2 and 3 are given along with an abbreviation and brief description of each algorithm. In Chapter 5, these algorithms are compared for the problem of signal estimation from unmodified STFT and STFTM. These examples show the problems which result when the STFTM algorithms converge to local minima in the distance functions instead of global minima.

Further comparisons of the performance of these algorithms are made in Chapters 6 and 7 for the problems of speech enhancement by spectral subtraction and time-scale modification of speech. Chapter 8 gives a summary of the performance of the estimation algorithms discussed in this thesis.
Chapter 2

Signal Estimation from Modified Short-Time Fourier Transform

2.1. Introduction

For those applications where the STFT has been modified [1]-[5], it is often desirable to estimate the processed signal directly from this modified STFT (MSTFT). For example, in speech enhancement by spectral subtraction [2,3], the STFT is modified by combining the STFT phase of the degraded speech with a modified STFT magnitude (MSTFTM) and then, a signal is reconstructed from this MSTFT. As another example, time varying filters such as those used in multi-microphone dereverberation of speech [4] are often implemented by multiplying the STFT of the signal by the Fourier transform of the filter impulse response for the current time interval and then reconstructing a signal from this MSTFT. In most applications where the STFT has been modified, this MSTFT is no longer valid in the sense that it is not the STFT of any signal. Consequently, for these applications, it is important to develop algorithms which estimate a signal with STFT close to a given MSTFT. In this chapter, we formulate this problem as a minimization of the mean squared error between the MSTFT and the STFT of the signal estimate.

2.2. Least Squares Error Estimation from the MSTFT

Let $x(n)$ and $X_w(mS,\omega)$ denote a real sequence and its STFT. The variable $S$ is a positive integer, which represents the sampling period of $X_w(n,\omega)$ in the variable $n$. Let the analysis window used in the STFT be denoted by $w(n)$, and with little loss of generality, $w(n)$
is assumed to be real, \( L \) points long, and non-zero for \( 0 \leq n \leq L - 1 \). From the definition of the STFT,

\[
X_w(mS,\omega) = F_l[x_w(mS,l)] = \sum_{l=-\infty}^{\infty} x_w(mS,l)e^{-j\omega l}
\]

where

\[
x_w(mS,l) = w(mS-l)x(l)
\]

and \( F_l[x_w(mS,l)] \) represents the Fourier transform of \( x_w(mS,l) \) with respect to the variable \( l \).

Let \( Y_w(mS,\omega) \) denote the given MSTFT and let \( y_w(mS,l) \) be given by

\[
y_w(mS,l) = \frac{1}{2\pi} \int_{-\pi}^{\pi} Y_w(mS,\omega)e^{j\omega l}d\omega
\]

An arbitrary \( Y_w(mS,\omega) \), in general, is not a valid STFT in the sense that there is no sequence whose STFT is given by \( Y_w(mS,\omega) \). In this section, we develop a new algorithm to estimate a sequence \( x(n) \) whose STFT \( X_w(mS,\omega) \) is closest to \( Y_w(mS,\omega) \) in the squared error sense.

Consider the following distance measure between \( x(n) \) and a given MSTFT \( Y_w(mS,\omega) \):

\[
D[x(n),Y_w(mS,\omega)] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} |X_w(mS,\omega) - Y_w(mS,\omega)|^2 d\omega
\]

The distance measure in Equation (2.4) is the squared error difference between \( X_w(mS,\omega) \) and \( Y_w(mS,\omega) \) integrated over all \( \omega \) and summed over all \( m \) and has been written as a function of \( x(n) \) and \( Y_w(mS,\omega) \) to emphasize that \( X_w(mS,\omega) \) is a valid STFT while \( Y_w(mS,\omega) \) is not necessarily a valid STFT. By Parseval's theorem, Equation (2.4) can be written as

\[
D[x(n),Y_w(mS,\omega)] = \sum_{m=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} [x_w(mS,l) - y_w(mS,l)]^2
\]

Since Equation (2.5) is quadratic in \( x(n) \), minimizing \( D[x(n),Y_w(mS,\omega)] \) with respect to \( x(n) \) leads to the following result:

\[
x(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n)y_w(mS,n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}
\]
This solution is similar in form to the standard overlap-add procedure [6,7], or the weighted overlap-add procedure [8,9]. The overlap-add procedure can be expressed as

$$x(n) = \sum_{m = -\infty}^{\infty} y_w(mS, n)$$

$$\sum_{m = -\infty}^{\infty} w(mS - n)$$

(2.7)

The weighted overlap-add procedure can be expressed as

$$x(n) = \sum_{m = -\infty}^{\infty} f(mS - n)y_w(mS, n)$$

(2.8)

for some “synthesis” filter $f(n)$. The major difference between Equations (2.6) and (2.7) is that Equation (2.6) specifies that $y_w(mS, n)$ should be windowed with the analysis window before being overlap-added and $w(mS - n)$ should be squared before summation over the variable $m$ for normalization. The difference between Equations (2.6) and (2.8) is that Equation (2.6) explicitly specifies what $f(n)$ is and has the normalization constant. In addition, the major difference between Equation (2.6), and Equations (2.7) and (2.8), is that Equation (2.6) was theoretically derived explicitly for the purpose of estimating a signal from the MSTFT while Equations (2.7) and (2.8) were derived to reconstruct a signal from its exact STFT and were sometimes used as an ad-hoc method to estimate a signal from the MSTFT. From the computational point of view, the differences cited above are minor in terms of both the number of arithmetic operations and the amount of on-line storage required. For example, Equation (2.6) can be implemented with little on-line storage and delay, in the same manner [9] as the standard overlap-add procedure of Equation (2.7) or the weighted overlap-add procedure of Equation (2.8). Since the algorithm represented by Equation (2.6) minimizes the distance measure of Equation (2.4), it will be referred to as LSEE-MSTFT, meaning least squares error estimation from the MSTFT.
In the standard overlap-add method, the window is usually normalized so that
\[
\sum_{m=-\infty}^{\infty} w(mS-n) \text{ is unity for all } n \text{ in order to reduce computation. As in the overlap-add method, the window in Equation (2.6) can be normalized so that } \sum_{m=-\infty}^{\infty} w^2(mS-n) \text{ is unity for all } n. \text{ Any non-zero window can be normalized in this manner for maximum window overlap } (S=1). \text{ For partial window overlap, however, the window is more restricted. Several windows which have this property for partial window overlap are discussed below.}

When the window shift \(S\) divides the window length \(L\) evenly, the rectangular window defined by

\[
w_r(n) = \begin{cases} 
\frac{\sqrt{S}}{\sqrt{L}} & , \quad 0 \leq n < L \\
0 & , \quad \text{otherwise}
\end{cases}
\) (2.9)

has the property

\[
\sum_{m=-\infty}^{\infty} w_r^2(mS-n) = \sum_{m=0}^{\frac{L}{S}-1} \frac{S}{L} = 1
\] (2.10)

We can further show with some algebra that if the window length \(L\) is a multiple of 4 times the window shift \(S\) then the sinusoidal window defined by:

\[
w_s(n) = \frac{2w_r(n)}{\sqrt{4a^2 + 2b^2}} \left[ a + b \cos \left( \frac{2\pi n}{L} + \phi \right) \right]
\) (2.11)

has the property given by Equation (2.10). In addition, we require that this class of sinusoidal windows be symmetric so that \(w(n) = w(L-1-n)\). This requirement can be satisfied by choosing \(\phi = \frac{\pi}{L}\). By choosing values for \(a\) and \(b\), windows similar to the Hamming window and the Hanning window can be obtained. Thus, the modified Hamming window used for the examples in Chapters 5, 6, and 7 will be defined as Equation (2.11) for \(a = .54, b = -.46, \) and
\[ \phi = \frac{\pi}{L}. \] The major difference between this definition and the standard definition of the Hamming window is that the period of the sine wave is \( L \) in the modified Hamming window as opposed to \( L-1 \) for the standard Hamming window. Similarly, a modified Hanning window can be defined as Equation (2.11) for \( a = .5, b = -.5, \) and \( \phi = \frac{\pi}{L} \). Use of these modified windows eliminates the need for normalizing by \( \sum_{m=-\infty}^{\infty} w^2(mS-n) \) in Equation (2.6) which reduces computation and/or storage requirements for partial window overlap.

Estimating \( x(n) \) based on Equation (2.6) minimizes the squared error between \( X_w(mS, \omega) \) and \( Y_w(mS, \omega) \) and therefore can be used directly to estimate a sequence from a MSTFT. As will be discussed in the next chapter, Equation (2.6) can also be used to develop an iterative algorithm that estimates a signal from the MSTFT.
Chapter 3

Signal Estimation from Modified STFT Magnitude

3.1. Introduction

In many applications where the STFT has been modified, it is desirable to estimate the signal from the modified STFT magnitude (MSTFTM). In these applications, either a better STFTM estimate is available or no STFT phase estimate is available. For example, in speech enhancement by spectral subtraction, the STFTM is modified to produce an estimate of the STFTM of the original speech and then combined with the STFT phase of the degraded speech. In this application, estimation directly from the MSTFTM would be preferable since only the STFT phase of the degraded signal is available. As another example, in time-scale modification of speech, the STFTM is time-scale modified to produce an estimate of the STFTM of time-scale modified speech. However, an estimate of the STFT phase of the time-scale modified speech is not easily produced and the time-scale modified speech must be estimated from the MSTFTM only. In most applications where the STFTM has been modified, this MSTFTM is no longer valid in the sense that it is not the STFTM of any signal. Consequently, for these applications, it is important to develop algorithms which estimate a signal with STFTM close to a given MSTFTM.

The methods for estimating signals with STFTM close to a given MSTFTM discussed in this chapter have been divided into three sections. The first section discusses an iterative algorithm based on the LSEE-MSTFT algorithm developed in Chapter 2. This algorithm is shown
to converge to a set consisting of the critical points of a distance measure between the given MSTFTM and the STFTM of the signal estimate. In the second section, other algorithms which minimize distance functions over all the windows of the MSTFTM simultaneously will be discussed. In the third section, algorithms which estimate the signal sequentially by minimizing a distance function over a single window at a time will be discussed.

3.2. Least Squares Error Estimation from MSTFTM

In this section, an algorithm will be developed, based on the LSEE-MSTFT algorithm of Chapter 2, which decreases the following distance measure in each iteration:

\[ D_1[x(n), Y_w(mS, \omega)] = \sum_{m=\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ |X_w(mS, \omega)| - |Y_w(mS, \omega)| \right]^2 d\omega \] (3.1)

Let \( x^i(n) \) denote the estimated \( x(n) \) after the \( i \)th iteration. The \( i+1 \) estimate \( x^{i+1}(n) \) is obtained by taking the STFT of \( x^i(n) \), replacing the magnitude of \( X_w^i(mS, \omega) \) with the given magnitude \( |Y_w(mS, \omega)| \) and then finding the signal with STFT closest to this modified STFT using the LSEE-MSTFT algorithm. The iterative algorithm, which is illustrated in Figure 3.1, results in the following update equation:

\[
x^{i+1}(n) = \frac{\sum_{m=\infty}^{\infty} w(mS-n) \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}_w^i(mS, \omega)e^{jn\omega} d\omega}{\sum_{m=\infty}^{\infty} w^2(mS-n)}
\] (3.2)

where

\[
\hat{X}_w^i(mS, \omega) = |Y_w(mS, \omega)| \frac{X_w^i(mS, \omega)}{|X_w^i(mS, \omega)|}
\] (3.3)

This algorithm can be shown (See Appendix A) to converge to a set consisting of the critical points of the simultaneous distance function \( D_1 \).

It is possible to develop ad-hoc methods to estimate \( x(n) \) from the MSTFTM by modify-
Given \(|Y_w(mS, \omega)|\)

Initial Estimate of \(x(n)\)

\(x^i(n)\)

Substitution of Given STFTM (\(|Y_w(mS, \omega)|\)):

\(\hat{X}_w^i(mS, \omega) = |Y_w(mS, \omega)| \cdot e^{j\omega t} X_w^i(mS, \omega)\)

\(\hat{X}_w^i(mS, \omega)\)

Signal Estimation from \(\hat{X}_w^i(mS, \omega)\):

\[
x^{i+1}(n) = m = -\infty \sum_{m = -\infty}^{\infty} w(mS - n) \cdot \hat{X}_w^i(mS, n)
\]

\[
= \frac{1}{m = -\infty \sum_{m = -\infty}^{\infty} w(mS - n)} \sum_{m = -\infty}^{\infty} w(mS - n)
\]

\(x^{i+1}(n)\)

Figure 3.1 - LSEE-MSTFTM Algorithm

The iterative algorithm in Figure 3.1. For example, suppose we use in one step of the iterative procedure the standard overlap-add method rather than the LSEE-MSTFT method in obtaining the next estimate \(x^{i+1}(n)\) from the MSTFT \(\hat{X}_w^i(mS, \omega)\). This results in the following update equation:

\[
x^{i+1}(n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}_w^i(mS, \omega) e^{j\omega t} d\omega
\]  

where \(\hat{X}_w^i(mS, \omega)\) is given by Equation (3.3). This algorithm will be called OA(overlap-add)-MSTFTM to distinguish it from the LSEE-MSTFTM algorithm. Although OA-MSTFTM requires fewer multiplications per iteration since one less windowing step is required, it is not
guaranteed to converge to the critical points of $D_1$ because Equation (3.4) is only an approximation of the minimum given by Equation (3.2). However, as will be shown in Chapters 5, 6, and 7, OA-MSTFTM does appear to reduce the distance function $D_1$ enough to produce a reasonable signal estimate for the applications considered in this thesis.

One major advantage of the LSEE-MSTFTM and OA-MSTFTM algorithms is the possibility of real time implementation with cascaded processors. Figure 3.2 shows an implementation of the LSEE-MSTFTM algorithm where the $i^{th}$ processor performs the $i^{th}$ iteration of the algorithm. In practice, the Fourier transforms in Equations (3.2) and (3.3) are replaced by Discrete Fourier Transforms (DFTs). To indicate this, the continuous variable $\omega$ has been replaced by the integer variable $k$ in the diagram. The sequence of operations performed by the $i^{th}$ processor is as follows:

```
Figure 3.2 - $i^{th}$ Processor of Cascade
```
1) The next $S$ points of the signal estimate $x^i(n)$ are shifted into the input buffer and simultaneously, the next $S$ points of the signal estimate $x^{i+1}(n)$ are shifted out of the output buffer and into the input buffer of the $i + 1^{st}$ processor and the first $S$ points of the output buffer are zeroed. These buffers are both $L$ points long so that a full window of data can be stored. At the same time, the magnitude for a single window of the MSTFT $|Y_w(m_0S - iL,k)|$ is stored in the first magnitude buffer and all the magnitude buffers are shifted into the next magnitude buffer. The last magnitude buffer is shifted into the next processor. Each of the magnitude buffers must be at least $K/2+1$ points long where $K$ is the length of the FFT to store all of the magnitude points.

2) The $L$ points of the input buffer are windowed and transformed to the Fourier domain with an FFT.

3) The magnitude of this FFT is then replaced by the magnitude in the first magnitude buffer.

4) The inverse FFT is then calculated and the result is windowed and added to the output buffer.

The major computational requirements of this procedure are the two FFTs and the magnitude replacement operation. In order to generate the output sequence in real time, each processor must perform these operations during the time interval $ST$, where $S$ is the number of points per window shift, and $T$ is the sampling interval. The delay between the output signal estimate and the input signal estimate introduced by each processor is equal to $LT$, where $L$ is the length of the output buffer. The cause of this delay can easily be seen by looking at the $i^{th}$ processor when the magnitude replacement is an identity operation (the FT magnitude of the current window is the desired FT magnitude). For this case, a given $S$ point subsequence of the input sequence $x^i(n)$ enters the first $S$ points of the input buffer at time interval 1. After
the window, FFT, magnitude replacement, IFFT, window, and add operations are completed, a windowed version of these $S$ points resides in the first $S$ points of the output buffer. In time interval $L/S$, these $S$ points of the input sequence reside in the last $S$ points of the input buffer. After the same operations are performed as in the first time interval, the last $S$ points of the output buffer will contain exactly the same $S$ points that entered the input buffer at time interval 1. At the beginning of time interval $L/S + 1$, these $S$ points are then shifted into the input buffer of the $i+1^{st}$ processor. Consequently, each processor delays the signal by $L/S$ time intervals multiplied by $ST$ seconds per time interval or $LT$ seconds. Since each processor introduces a delay of $L/S$ time intervals, $L/S$ magnitude buffers are required to keep the signal estimate synchronized with the desired magnitudes.

This implementation can be easily modified to suit the OA-MSTFTM algorithm by increasing the length of the output buffer to $K$ points where $K$ is the FFT length. This increases the delay of each processor to $K/S$ time intervals and consequently increases the required number of magnitude buffers to $K/S$. This tends to make OA-MSTFTM less desirable from an implementational point of view since more delay is introduced per iteration, and more memory is required in each processor.

3.3. Other Simultaneous Estimation Methods

In the following two sections, two standard minimization techniques, steepest descent and conjugate gradient, will be developed for decreasing a simultaneous distance function of the form:

$$D_p[x(n), |Y_w(mS, \omega)|] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ |X_w(mS, \omega)|^p - |Y_w(mS, \omega)|^p \right]^2 d\omega \quad (3.5)$$

For maximum window overlap ($S=1$), Nawab has shown [10] that if a sequence has fewer than $L-1$ consecutive zero samples between any two non-zero samples, then $x(n)$ can be
exactly reconstructed from its STFT magnitude $|X_{w}(mS,\omega)|$ to within a sign ambiguity. If $x(n)$ has $L-1$ or more consecutive zero samples, we can break it up into subsequences that have fewer than $L-1$ consecutive zero samples and reconstruct each subsequence to within a sign ambiguity. Thus, for maximum window overlap the distance measure $D_p$ can have up to $2^M$ global minima where $M$ is the number of subsequences. For speech, if the window is long enough so that sign changes following $L-1$ or more consecutive zeros are not perceived by the ear, then convergence to any one of these global minima will be considered to be acceptable performance of the algorithm. In addition to these global minima, $D_p$ can possess a multitude of local minima to which any of the simultaneous minimization algorithms could converge. Thus, the only way we can be assured of converging to a global minimum is to start with an initial estimate close to the global minimum or minimize distance functions that have only global minima.

As shown in the previous section, the LSEE-MSTFTM algorithm converges to a set consisting of the critical points of the distance function of Equation (3.5) for $p=1$. In order to decrease distance functions of the form given by Equation (3.5) using the steepest descent and conjugate gradient methods, it is necessary for the required line search to be computationally tractable. In order to meet this requirement, $p=2$ was chosen so that the line search reduces to rooting a cubic equation. The following two sections discuss the derivation and implementation of the steepest descent and conjugate gradient simultaneous estimation algorithms.

3.3.1. Simultaneous Steepest Descent

The method of steepest descent obtains a new estimate by searching for a minimum along a line in the direction of the negative of the gradient evaluated at the current estimate. Thus, if we define $x^t(n)$ as the current estimate of $x(n)$, then the update equation for the new esti-
mate \( (x^{i+1}(n)) \) can be written as:

\[
x^{i+1}(n) = x^i(n) - \alpha g^i(n)
\]  

(3.6)

The sequence \( g^i(n) \) is the gradient of the distance function \( D_2 \) evaluated at \( x^i(n) \) and can be written as:

\[
g^i(n) = 4 \sum_{m=-\infty}^{\infty} w(mS-n) \hat{g}^i(mS,n)
\]  

(3.7)

\[
\hat{g}^i(mS,n) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| X^i_m(mS,\omega) \right|^2 - \left| Y^i_m(mS,\omega) \right|^2 X^i_m(mS,\omega) e^{j\omega n} d\omega
\]  

(3.8)

From Equation (3.7), we see that the gradient \( g^i(n) \) can be formed by a window, overlap and add of the gradients for the individual windows \( \hat{g}^i(mS,n) \). The gradient for the \( m^{th} \) window, \( \hat{g}^i(mS,n) \), can be computed using FFTs in memory on the order of the window length \( (L) \). Thus, the gradient can be formed on off-line storage since all of the samples aren't computed simultaneously.

The parameter \( \alpha \) is chosen to minimize

\[
D_2[x^i(n) - \alpha g^i(n) , |Y^i_m(mS,\omega)|]
\]  

(3.9)

Since the distance function \( D_2 \) along a straight line is a one dimensional quartic equation, the directional derivative is a one dimensional cubic which can be rooted analytically to find the minimum of \( D_2 \) along the line. This is the prime motivation for choosing \( p = 2 \) as opposed to some other value for the distance function of Equation (3.5). The directional derivative \( (D') \) at the point \( x^{i+1}(n) \) in the direction \( -g^i(n) \) is:

\[
D' = 4 \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ -\alpha^3 |G^i_m(mS,\omega)|^4 + \alpha^2 |G^i_m(mS,\omega)|^2 G^i_m(mS,\omega)X^i_m(mS,\omega)
\]

\[
- \alpha \left[ |G^i_m(mS,\omega)|^2 \left( 2 |X^i_m(mS,\omega)|^2 - |Y^i_m(mS,\omega)|^2 \right) + \left( G^i_m(mS,\omega)X^i_m(mS,\omega) \right)^2 \right]
\]  

(3.10)
\[ + G_w(mS, \omega) R_w(mS, \omega) \left[ |X_w(mS, \omega)|^2 - |Y_w(mS, \omega)|^2 \right] d\omega \]

The coefficients of Equation (3.10) can be determined using FFTs and memory on the order of the window length. Once the roots are determined, the minimum can be determined by substituting the roots for \( \alpha \) in the following equation:

\[ \bar{D} = a_3 \alpha^4 + a_2 \alpha^3 + a_1 \alpha^2 + a_0 \alpha \]  
(3.11)

Where the coefficients \( (a_n) \) are the same as in Equation (3.10) for \( \alpha'' \). For this equation, the current estimate yields a distance \( \bar{D} \) of zero and a decrease in the distance function yields a negative value. Thus, the root that is chosen should yield a negative value in order to ensure that the distance function is decreased. Since up to two minima of \( \bar{D} \) may exist, there are several ways to choose which one to use. The most obvious method is to use the root which produces the smallest value of \( \bar{D} \). However, if we start the iteration with certain initial conditions, we may want to converge to the minima nearest the initial conditions. If this is the case, we should choose the root which yields a negative \( \bar{D} \) and is smallest in absolute value. This will prevent the algorithm from jumping from the vicinity of one relative minimum to the vicinity of another relative minimum. This leads to the following algorithm:

1) Calculate the gradient using Equations (3.7) and (3.8).

2) Find the coefficients of Equation (3.10).

3) Choose alpha as the root of Equation (3.10) which yields a negative \( \bar{D} \) in Equation (3.11) and is smallest in absolute value.

4) Update the signal estimate using Equation (3.6).

5) Repeat steps 1) - 4) until the distance has been reduced to the desired level.
3.3.2. Simultaneous Conjugate Gradient

Since the steepest descent algorithm tends to converge slowly near the solution point, we would like to develop an algorithm which converges faster without requiring unreasonable quantities of memory or computation. The conjugate gradient algorithm introduces information about the Hessian (second order derivatives) which improves the convergence rate and doesn't require the Hessian matrix to be inverted as in Newton's method. The first step of this algorithm is identical to a step of the simultaneous steepest descent algorithm of Section 3.3.1. Thereafter, the new direction is chosen as a linear combination of the current gradient and the previous direction. The new estimate is then obtained by searching for a minimum in this direction. The direction update can be written as [13]:

\[
    d^i(n) = -g^i(n) + \beta d^{i-1}(n) \tag{3.12}
\]

\[
    \beta = \frac{\sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} g^i(k) H^{i-1}(k,l) d^{i-1}(l)}{\sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} d^{i-1}(k) H^{i-1}(k,l) d^{i-1}(l)} \tag{3.13}
\]

Where \( H^{i-1}(k,l) \) is the Hessian evaluated at \( x^{-1}(n) \) which is defined as:

\[
    H^{i-1}(k,l) = \left. \frac{\partial^2 D_2[x(n), \{Y_w(mS, \omega)\}]}{\partial x(k) \partial x(l)} \right|_{x(n) = x^{-1}(n)} \tag{3.14}
\]

By substituting Equation (3.14) for \( H^{i-1}(k,l) \) in Equation (3.13) and evaluating \( H_{i-1}(k,l) \), the following expression for \( \beta \) can be obtained:

\[
    \beta = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \text{Re} \left[ G_w^i(mS, \omega) X_w^{-1\ast}(mS, \omega) \right] \text{Re} \left[ D_w^{i-1}(mS, \omega) X_w^{-1\ast}(mS, \omega) \right] 
    + G_w^i(mS, \omega) D_w^{i-1\ast}(mS, \omega) [\text{Re} X_w^{-1}(mS, \omega) |^2 - \text{Im} Y_w(mS, \omega) |^2] \right\} d\omega

\[
    \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ 2 \left( \text{Re} \left[ D_w^{i-1}(mS, \omega) X_w^{-1\ast}(mS, \omega) \right] \right)^2 \right\}
\]
\[ + |D^{i-1}_w(mS, \omega)|^2 \left( |X^{i-1}_w(mS, \omega)|^2 - |Y_w(mS, \omega)|^2 \right) \right] d\omega \]

The signal estimate is then updated using the following equation:

\[ x^{i+1}(n) = x^n(n) + \alpha d^i(n) \]  \hspace{1cm} (3.16)

Where \( \alpha \) is chosen as the appropriate root of the directional derivative:

\[ \bar{D}^i = 4 \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} \left( \alpha^3 |D_w(mS, \omega)|^4 + \alpha^2 |D_w(mS, \omega)|^2 \left[ D_w^e(mS, \omega) X_w^i(mS, \omega) \right] \\
+ \alpha \left[ |D_w^i(mS, \omega)|^2 \left( 2 |X_w^i(mS, \omega)|^2 - |Y_w(mS, \omega)|^2 \right) + \left( D_w^e(mS, \omega) X_w^i(mS, \omega) \right)^2 \right] + D_w^e(mS, \omega) X_w^i(mS, \omega) \left[ |X_w^i(mS, \omega)|^2 - |Y_w(mS, \omega)|^2 \right] \right) d\omega \]  \hspace{1cm} (3.17)

This leads to the following algorithm:

1) Perform simultaneous steepest descent step of Section 3.3.1. Set \( d^0(n) = -g^0(n) \).
2) Calculate the gradient using Equation (3.7).
3) Calculate \( \beta \) using Equation (3.15).
4) Update the direction using Equation (3.12).
5) Find the coefficients of Equation (3.17).
6) Choose \( \alpha \) as the root of Equation (3.17) which yields a negative \( \bar{D} \) in Equation (3.11) and is smallest in absolute value.
7) Update the signal estimate using Equation (3.16).
8) Repeat steps 1) - 7) up to \( N \) times, where \( N \) is the length of the sequence \( x(n) \).
9) Repeat steps 1) - 8) until the distance has been reduced to the desired level.

As discussed in Luenberger [13], the conjugate gradient algorithm should be restarted
with a steepest descent step every $K$ steps, where $K$ is the dimension of the problem. The dimension of this problem is $N$, the length of the sequence $x(n)$, as specified in step 8) above. A line search is performed to find $\alpha$ rather than computing $\alpha$ from the gradient vector, direction vector, and Hessian as in Luenberger since a line search is relatively cheap computationally and since the method proposed in Luenberger is not guaranteed to be globally convergent for this non-quadratic distance function. When computing the gradient, cubic equation coefficients, and $\beta$ for the simultaneous steepest descent or simultaneous conjugate gradient algorithms, the FFT length should be at least $2L - 1$ where $L$ is the window length to prevent aliasing of the convolution of two $L$ point sequences performed by the FFT.

3.4. Sequential Estimation Methods

Sequential estimation methods break the problem of estimating a signal with STFTM close to a given MSTFTM into two parts in order to reduce the dimension of the minimization problem. First, estimates for individual windows of the MSTFTM are produced. Then, these estimates are combined in some manner to produce the estimate of the entire signal. Although this approach isn't guaranteed to minimize a distance measure over all of the windows simultaneously, some of these methods obtain estimates comparable in quality to the simultaneous methods described earlier in this chapter. In Section 3.4.1, methods for estimating a signal with Fourier Transform Magnitude (FTM) for a single window close to a given modified Fourier Transform Magnitude (MFTM) for a single window will be developed. In Section 3.4.2, methods for combining these estimates for individual windows into an estimate of the entire signal will be discussed.
3.4.1. Signal Estimation for a Single Window from MFTM

In this section algorithms for estimating a signal over a single window from MFTM will be developed. These algorithms decrease the following distance measure

\[ D_p [x_w (m_0 S, n), Y_w (m_0 S, \omega)] = \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \left( |X_w (m_0 S, \omega)|^p - |Y_w (m_0 S, \omega)|^p \right)^2 d\omega \]  

(3.18)

where \( p \) takes on values of one or two. For distance measures of this form, we can show that a unique global minimum doesn’t always exist. Suppose \( \hat{x}_w (m_0 S, n) \) is a relative minimum of the distance function such that no other \( x_w (m_0 S, n) \) has a smaller distance for a given \( |Y_w (m_0 S, \omega)| \). Since \( \hat{x}_w (m_0 S, n) \) is a sequence of length \( L \), \( \hat{X}_w (m_0 S, x) \) can be factored into no more than \( L - 1 \) zeros. As many as \( 2^{L-1} \) sequences can be formed with the same \( z \)-transform magnitude by replacing the zeros of \( \hat{X}_w (m_0 S, x) \) by their conjugate reciprocals. Further, the negative of a sequence also has the same magnitude. Thus, there can be up to \( 2^L \) sequences with the same magnitude. So, Equation (3.18) can have up to \( 2^L \) relative minima with the same distance measure.

In order to ensure that a unique global minimum exists, we must add some constraints to the problem. One method of adding constraints is to fix some of the values of \( x_w (m_0 S, n) \) during the minimization. Nawab has shown [10] that if the first \( \left\lfloor \frac{L}{2} \right\rfloor \) or more points are fixed* and \( \hat{x}_w (m_0 S, m_0 S - L + 1) \) is nonzero, then only one sequence exists with the magnitude \( |\hat{X}_w (m_0 S, \omega)| \) which implies that \( \hat{x}_w (m_0 S, n) \) must be a unique global minimum. If \( \hat{x}_w (m_0 S, m_0 S - L + 1) \) is zero, the equivalent window length \( (L') \) is the window length \( (L) \) minus the number of consecutive zeros starting at \( \hat{x}_w (m_0 S, m_0 S - L + 1) \). For this case, the first \( \left\lfloor \frac{L'}{2} \right\rfloor \) or more points must be fixed.

* \( \lfloor x \rfloor \) is defined as the least integer greater than or equal to \( x \)
To minimize $D_2[x_w(m_0S,n), |Y_w(m_0S,\omega)|]$ for a given $|Y_w(m_0S,\omega)|$, standard minimization methods such as steepest descent and conjugate gradient can be employed. The primary advantage to using $p=2$ in the distance function of Equation (3.18) for these methods is that the line search reduces to rooting a cubic equation which can be performed analytically reducing the computational burden substantially. These two methods will be developed in the following two sections. A third method, proposed by Nawab, which minimizes the distance function $D_1[x_w(m_0S,n), |Y_w(m_0S,\omega)|]$ will be discussed in Section 3.4.1.3.

3.4.1.1. Sequential Steepest Descent

The derivation of the sequential steepest descent algorithm for distance function $D_2$ is very similar to the simultaneous steepest descent algorithm presented in Section 3.3.1. Thus, if we define $z^l(n)$ as the current estimate of $x_w(m_0S,n)$ with the first $M$ samples fixed, then the new estimate $(z^{l+1}(n))$ can be written as:

$$z^{l+1}(n) = z^l(n) - \alpha g^l(n)$$  \hspace{1cm} (3.19)

Where $\alpha$ is chosen to minimize $D_2[z^l(n) - \alpha g^l(n), |Y_w(m_0S,\omega)|]$. The sequence $g^l(n)$ is the gradient of the distance function evaluated at $z^l(n)$ which can be written as:

$$g^l(n) = \begin{cases} \frac{1}{2\pi} \int_{-\pi}^{\pi} 4 \left[ |Z^l(\omega)|^2 - |Y_w(m_0S,\omega)|^2 \right] Z^l(\omega) e^{jw_n} d\omega, & m_0S - L + M + 1 \leq n \leq m_0S \\ 0, & \text{otherwise} \end{cases}$$  \hspace{1cm} (3.20)

By expressing the gradient in the Fourier domain, the convolutions in the time domain are replaced by multiplies. Consequently, the gradient can be efficiently computed in the Fourier domain and then transformed back into the time domain and zeroed for the appropriate values of $n$. To find $\alpha$, we must minimize the distance function along the line $z^{l+1}(n) = z^l(n) - \alpha g^l(n)$. Since the distance function along a straight line is a one dimensional
quartic, the directional derivative will be a one dimensional cubic which can be rooted analytically to find the minimum. This is the motivation for choosing $p = 2$ as opposed to some other value for the distance measure of Equation (3.18). The directional derivative $(\bar{D}')$ at the point $z^{l+1}(n)$ in the direction $-g^l(n)$ is:

\[
\bar{D}' = \sum_{n=-\infty}^{\infty} g^l(n) g^{l+1}(n) \tag{3.21}
\]

By substituting Equation (3.20) for $g^{l+1}(n)$, we obtain:

\[
\bar{D}' = \frac{1}{2\pi} \int_{-\pi}^{\pi} 4 \left[ |Z^{l+1}(\omega)|^2 - |Y^c (m_0 S, \omega)|^2 \right] G^l(\omega) Z^{l+1*}(\omega) d\omega \tag{3.22}
\]

Finally, by substituting the Fourier domain equivalent of Equation (3.19) for $Z^{l+1}(\omega)$ we obtain a cubic in $\alpha$:

\[
\bar{D}' = 4 \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{-\alpha^3 |G^l(\omega)|^4 + \alpha^2 |G^l(\omega)|^2 G^l(\omega) Z^m(\omega) \right. \\
- \alpha \left[ (G^l(\omega) Z^m(\omega))^2 + |G^l(\omega)|^2 \left( 2 |Z^l(\omega)|^2 - |Y^c (m_0 S, \omega)|^2 \right) \right] \\
+ \left( |Z^l(\omega)|^2 - |Y^c (m_0 S, \omega)|^2 \right) G^l(\omega) Z^m(\omega) \right\} d\omega \tag{3.23}
\]

Once the roots of Equation (3.23) are found, the minimum can be determined by substituting the roots for $\alpha$ in the following equation:

\[
\bar{D} = a_3 \alpha^4 + a_2 \alpha^3 + a_1 \alpha^2 + a_0 \alpha \tag{3.24}
\]

Where the coefficients ($a_n$) are the same as in Equation (3.23) for $\alpha^n$. As in the case of simultaneous steepest descent, $\alpha$ is chosen as the root which yields a negative $\bar{D}$ and is smallest in absolute value. This leads to the following algorithm:

1) Calculate $g^l(n)$ using Equation (3.20).

2) Find the coefficients of Equation (3.23).
3) Find the roots of Equation (3.23).

4) Choose $\alpha$ as the root which yields a negative $D$ in Equation (3.24) and is smallest in absolute value. If no root yields a negative $D$, then quit (We can't get any closer to the minimum due to round-off errors).

5) Update the signal estimate using Equation (3.19).

6) Repeat steps 1) - 5) until the distance has been reduced to the desired level.

3.4.1.2. Sequential Conjugate Gradient

Since the steepest descent algorithm tends to have a slow convergence rate near the solution, we would like to develop a conjugate gradient method for the sequential estimation problem. Following the simultaneous conjugate gradient derivation of Section 3.3.2, the direction update can be written as:

$$d^i(n) = -g^i(n) + \beta d^{i-1}(n)$$  \hspace{1cm} (3.25)

$$\beta = \frac{\sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} g^i(k)H^{i-1}(k,l)d^{i-1}(l)}{\sum_{k=-\infty}^{\infty} \sum_{l=-\infty}^{\infty} d^{i-1}(k)H^{i-1}(k,l)d^{i-1}(l)}$$  \hspace{1cm} (3.26)

Where $H^{i-1}(k,l)$ is the Hessian evaluated at $z^{i-1}(n)$ which is defined as:

$$H^{i-1}(k,l) = \left. \frac{\partial^2 D[z(n), Y_w(m_0S, \omega)]}{\partial z(k) \partial z(l)} \right|_{z(n)=z^{i-1}(n)}$$  \hspace{1cm} (3.27)

$$= \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ (Z^{i-1}(\omega))^2 e^{j\omega(k+l)} + 2|Z^{i-1}(\omega)|^2 - |Y_w(m_0S, \omega)|^2 \right] e^{j\omega(k-l)} d\omega,$$

$$m_0S - L + M + 1 \leq k, l \leq m_0S$$

0, otherwise

By substituting Equation (3.27) for $H^{i-1}(k,l)$ in Equation (3.26) the following expression for
β can be obtained:

\[
\beta = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ 2 \text{Re} \left[ G'(\omega)Z^{l-1*}(\omega) \right] \text{Re} \left[ D^{l-1}(\omega)Z^{l-1*}(\omega) \right] \\
+ G'(\omega)D^{l-1*}(\omega) \left[ |Z^{l-1}(\omega)|^2 - |y_w(m_0,\omega)|^2 \right] \right\} d\omega
\]

\[
\frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ 2 \left( \text{Re} \left[ D^{l-1}(\omega)Z^{l-1*}(\omega) \right] \right)^2 \\
+ |D^{l-1}(\omega)|^2 |Z^{l-1}(\omega)|^2 - |y_w(m_0,\omega)|^2 \right\} d\omega
\]

The line search required to find α is the same as in Section 3.4.1.1 except that we search in the direction \( d^l(n) \) rather than the negative gradient. Thus, the problem reduces to finding the roots of the following directional derivative:

\[
\overline{D}' = 4 \frac{1}{2\pi} \int_{-\pi}^{\pi} \left\{ \alpha^3 |D'(\omega)|^4 + \alpha^2 |D'(\omega)|^2 D'(\omega)Z'^*(\omega) \\
+ \alpha \left[ (D'(\omega)Z'^*(\omega))^2 + |D'(\omega)|^2 \left( 2 |Z'(\omega)|^2 - |y_w(m_0,\omega)|^2 \right) \right] \\
+ D'(\omega)Z'^*(\omega) \left( |Z'(\omega)|^2 - |y_w(m_0,\omega)|^2 \right) \right\} d\omega
\]

The parameter α is chosen as the appropriate root as in Section 3.4.1.1 and the signal estimate is updated as follows:

\[
z^{l+1}(n) = z^l(n) + \alpha d^l(n)
\]

This leads to the following iterative algorithm:

1) Perform a steepest descent step of Section 3.4.1.1. Set \( d^0(n) = -g^0(n) \).

2) Calculate the gradient using Equation (3.20).

3) Calculate β using Equation (3.28).
4) Update the direction using Equation (3.25).

5) Find the coefficients of Equation (3.29).

6) Choose $\alpha$ as in step 4) of the steepest descent algorithm of Section 3.4.1.1 using the coefficients of Equation (3.29) in Equation (3.24).

7) Update the signal estimate using Equation (3.30).

8) Repeat steps 2) - 7) up to $L - M$ times.

9) Repeat steps 1) - 3) until the distance has been reduced to the desired level.

As discussed in Luenberger [13], the conjugate gradient algorithm should be restarted with a steepest descent step every $K$ steps, where $K$ is the dimension of the problem. The dimension of this problem is the window length ($L$) minus the number of points fixed ($M$) as specified in step 8) above.

3.4.1.3. Nawab's Algorithm

This minimization method was developed by Nawab [10] for the sequential estimation problem from STFTM. In this algorithm, the next signal estimate for a given window $x_{w}^{l+1}(m_0S, n)$ is obtained by setting the Fourier transform magnitude of the current estimate $x_{w}^{l}(m_0S, n)$ to the given magnitude $|Y_{w}(m_0S, \omega)|$, taking the inverse Fourier transform, and then enforcing the time domain constraints. The time domain constraints are enforced by setting the first $M$ points to the values estimated from previous windows and by setting the points from the inverse FFT not covered by the window to zero. This iterative algorithm results in the following update equation:

$$
\begin{align*}
    x_{w}^{l+1}(m_0S, n) = \begin{cases} 
        \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{X}_{w}^{l}(m_0S, \omega)e^{j\omega n} d\omega, & m_0S - L + M < n \leq m_0S \\
        c(n), & \text{otherwise}
    \end{cases}
\end{align*}
$$

(3.31)
where

\[ \hat{X}_n^x(m_0S, \omega) = \frac{Y_n(m_0S, \omega) \cdot \frac{X_n^x(m_0S, \omega)}{|X_n^x(m_0S, \omega)|}} \] (3.32)

and \( c(n) \) consists of the points estimated from previous windows for \( m_0S - L < n \leq m_0S - L + M \) and is zero for all other values of \( n \). This algorithm can be shown (see Appendix B) to decrease the distance measure of Equation (3.18) for \( p = 1 \) \( (D_1) \) in each iteration and converges to a set consisting of the critical points of the distance measure \( D_1 \) subject to the constraints \( c(n) \).

3.4.2. Signal Estimation using Estimates from MFTM Algorithms

In this section, we discuss two methods for combining the estimates for individual windows produced by the MFTM algorithms to generate an estimate of the entire signal. The first method, developed by Nawab [10], constructs the entire signal by piecing together the estimated points from each window in a manner similar to the overlap-save method [14] for constructing a signal. The second method treats the signal estimate for each window as an estimate of the phase for that window which is then combined with the given MSTFTM to produce a MSTFT. The entire signal is then estimated from the MSTFT using the LSEE-MSTFT algorithm developed in Chapter 2.

Nawab's method produces the entire signal estimate by starting with \( L - S \) initial data points assumed known where \( L \) is the window length and \( S \) is the window shift. Then, the remaining \( S \) points covered by the current window are estimated using one of the MFTM algorithms discussed in Section 3.4.1. The window shift \( S \) is chosen so that the number of points fixed, \( L - S \), is greater than \( \left\lceil \frac{L}{2} \right\rceil \) so that a unique minimum exists when the signal is non-zero.

Once an estimate for these \( S \) points has been obtained, they are fixed and become part of the
$L - S$ points used as constraints during estimation of the last $S$ points covered by the next window. This procedure can be expressed as:

$$x^n(n) = x^{n-1}(n) + a(mS - n)MFTM_S \left( b(mS - n)x^{n-1}(n), |Y_w(mS, \omega)| \right)$$  \hspace{1cm} (3.33)

where

$$b(n) = \begin{cases} w(n) & S \leq n < L \\ 0 & \text{else} \end{cases}$$

$$a(n) = \begin{cases} \frac{1}{w(n)} & 0 \leq n < S \\ 0 & \text{else} \end{cases}$$

$$x^0(n) = u(-n)x(n)$$

The sequence $x^0(n)$ corresponds to the points which must be known to start the algorithm.

The notation $MFTM_S \left( b(mS - n)x^{n-1}(n), |Y_w(mS, \omega)| \right)$ represents the output of one of the MFTM algorithms where $b(mS - n)x^{n-1}(n)$ corresponds to the constraints and the $S$ points of the initial estimate and $|Y_w(mS, \omega)|$ is the given MFTM for the $m^{th}$ window. The output includes both the constraints and the new signal estimate for the last $S$ points of the $m^{th}$ window. The sequence $x^n(n)$ is obtained by placing these points adjacent to those previously estimated by multiplying by $a(mS - n)$ to divide the window out and select the $S$ estimated points and then adding this to $x^{n-1}(n)$. Since the window is divided out, rectangular windows are usually used for this procedure. This method will be called the overlap-save method to distinguish it from the following method.

The entire signal can also be estimated by regarding the result of the estimation process for individual windows as an estimate of the STFT phase. This STFT phase is then combined with the MSTFTM to produce a MSTFT. The signal estimate can then be generated from this MSTFT using the LSEE-MSTFT algorithm developed in Chapter 2. Thus, the estimation process proceeds by starting with $M$ data points assumed known. Then, the $L - M$ remaining
points for the current window are estimated using one of the MFTM algorithms. The phase of this estimate is then combined with the given MFTM and the inverse FT is then windowed with the analysis window and added to the estimate obtained so far. Then, the next $S$ points are fixed where $S$ is the window shift and used as part of the first $M$ points assumed known in the next window. This procedure can be expressed as:

$$x^m(n) = x^{m-1}(n) + w(mS-n) \frac{1}{2\pi} \int_{-\pi}^{\pi} |Y_w(mS,\omega)| \frac{Z^m(\omega)}{|Z^m(\omega)|} e^{j\omega n} d\omega \quad (3.34)$$

$$x^m(n) = \text{MFTM}_{L-M}(v(mS-n)x^{m-1}(n), |Y_w(mS,\omega)|)$$

$$w(n) = \frac{L}{S} - 1 \sum_{m=1}^{L/M} w^2(n-mS)$$

$$v(n) = \begin{cases} \frac{L}{S} - 1 \sum_{m=1}^{L/M} w^2(n-mS) & L-M \leq n < L \\ 0 & \text{else} \end{cases}$$

$$x^0(n) = x(n) \sum_{m=-\infty}^{0} w^2(mS-n)$$

The sequence $x^0(n)$ corresponds to the points which must be known to start the algorithm windowed by an appropriate function to compensate for the effect of $v(n)$. The sequence $v(n)$ is chosen to undo the effects of the windows from previous estimates. One advantage of this method is that the first $M$ points in the current window instead of $L-S$ points are fixed. This allows $M$ to be smaller than $L-S$ so that only points that are the result of the addition of the estimates of several previous windows are fixed. So, none of the fixed points are the result of a single estimate near the ends of the window which tend to be worse estimates. Consequently, this method is expected to perform somewhat better than the overlap-save method.
Chapter 4

Criteria for Performance Comparison

4.1. Introduction

In this chapter, the criteria for comparison of the signal estimation algorithms presented in Chapter 2 and Chapter 3 are discussed. In the next section, a table containing an abbreviation and a short description for each of these algorithms will be presented. In the following section, the criteria for comparison of these algorithms for the applications of reconstruction from unmodified STFT, speech enhancement by spectral subtraction, and speed transformation of speech will be discussed.

4.2. Algorithms

In Table 4.1, an abbreviation is given for each of the methods discussed in Chapters 2 and 3 with a short description of each algorithm. These abbreviations will be used to refer to the algorithms in later chapters.

4.3. Comparison Criteria

The criteria used for comparison of the algorithms shown in Table 4.1 consist of both objective and subjective criteria. The objective criteria will consist of the two simultaneous STFTM distance measures $D_1$ and $D_2$ given by Equation (3.18) for $p=1$ and $p=2$ respectively. These distance measures were chosen since they measure the error over all of the windows of the STFTM and because several of the simultaneous methods are guaranteed to converge to the critical points of these distance measures. The subjective criteria will consist of comparison
of the waveforms and spectrograms of the estimates produced by each of the algorithms and informal listening comparisons.
<table>
<thead>
<tr>
<th>Description</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate phase from the MISTF using ISEE-MISTF. Phase for each frame of the MISTF. Combine estimated phase with magnitude to form MISTF.</td>
<td>ISEE-MISTF and Newb's Algorithm. Use Newb's iterative algorithm to estimate</td>
</tr>
<tr>
<td>Estimate phase from the MISTF using ISEE-MISTF. For each frame of the MISTF, combine estimated phase with magnitude to form MISTF.</td>
<td>ISEE-CC and Conjugate Gradien. Use conjugate gradient to estimate phase.</td>
</tr>
<tr>
<td>Estimation to estimate phase.</td>
<td>Overlap-Save and Newb's Algorithm. Same as OS-CC except use Newb's iterative.</td>
</tr>
<tr>
<td>Fix nearby estimated phase and estimate remaining points covered by next window. Use conjugate gradient to estimate remaining points covered by window.</td>
<td>Overlap-Save Conjugate Gradien. Assume remaining points of window are known.</td>
</tr>
<tr>
<td>Minimize magnitude error estimation sequentially frame by frame.</td>
<td>Minimize magnitude error estimation sequentially frame by frame.</td>
</tr>
<tr>
<td>(Conjugate Gradien) Use conjugate gradient algorithm for minimization.</td>
<td></td>
</tr>
<tr>
<td>from the MISTF. Then Iterate.</td>
<td>Overlap-MISTF. Use overlap-added to estimate a signal.</td>
</tr>
<tr>
<td>Overlap-Add estimation from MISTF (set magnitude of STFT of current estimate to given</td>
<td>Overlap-MISTF. Then Iterate.</td>
</tr>
<tr>
<td>estimate in given MISTF to form MISTF. Then use ISEE-MISTF to estimate a signal.</td>
<td>Least Squares Error Estimation from MISTF (set magnitude of STFT of current</td>
</tr>
<tr>
<td>Estimate Magnitude Error estimation over all frames simultaneously.</td>
<td></td>
</tr>
</tbody>
</table>

Table 4.1 - Signal Estimation Algorithms
Chapter 5

Signal Reconstruction from Unmodified STFT

5.1. Introduction

In this chapter, we will show that under certain conditions, all of the methods presented in Chapters 2 and 3 can reconstruct the original signal from appropriate information from its STFT. In Section 5.2, we will show that the LSEE-MSTFT algorithm recovers the signal exactly from its STFT. In Section 5.3, we will show that under certain conditions, most of the methods presented in Chapter 3 can reconstruct the original signal from its STFT.

5.2. Signal Reconstruction from Unmodified STFT

To show that the LSEE-MSTFT algorithm exactly recovers the original signal from its STFT, $y_w(mS,n)$ is replaced by $w(mS-n)y(n)$ in Equation (2.6) which gives us the signal estimate:

\[
x(n) = \sum_{m=-\infty}^{\infty} w^2(mS-n)y(n) = \frac{\sum_{m=-\infty}^{\infty} w^2(mS-n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} y(n) = \frac{\sum_{m=-\infty}^{\infty} w^2(mS-n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}
\]  

(5.1)

Consequently, as long as $\sum_{m=-\infty}^{\infty} w^2(mS-n)$ is non-zero for all $n$, then the original sequence $y(n)$ is exactly recovered.

5.3. Signal Reconstruction from Unmodified STFT Magnitude

Two problems arise when the algorithms presented in Chapter 3 are employed in an attempt to exactly recover the original signal from its STFTM. The first problem is that a
unique global minimum doesn't necessarily exist unless more information than the STFTM is available. This is not important if the distance measure selected is a good approximation to the auditory system's distance measure since all of these global minima will sound alike. The second problem is the possibility of the existence of local minima in the chosen distance measure to which the algorithms could converge. This problem is much more serious than the first problem since the distance measure at these local minima could be considerably greater than at the global minima leading to an inferior signal estimate. In the next section, examples of convergence rates for the algorithms developed in Chapter 3 will be given. In the following section, some examples of reconstruction from STFTM will be presented.

5.3.1. Convergence Rate Examples

In this section, the simultaneous and sequential estimation algorithms presented in Chapter 3 will be compared in terms of convergence rate. First, the convergence rate for LSEE-MSTFTM, OA-MSTFTM, simultaneous steepest descent, and simultaneous conjugate gradient will be compared. Then, the convergence rate for steepest descent, conjugate gradient, and Nawab's method will be compared for a single window of speech.

Figure 5.1 shows the time waveform of the speech sentence "Line up at the screen door." The STFTM of this sentence was calculated using a 256 point modified Hamming window, a window shift (S) of 64, and an FFT length of 512. Then, each simultaneous algorithm was used to estimate a signal with STFTM close to this unmodified STFTM starting with an initial estimate consisting of the original speech sentence with enough white Gaussian noise added to produce a 10 dB signal to noise ratio. Figure 5.2 shows distance measures $D_1$ and $D_2$ versus iteration number for LSEE-MSTFTM, OA-MSTFTM, simultaneous steepest descent, and simultaneous conjugate gradient. Both the LSEE-MSTFTM method and the OA-MSTFTM method converge much faster than the conjugate gradient method for both distance measures.
However, the OA-MSTFTM method starts to increase distance measures $D_1$ and $D_2$ at iteration 61. As expected, conjugate gradient converges much faster than steepest descent, and since they both minimize the same distance measure, the steepest descent method will no longer be considered. This tends to indicate that the LSEE-MSTFTM method is the best method to use in terms of convergence for this example. For the other examples later in this thesis, it always converges faster than the conjugate gradient method and is comparable to the initial convergence rate of the OA-MSTFTM method. However, as in this example, the OA-MSTFTM method eventually starts to increase the distance measures when the distance nears the distance at the stationary points of the LSEE-MSTFTM method.

In Figure 5.3, a 256 point rectangular windowed section of the speech sentence "Line up at the screen door." is displayed. The Fourier transform magnitude of this window of speech was calculated using a 512 point FFT with zero-padding and then steepest descent, conjugate gradient, and Nawab's method were used to estimate the last 64 points covered by the window starting with the correct first 192 points of the window and an initial estimate of zero for the last 64 points. Figure 5.4 shows the error versus iteration number for these three methods for sequential distance measures $D_1$ and $D_2$. All three of these algorithms converge to the global minimum for this example. As expected, the conjugate gradient algorithm converges much faster than steepest descent and since they both minimize the same distance measure, steepest descent will no longer be considered. Although conjugate gradient and steepest descent initially decrease the distance measures more, Nawab's method eventually surpasses them in both distance measures and more importantly, maintains its exponential convergence rate. In this example, the exact initial points were available and the example was chosen so that these methods converged to the global minimum. In Section 5.3.2, the effects of starting with incorrect initial points and convergence to local minima will be seen for reconstruction of the entire speech sentence from the STFTM.
5.3.2. Reconstruction Examples

Table 5.1 gives a performance comparison for the methods discussed in Chapter 3. The FFT length, window length, and window shift were 512, 256, and 64 respectively for all entries in this table. For the LSEE-Nawab method, the number of points estimated in each window was 128.

For comparison, the section of the waveform for the sentence "Line up at the screen door" corresponding to "the screen" for the original sentence and the estimates produced by these algorithms is displayed in Figures 5.5, 5.7, and 5.9. Similarly, the spectrograms of the original sentence and the spectrogram for each of the signal estimates is displayed in Figures 5.6, 5.8, and 5.10.

Of the simultaneous methods, LSEE-MSTFTM and OA-MSTFTM produce the most periodic waveform with no clicks. In terms of distance measures, LSEE-MSTFTM produces the smallest $D_1$ and simultaneous conjugate gradient (CG) produces the smallest $D_2$. This is expected since LSEE-MSTFTM converges to the critical points of $D_1$ and simultaneous conjugate gradient converges to the critical points of $D_2$.

Of the sequential methods, LSEE-Nawab is superior in all categories. The other sequential methods tend to produce glitches in the waveform which results in wideband noise in the spectrogram and a perception of clicks in the listening test.

Overall, LSEE-MSTFTM, OA-MSTFTM, and LSEE-Nawab produce the best estimates in terms of waveform periodicity, absence of wideband noise, and listening tests. It is important to note that while the waveform reconstructed by these algorithms is very different from the original due to convergence to local minima, the spectrograms are very close to the original and the estimates sound very close to the original. A small amount of reverberation is detectable in the reconstructions due to the nonstationarity of the speech waveform over the window
length and convergence to local minima.

5.4. Summary

We have shown that if the unmodified STFT of a sequence is available, the LSEE-MSTFT method will recover the original sequence exactly. If the unmodified STFTM of a sequence and an initial estimate which is close enough to the original sequence are available, then all of the methods of Chapter 3 can reconstruct the original sequence. However, if no good initial estimate is available, none of these methods will in general reconstruct the original sequence exactly. LSEE-MSTFTM, OA-MSTFTM, and LSEE-Nawab all reconstruct a signal with STFTM close enough to the STFTM of the original sequence so that the signal estimates are very close to the original signal in listening tests. The other methods fail due to convergence to local minima in the distance function which are far from the global minimum (or minima).
Figure 5.2a - Simultaneous Distance $D_1$ for a) LSEE-MSTFTM, b) OA-MSTFTM, c) Conjugate Gradient, and d) Steepest Descent

Figure 5.2b - Simultaneous Distance $D_2$ for a) LSEE-MSTFTM, b) OA-MSTFTM, c) Conjugate Gradient, and d) Steepest Descent
Figure 5.3 - 256 point rectangular window of speech
Figure 5.4a - Sequential Distance $D_1$ for a) Steepest Descent, b) Conjugate Gradient, and c) Nawab's method

Figure 5.4b - Sequential Distance $D_2$ for a) Steepest Descent, b) Conjugate Gradient, and c) Nawab's method
<table>
<thead>
<tr>
<th>Situation</th>
<th>Disturbance</th>
<th>Periodicity</th>
<th>$H_m$</th>
<th>$H_c$</th>
<th>$D_1$</th>
<th>$D_2$</th>
<th>Type</th>
<th>Window</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>No distortion</strong></td>
<td>no chicks</td>
<td>no wideband</td>
<td>40400</td>
<td>1170</td>
<td>30</td>
<td>40</td>
<td>Recl.</td>
<td>1850</td>
<td>ISTE-NEWab</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ISTE-CC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OS-NEWab</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OS-CC</td>
</tr>
<tr>
<td><strong>Different</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Sequential</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Some chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Many chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Low chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>No distortion</strong></td>
<td>no chicks</td>
<td>no wideband</td>
<td>6970</td>
<td>1160</td>
<td>100</td>
<td>180</td>
<td>Recl.</td>
<td>1185</td>
<td>ISTE-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OA-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ISTE-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Different</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Some chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Many chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Low chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>No distortion</strong></td>
<td>no chicks</td>
<td>no wideband</td>
<td>6590</td>
<td>1380</td>
<td>100</td>
<td>100</td>
<td>Mod.</td>
<td>1760</td>
<td>ISTE-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OA-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ISTE-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Different</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Some chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Many chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Low chicks</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>No distortion</strong></td>
<td>no chicks</td>
<td>no wideband</td>
<td>324</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>Mod.</td>
<td>1760</td>
<td>ISTE-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>OA-MSTTM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ISTE-MSTTM</td>
</tr>
</tbody>
</table>

*Table S.1 - Performance Comparison for Reconstruction From S.I.T. Amplitude*
Figure 5.6a - STFTM of original signal

Figure 5.6b - STFTM of LSEE-MSFTM estimate

Figure 5.6c - STFTM of OA-MSFTM estimate

Figure 5.6d - STFTM of Simultaneous Conjugate Gradient estimate
Figure 5.7 - (a) Original signal, and signal estimates for (b) OS-CC, and (c) OS-Navab
Figure 5.8a - STFTM of original signal

Figure 5.8b - STFTM of OS-CG estimate

Figure 5.8c - STFTM of OS-Nawab estimate
Figure 5.9 - a) Original signal, and signal estimate for b) LSEE-CG, and c) LSEE-Nawab
Chapter 6

Speech Enhancement by Spectral Subtraction

6.1. Introduction

Several methods of enhancing speech degraded by additive noise have been developed which modify the STFT of the noisy signal and then estimate a signal from this modified STFT. In this chapter, we will apply the methods developed in Chapters 2 and 3 for estimating a signal from MSTFT or MSTFTM to speech enhancement by spectral subtraction.

Spectral subtraction is based on the availability of a signal $y(n)$ which consists of the desired signal $s(n)$ degraded by uncorrelated additive noise $d(n)$. Expressing this relation in terms of the STFTs of these signals, we have

$$ Y_w(mS,\omega) = S_w(mS,\omega) + D_w(mS,\omega) \quad (6.1) $$

and

$$ |Y_w(mS,\omega)|^2 = |S_w(mS,\omega)|^2 + |D_w(mS,\omega)|^2 + 2\text{Re}[S_w(mS,\omega)D_w^*(mS,\omega)] \quad (6.2) $$

Thus, with zero mean noise and a given $|S_w(mS,\omega)|^2$, the expected value of the estimate

$$ |\hat{S}_w(mS,\omega)|^2 = |Y_w(mS,\omega)|^2 - E\left[|D_w(mS,\omega)|^2\right] \quad (6.3) $$

is $|S_w(mS,\omega)|^2$ giving us an unbiased estimate. Assuming a Gaussian density for $D_w(mS,\omega)$, the variance of this estimate is:

$$ \text{Var}\left[|\hat{S}_w(mS,\omega)|^2\right] = \text{Var}\left[|Y_w(mS,\omega)|^2\right] $$

$$ = E^2\left[|D_w(mS,\omega)|^2\right] - E\left[\left(D_w(mS,\omega)\right)^2\right]^2 $$

$$ + 2|S_w(mS,\omega)|^2E\left[|D_w(mS,\omega)|^2\right] $$
\[ +2 \text{Re}\left( S_w(mS, \omega) \right)^2 \text{E}\left[ |D_w(mS, \omega)|^2 \right]^* \]

Since the second term can vary between 0 and \( E^2 \left[ |D_w(mS, \omega)|^2 \right] \) and the fourth term can vary between \(-2|S_w(mS, \omega)|^2 E \left[ |D_w(mS, \omega)|^2 \right] \) and \( 2|S_w(mS, \omega)|^2 E \left[ |D_w(mS, \omega)|^2 \right] \), the variance is at least \( E^2 \left[ |D_w(mS, \omega)|^2 \right] \). Thus, the standard deviation of this estimate is greater than or equal to the mean value of the noise STFT magnitude squared. Consequently, the values subtracted from each STFT frame will differ significantly from the actual noise power at each frequency in that frame. This leads to a signal estimate with bursts of energy at frequencies where the noise power is much greater than the value subtracted resulting in "musical tones" in the signal estimate.

Another problem is that the estimate \( |\hat{S}_w(mS, \omega)|^2 \) in Equation (6.3) is negative when \( E \left[ |D_w(mS, \omega)|^2 \right] \) is greater than \( |Y_w(mS, \omega)|^2 \). This problem is usually resolved by setting \( |\hat{S}_w(mS, \omega)|^2 \) to zero in this case. Although this gives us a biased estimate, it also reduces the variance somewhat since the estimate can no longer become negative. Due to the large variance of the estimate and the noise that still remains, the estimate

\[
|\hat{S}_w(mS, \omega)|^2 = \begin{cases} |Y_w(mS, \omega)|^2 - kE \left[ |D_w(mS, \omega)|^2 \right], & kE \left[ |D_w(mS, \omega)|^2 \right] \leq |Y_w(mS, \omega)|^2 \\ 0, & \text{otherwise} \end{cases}
\]

is usually used where \( k \) is chosen to be greater than one. The parameter \( k \) allows more than the expected noise power to be subtracted to account for the large variance of the noise power in a single frame. Although, choosing \( k > 1 \) increases the bias of the estimate, it reduces the variance of the estimate as well due to the clamping of negative estimates to zero. However, due to the large noise variance, the noise cannot be entirely subtracted without removing a significant portion of the desired signal.
The usual implementation of this method combines this new magnitude estimate with the noisy phase estimate and then takes the inverse FT and uses an overlap-add procedure to produce the time domain estimate. Since the STFT has been modified, we can apply the LSEE-MSTFT algorithm presented in Chapter 2 to find the time domain sequence with STFT closest to this MSTFT. The results of this method will be presented in Section 6.2.

Since, in spectral subtraction, we have a better estimate of the STFT magnitude but not the phase, we would expect methods capable of estimating a signal directly from MSTFTM such as those presented in Chapter 3 to produce somewhat better results. However, since the primary artifact in this method, "musical tones", is caused by a poor estimate of the STFTM of the desired signal, estimating a signal directly from the magnitude will not decrease the "musical tones" artifact. In addition, after spectral subtraction, the only regions of the spectrum due to the signal that remain are the regions which had a high signal-to noise ratio in the STFTM of the degraded signal. Since the STFT phase is less affected by the additive noise in regions of the spectrum where the STFTM SNR is high, it is expected that the phase is approximately correct in these regions. Further, the result by Wang [15] indicates that small improvements in the phase estimate will not significantly improve the final signal estimate. Since signal estimation methods from modified STFTM are not expected to reduce the primary artifact of spectral subtraction, the primary reason for applying them to spectral subtraction will be for comparison of these algorithms. The results of these methods will be presented in Section 6.3.

6.2. Spectral Subtraction Signal Estimation from MSTFT

In this section, we will present the results of applying the LSEE-MSTFT algorithm to speech enhancement by spectral subtraction and compare them to the results obtained by the standard overlap-add algorithm. These methods are applied by calculating the MSTFT $\hat{S}_w(mS,\omega)$ from the STFT of the noisy signal $Y_w(mS,\omega)$ as follows:
$$\hat{S}_w(mS, \omega) = \sqrt{|\mathcal{S}_w(mS, \omega)|^2 \frac{Y_w(mS, \omega)}{|Y_w(mS, \omega)|}}$$  \hspace{1cm} (6.6)

where $|\mathcal{S}_w(mS, \omega)|^2$ is given by Equation (6.5). The signal estimate can then be obtained from $\hat{S}_w(mS, \omega)$ using the LSEE-MSTFT algorithm or the overlap-add algorithm.

Figure 6.1a shows the portion of the time waveform of the speech sentence "Line up at the screen door." corresponding to "the screen." Figure 6.1b shows the same portion of waveform with enough white Gaussian noise added to produce a 10 dB signal to noise ratio (SNR). The signal to noise ratio was calculated as:

$$\text{SNR} = 10 \log \left( \frac{\sum_{n=-\infty}^{\infty} s^2(n)}{\sum_{n=-\infty}^{\infty} d^2(n)} \right)$$  \hspace{1cm} (6.7)

Then, the MSTFT $\hat{S}_w(mS, \omega)$ was calculated using a 256 point modified Hamming window, a window shift of 64 points, an FFT length of 512 points, and a $k$ value of 4 in Equation (6.5). Figures 6.1c and 6.1d show the results of estimating the signal from the MSTFT $\hat{S}_w(mS, \omega)$ using the overlap-add method and the LSEE-MSTFT method. Due to the similarity in these algorithms, the signal estimates look and sound very similar. Figure 6.2 shows the STFTM of the noisy speech, the estimate of the STFTM of the signal ($|\mathcal{S}_w(mS, \omega)|$), and the results of the two algorithms. Comparison of the STFTM of the signal estimates with the MSTFTM $|\mathcal{S}_w(mS, \omega)|$ shows that even though the correct phase was not used, the STFTM of the signal estimates is very close to $|\mathcal{S}_w(mS, \omega)|$. This indicates that the phase of the noisy signal was fairly close to the correct phase for the frequencies with high SNR. Smearing of the STFTM of both estimates in time is also evident in Figure 6.2. The maximum time smearing possible with these algorithms is the length of the synthesis window. Consequently, the overlap-add algorithm exhibits slightly more time smearing of the short bursts of energy in $|\hat{S}_w(mS, \omega)|$ since the length of the synthesis window is effectively the length of the FFT. This figure also shows
the artifacts of the spectral subtraction method as bursts of energy where the STFTM of the
noisy speech had a value much greater than the expected value. These bursts of energy can be
easily heard in the background since the energy at neighboring frequencies has been removed
by spectral subtraction. Since the STFTM of the signal estimates is so close to the desired
MSTFTM in this example, very little improvement is expected from estimating the signal
directly from the MSTFTM $|\hat{S}_w(mS,\omega)|$ using the methods of Chapter 3.

The top portion of Table 6.1 shows the simultaneous distance measures $D_1$ and $D_2$ of
Chapter 3 for the estimates produced by LSEE-MSTFT and overlap-add. These distance
measures were used to measure the distance between the STFTM of the signal estimates and
the desired MSTFTM $|\hat{S}_w(mS,\omega)|$. For this example, it can be seen that LSEE-MSTFT pro-
duces both a smaller $D_1$ and a smaller $D_2$.

6.3. Spectral Subtraction Signal Estimation from MSTFTM

In this section, the methods presented in Chapter 3 for estimating a signal from MSTFTM
will be applied to the problem of estimating a signal with STFTM close to the spectral subtrac-
tion MSTFTM estimate $|\hat{S}_w(mS,\omega)|$ given by Equation (6.5). The results of the simultaneous
algorithms will be presented in the next section. In the following section, the results of the
sequential algorithms will be presented.

6.3.1. Simultaneous Algorithms

This section presents the estimates produced by the simultaneous algorithms for spectral
subtraction. The value used for $|\hat{S}_w(mS,\omega)|$ was the same as in Section 6.2. The window type
and number of iterations for each algorithm are shown in Table 6.1, the other common
parameters are an FFT length of 512, a window length of 256, a window shift of 64, and an
initial estimate consisting of the noisy signal. The signal estimates produced by the simultane-
ous methods are shown in Figure 6.3. The spectrograms of these estimates are shown in Figure 6.4.

For the LSEE-MSTFTM method, since the initial estimate was the noisy signal, the first iteration of this method is the same as using the LSEE-MSTFT method. Although the simultaneous distance measure $D_1$ continues to decrease after the first iteration, the estimate sounds the same.

Similarly, for the OA-MSTFTM method, the first iteration is the same as using the overlap-add method. However, for this method, $D_1$ increases somewhat after the first iteration although the estimate continues to sound about the same.

There is still some noise in the conjugate gradient estimate which is not present in the estimates of Section 6.2 since the conjugate gradient method did not completely converge after 50 iterations. The simultaneous conjugate gradient spectral subtraction estimate doesn’t exhibit as much aperiodicity as the simultaneous conjugate gradient estimate in Chapter 5 since the initial estimate was much closer to the global minimum in this case. As in Chapter 5, this method converged the slowest of the simultaneous algorithms and produced the worst estimate.

The spectrograms of these estimates (Figure 6.4) show that LSEE-MSTFTM and OA-MSTFTM come closest to the desired MSTFTM. As with LSEE-MSTFT and OA-MSTFT, the STFTMs of the estimates are blurred slightly in time due to the length of the synthesis window. Consequently, the OA-MSTFTM estimate is blurred slightly more in time than the LSEE-MSTFTM estimate.

6.3.2. Sequential Algorithms

This section presents the estimates produced by the sequential algorithms for spectral subtraction. The value used for $|\hat{S}_w(mS, \omega)|$ was the same as in Section 6.2. The signal estimates were calculated using the parameters shown in Table 6.1 with an FFT length of 512 points, a
window length of 256 points, a window shift of 64 points, and 128 points estimated per window for the LSEE-Nawab method.

Figure 6.5 shows the estimates produced by the OS-CG and OS-Nawab methods. The STFTMs of these estimates are shown in Figure 6.6 and the distances $D_1$ and $D_2$ are given in Table 6.1. As in signal estimation from unmodified STFTM, both of these methods affect the periodicity of the signal which shows up as vertical bars in the spectrograms shown in Figure 6.6. Due to this problem, especially evident in the estimate produced by the conjugate gradient algorithm, both of these methods perform worse than the methods for signal estimation from MSTFT.

Figure 6.7 shows the estimates produced by the LSEE-CG method and the LSEE-Nawab method. The spectrogram of the LSEE-CG estimate, shown in Figure 6.8 indicates that this estimate is much better than the OS-CG estimate. However, it still suffers from aperiodicity which is manifested in a lack of continuity in the pitch harmonics in the spectrogram. The LSEE-Nawab method produces the best estimate of the sequential algorithms and sounds the same as the estimate produced by the LSEE-MSTFT algorithm in Section 6.2. However, distance measures $D_1$ and $D_2$ are greater for this method than for LSEE-MSTFT or LSEE-MSTFTM.

6.4. Conclusions

For spectral subtraction, the phase of the noisy speech is approximately correct for the frequencies with high SNR. This allows the LSEE-MSTFT and overlap-add method to estimate a signal with STFTM close to the estimated MSTFTM $|\hat{S}_w(mS,\omega)|$. Since these methods estimate a signal with STFTM so close to the desired MSTFTM, no noticeable improvement is realized by using the methods for estimating a signal directly from MSTFTM.
<table>
<thead>
<tr>
<th>Distortion</th>
<th>Noise</th>
<th>Window</th>
<th>Phase</th>
<th>Method</th>
<th>Iter.</th>
<th>Type</th>
<th>Window</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Noise</td>
<td>no wideband</td>
<td>no glitch, periodic</td>
<td>ISEE-Nim</td>
<td>30</td>
<td>Mod.</td>
<td>ISEE-Nim</td>
</tr>
<tr>
<td>Direction</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Same</td>
<td>Noise</td>
<td>no wideband</td>
<td>no glitch, periodic</td>
<td>ISEE-NWIP</td>
<td>40</td>
<td>Mod.</td>
<td>ISEE-NWIP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Many</td>
<td>Noise</td>
<td>no wideband</td>
<td>no glitch, periodic</td>
<td>OS-NWIP</td>
<td>30</td>
<td>Rect.</td>
<td>OS-NWIP</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>much wideband</td>
<td>many glitches, periodic</td>
<td>OS-CG</td>
<td>40</td>
<td>Rect.</td>
<td>OS-CG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>low wideband</td>
<td>few glitches, periodic</td>
<td>CG</td>
<td>30</td>
<td>Mod.</td>
<td>CG</td>
</tr>
<tr>
<td>Low</td>
<td>Noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>high wideband</td>
<td>no glitch, periodic</td>
<td>O-A-MSTIF</td>
<td>40</td>
<td>Mod.</td>
<td>O-A-MSTIF</td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>no wideband</td>
<td>no glitch, periodic</td>
<td>1SEE-MSTIF</td>
<td>20</td>
<td>Mod.</td>
<td>1SEE-MSTIF</td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>no wideband</td>
<td>no glitch, periodic</td>
<td>MSITF</td>
<td>20</td>
<td>Mod.</td>
<td>MSITF</td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td>all wideband</td>
<td>no glitch, periodic</td>
<td>MSITF</td>
<td>20</td>
<td>Mod.</td>
<td>MSITF</td>
</tr>
<tr>
<td></td>
<td>Noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1: Performance Comparison for Spectral Subtraction
Figure 6.2a - STFTM of Noisy signal

Figure 6.2b - Desired MSTFTM

Figure 6.2c - STFTM of Overlap-add estimate

Figure 6.2d - STFTM of LSEE-MSTFT estimate
Figure 6.3 - (a) Original signal, (b) Noisy signal, (c) Signal estimate for e) IS-E-MISEM, (d) OA-MISEM, and (e) Simultaneous Conjugate Gradient.
Figure 6.4a - Desired MSTFTM

Figure 6.4b - STFTM of LSEE-MSTFTM estimate

Figure 6.4c - STFTM of OA-MSTFTM estimate

Figure 6.4d - STFTM of Simultaneous Conjugate Gradient estimate
Figure 6.6a - Desired MSTFTM

Figure 6.6b - STFTM of OS-CG estimate

Figure 6.6c - STFTM of OS-Nawab estimate
Figure 6.7: a) Original signal, b) Noisy signal, c) signal estimates for LSEE-CG, and d) LSEE-Nawab.
Figure 6.8a - Desired MSTFTM

Figure 6.8b - STFTM of LSEE-CG estimate

Figure 6.8c - STFTM of LSEE-Nawab estimate
Chapter 7

Time-Scale Modification of Speech

7.1. Introduction

One method of decomposing a speech signal $y(n)$ is to represent it as the convolution of an excitation function with a time-varying vocal tract impulse response. Consequently, the STFT magnitude of this speech signal $|Y_w(mS,\omega)|$ can be written as the product of a component due to the excitation function $|P_w(mS,\omega)|$ and a component due to the vocal tract impulse response $|H_w(mS,\omega)|$. This decomposition is valid if the analysis window is long enough to include several vocal tract impulse responses and short enough so that the speech signal is approximately stationary over the window length. Under these conditions, the function $|P_w(mS,\omega)|$ will correspond to the rapidly varying portion of $|Y_w(mS,\omega)|$ with $\omega$, taking on an harmonic structure for voiced speech or noise for unvoiced speech. The function $|H_w(mS,\omega)|$ will correspond to the slowly varying portion of $|Y_w(mS,\omega)|$ with $\omega$, and will include the formant information of the speech signal. Since the speech signal is assumed to be approximately stationary over the window length, $|P_w(mS,\omega)|$ and $|H_w(mS,\omega)|$ will change slowly with the time index $mS$ as the pitch period and vocal tract impulse response change.

The goal of time-scale modification is to modify the rate at which $|P_w(mS,\omega)|$ and $|H_w(mS,\omega)|$ vary with time, hence the rate at which $|Y_w(mS,\omega)|$ varies with time, without affecting the spectral characteristics. Several methods which attempt to accomplish this have previously been developed and operate in either the time domain or the STFT domain. Among these methods are Fairbanks' method [16] which operates in the time domain and Portnoff's
method [1] which exploits properties of the STFT of speech signals.

Fairbanks' method entails removing sections from the original speech if an increase in the articulation rate is desired or replicating sections if a decrease in the articulation rate is desired. The section length should be chosen to be equal to or longer than the pitch period and shorter than the length of the phoneme. This allows the pitch period to be preserved locally while the phoneme is compressed or elongated. However, the transitions between sections for this method cause an annoying distortion with a period of the section length since the resulting signal is no longer periodic at these transitions. This distortion of the periodicity of the signal prevents the desired time-scale modified $|P_w(mS, \omega)|$ from being produced. Various time domain methods have improved on Fairbanks' original method somewhat but haven't completely eliminated this problem.

Portnoff's method employs properties of the STFT to generate time-scale modified speech. The STFTM of the desired signal is produced by time-scale modification of the STFTM of the original speech signal. The STFT phase is generated by adding appropriately modified versions of the phase modulation component and frequency modulation component of the phase. The phase modulation component is time-scale modified by the same factor as the STFTM. The frequency modulation component is time-scale modified, unwrapped, and divided by the time-scale modification factor. The phase modulation component is assumed to vary slowly enough so that the difference from one frame to the next is zero. Consequently, changes in the phase modulation component from frame to frame will cause errors in the phase estimate. Another problem with this method is the necessity of accurately unwrapping the frequency modulation component which causes degradation when errors occur.

An alternate approach to the problem of time-scale modification of speech is to estimate the desired signal directly from the time-scale modified STFTM using the methods presented in
Chapter 3. This can be accomplished by calculating the STFTM of the original speech \(|Y_w(mS_1,\omega)|\) at a window shift \(S_1\) different from the window shift \(S_2\) of the STFT of the signal estimate \(|X_w(mS_2,\omega)|\). This results in a time scale modification of \(S_1 : S_2\). Thus, when \(S_1\) is twice \(S_2\) the time-scale of the STFTM of the signal estimate is compressed by a factor of two which speeds up the apparent articulation rate by a factor of two. To see this in more detail consider Figure 7.1a. In this figure, the analysis window shift \((S_1)\) used for calculating \(|Y_w(mS_1,\omega)|\) is 128 and the synthesis window shift \((S_2)\) used in the signal estimation algorithm is 64. The window length for this example is 256 points. Suppose a section of the original waveform corresponding to \(y_a\) has a spectrum of \(Y_a\). Then, the spectrum \(Y_1\) contributes to the windowed portions of speech covered by windows \(w_1\) and \(w_2\). Consequently, during the synthesis stage, the algorithms will produce a signal with a spectrum under windows \(w_1\) and \(w_2\) close to \(Y_1\). Since none of the windows adjacent to \(w_1\) and \(w_2\) cover \(y_1\), the spectrum \(Y_1\) is constrained to the region of the signal estimate \(x(n)\) corresponding to \(x_1\) in order to minimize the error. Thus, the 128 point section of the original waveform \(y_1\) is compressed to the 64 point section of waveform \(x_1\) while maintaining approximately the same spectrum \(Y_1\).

Similarly, when \(S_1\) is half \(S_2\) the time-scale of the STFTM of the signal estimate is expanded by a factor of two which slows down the apparent articulation rate by a factor of two. To see this in more detail consider Figure 7.1b. In this figure, the analysis window shift \((S_1)\) used for calculating \(|Y_w(mS_1,\omega)|\) is 32 and the synthesis window shift \((S_2)\) used in the signal estimation algorithm is 64. The window length for this example is 256 points. In the analysis stage, windows \(w_{-2}\) through \(w_5\) cover the section of speech \(y_1\). Consequently, in the synthesis stage, all of the windows \(w_{-2}\) through \(w_5\) will cover a portion of the signal estimate with a spectrum that contains \(Y_1\) in order to minimize error. Since \(w_{-3}\) and \(w_6\) are the closest windows to this range which do not cover \(y_1\), the sections of the signal estimate \(x(n)\) which contain the spectrum \(Y_1\) are constrained to be between the sections covered by \(w_{-3}\) and \(w_6\) to
minimize error. Thus, the spectrum $Y_1$ is spread out so that it occurs in sections $x_{-1}$ through $x_3$. Similarly, the spectrum $Y_0$ occurs in sections $x_{-2}$ through $x_2$ and $Y_2$ occurs in sections $x_0$ through $x_4$. So, each spectrum $Y_n$ occurs in section $x_n$ and is also smeared into sections $x_{n-2}$, $x_{n-1}$, $x_{n+1}$, and $x_{n+2}$. In general, the amount of smearing will be approximately

$$\Delta=(r_e-1)L$$

(7.1)

where $\Delta$ is the amount of smear, $r_e$ is the expansion ratio, and $L$ is the window length. Thus, for this example, $r_e$ is 2, $L$ is 256, and $\Delta$ is 256. Consequently, the amount of smear of the spectrum $Y_n$ on each side of section $x_n$ will be half of $\Delta$ or 128 points corresponding to smearing into sections $x_{n-2}$, $x_{n-1}$, $x_{n+1}$, and $x_{n+2}$ as discussed earlier. When $r_e$ is less than or equal to 1, the amount of smearing is zero as was indicated in the 2:1 compression example discussed earlier ($r_e=.5$). This discussion assumes that the distance measure has been minimized so that a global minimum has been reached and is therefore a best case analysis. A worst case analysis for the three best algorithms (LSEE-MSTFTM, OA-MSTFTM, and LSEE-Nawab) is given in the following paragraphs.

For LSEE-MSTFTM, OA-MSTFTM, and LSEE-Nawab, the maximum amount of smearing of the spectrum is governed by the synthesis window length. For example, consider 1:2 expansion shown in Figure 7.1b. In this figure, all of the windows $w_{-2}$ through $w_5$ cover section $y_1$. Consequently, in the synthesis step of these algorithms, the energy in the spectrum $Y_1$ could, at worst, be spread out over all of these windows, namely, from section $x_{-4}$ to section $x_{6}$. In general, for the worst case, the amount of smear is

$$\Delta=(r_e+1)L-2S_2$$

(7.2)

where $S_2$ is the synthesis window shift, $L$ is the effective synthesis window length, and the other variables are the same as in Equation (7.1). For LSEE-MSTFTM, and LSEE-Nawab, $L$ is just the synthesis window length. For OA-MSTFTM, $L$ is the length of the FFT since no explicit synthesis window is used. Consequently, for OA-MSTFTM, the worst case amount of
smear can be much larger than for LSEE-MSTFTM and LSEE-Nawab.

In the next section, the simultaneous and sequential methods developed in Chapter 3 will be used to estimate signals directly from time-scale modified STFTM. The examples presented in this section include a time-scale compression of 2:1 and a time-scale expansion of 35:64. These two examples are presented to show that these methods perform equally well for both integer and non-integer modification ratios which is not the case for such methods as Portland's. A third example, which performs a 1:2 time-scale expansion of the 2:1 time-scale compressed speech is presented for the algorithms which produce high quality speech. This example is important for several reasons. First, any errors which the algorithm makes will be much more audible since the algorithm has been applied twice and the original sequence is available for comparison. Second, this example demonstrates the possibility of decreasing the data rate over communication channels by performing time-scale compression before voice transmission followed by time-scale expansion.

7.2. Signal Estimation from Time-Scale Modified STFTM

In this section, the algorithms developed in Chapter 3 will be used to estimate a signal with STFTM close to a given time-scale modified STFTM. The performance of these methods will be compared for 2:1 time-scale compression and 35:64 time-scale expansion. The 2:1 time-scale compression was accomplished by calculating the STFTM of the original speech using a window shift of 128 points and then modifying the time-scale so that each frame was 64 points apart. Similarly, the 35:64 time-scale expansion was accomplished by calculating the STFTM of the original speech using a window shift of 35 points and then modifying the time-scale so that each frame was 64 points apart. A final example shows the performance of the three "best" algorithms (LSEE-MSTFTM, LSEE-Nawab, and OA-MSTFTM) for 2:1 time-scale compression followed by 1:2 time-scale expansion.
Table 7.1 shows a performance comparison of the simultaneous and sequential algorithms for 2:1 time-scale compression. All of the algorithms used a window length of 256, a window shift of 64, and an FFT length of 512. For the LSEE-Nawab method, the number of points estimated in each window was 128. This table shows that LSEE-MSTFTM, LSEE-Nawab, and OA-MSTFTM produce the best signal estimates. It is interesting to note that while CG produces a smaller $D_2$ than all of the other methods, the signal estimate that it produces is inferior to LSEE-MSTFTM, LSEE-Nawab, and OA-MSTFTM.

Figures 7.2-7.3 show the waveforms and spectrograms for the signal estimates produced by the simultaneous algorithms. The waveforms and spectrograms for the estimates produced by LSEE-MSTFTM and OA-MSTFTM are the best of the simultaneous algorithms. The signal estimate produced by CG contains a few glitches in the time waveform resulting in broadband noise in the spectrogram.

Figures 7.4-7.7 show the waveforms and spectrograms for the signal estimates produced by the sequential algorithms. By comparing these figures, it can easily be seen that LSEE-Nawab is the best of the sequential methods for this example since it comes much closer to the desired spectrogram, and has no glitches in the time waveform. The LSEE-Nawab method produces a signal estimate with quality comparable to LSEE-MSTFTM and OA-MSTFTM.

By measuring the length of the /k/ burst in the word "screen", it can be seen that almost no smearing of the spectrum in time occurs for LSEE-MSTFTM and LSEE-Nawab as would be predicted by the best case smearing estimate discussed in the previous section. For OA-MSTFTM, a little smearing occurs, probably due to the longer effective synthesis window.

Table 7.2 shows a performance comparison of the simultaneous and sequential algorithms for 35:64 time-scale expansion. All of the algorithms used a window length of 256, a window shift of 64, and an FFT length of 512. For the LSEE-Nawab method, the number of points
estimated in each window was 128. This table shows that LSEE-MSTFTM, LSEE-Nawab, and
OA-MSTFTM produce the best signal estimates. Again, CG produces a smaller $D_2$ than all of
the other methods, however, the signal estimate that it produces is inferior to LSEE-MSTFTM,
LSEE-Nawab, and OA-MSTFTM.

Figures 7.8-7.9 show the waveforms and spectrograms for the signal estimates produced
by the simultaneous algorithms. The waveforms and spectrograms for the estimates produced
by LSEE-MSTFTM and OA-MSTFTM are the best of the simultaneous methods. The signal
estimate produced by CG contains a few glitches in the time waveform resulting in broadband
noise in the spectrogram.

Figures 7.10-7.13 show the waveforms and spectrograms for the signal estimates produced
by the sequential algorithms. As with 2:1 compression, it can be seen from the figures that
LSEE-Nawab is the best of the sequential methods for this example since it comes much closer
to the desired spectrogram, and has no glitches in the time waveform. Again, the LSEE-Nawab
method produces a signal estimate with quality comparable to LSEE-MSTFTM and
OA-MSTFTM.

By measuring the length of the /k/ burst in the word "screen", it can be seen that approxi-
mately 220 points of smearing occurs for LSEE-MSTFTM, 280 points for LSEE-Nawab, and
310 points for OA-MSTFTM. These approximate amounts of smearing are much closer to the
best case smearing of 212 points than to the worst case smearing of 596 points for this example.

Table 7.3 shows a performance comparison of the three best methods (LSEE-MSTFTM,
OA-MSTFTM, and LSEE-Nawab) for 2:1 time-scale compression followed by 1:2 time-scale
expansion of the compressed speech. This example is particularly important since any errors
made by the algorithms will be much more apparent because the algorithms have been applied
twice to obtain the final estimate. In addition, this example shows one method of reducing
bandwidth for voice transmission by performing time-scale compression of the speech before transmission, and then performing time-scale expansion of the speech after transmission. Figure 7.14 shows the waveforms of the original speech and the final estimates produced by LSEE-MSTFTM, OA-MSTFTM, and LSEE-Nawab. From these waveforms, especially in the /k/ region of the word "screen", the primary degradation caused by compression followed by expansion can be seen. The /k/ burst has been lengthened in the estimates with respect to the original. This effect can be seen throughout the sentence as a slight smearing in time of the spectrograms of the final estimates shown in Figures 7.15-7.17. This time smearing of the spectrograms causes the speech to sound slightly reverberant. As discussed earlier, the time smearing occurs primarily in the 1:2 expansion step. By measuring the /k/ burst for the original and reconstructed versions, the amount of time smearing was determined to be approximately 260 points for LSEE-MSTFTM, 300 points for LSEE-Nawab, and 340 points for OA-MSTFTM. These values are much closer to the best case smearing estimate of 256 points than to the worst case estimate of 640 points. Since the amount of smearing doesn't differ very much between estimates, these three estimates sound approximately the same.

7.3. Conclusions

For time-scale modification of speech, LSEE-MSTFTM, OA-MSTFTM, and LSEE-Nawab produce the best signal estimates. For time-scale compression, very little time smearing of the spectrum occurs. For time-scale expansion, time smearing of the spectrum is well predicted by the best case estimate given in Equation (7.1).
Figure 7.1a - 2:1 Compression

Figure 7.1b - 1:2 Expansion
<table>
<thead>
<tr>
<th>Description</th>
<th>Noise</th>
<th>Periodic</th>
<th>ISEE-Navd.</th>
<th>ISEE-CG</th>
<th>OS-Navdp</th>
<th>OS-CG</th>
<th>Sequential Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Dist. No Chick.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Dist. Few Chick.</td>
<td>Noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Dist. Many Chick.</td>
<td>Noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening Test</td>
<td>Speculum</td>
<td></td>
<td>Waveform</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 7.1 - Performance Comparison for 21 Time-Scale Compression**
Figure 7.2: Original signal 2:1 compression for (b) LEF-MSTFM.

(a) OA-MSTFM, and (d) Simultaneous Conjugate Gradient.
Figure 7.3a - Desired MSTFTM

Figure 7.3b - STFTM of LSEE-MSTFTM estimate

Figure 7.3c - STFTM of OA-MSTFTM estimate

Figure 7.3d - STFTM of Simultaneous Conjugate Gradient estimate
Figure 7.4 - (a) Original signal, (b) compression for b, (c) OS-CG, and (d) OS-NA韦b.
Figure 7.5a - Desired MSTFTM

Figure 7.5b - STFTM of OS-CG estimate

Figure 7.5c - STFTM of OS-Nawab estimate
FIGURE 7.5 - (a) Original signal, (b) compression for p, (c) LSEE-CC, and (d) LSEE-Navb
Figure 7.7a - Desired MSFTM

Figure 7.7b - STFTM of LSEE-CG estimate

Figure 7.7c - STFTM of LSEE-Nawab estimate
| Distortion | Noise | Periodic | Wavelet | Iter | Type | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NEP | ISE-NWEP | ISE-NWEP |
|-----------|-------|----------|---------|------|-----|-------|--------|-------|---------|----------|----------|--------|---------|----------|---------|
| No distortion | no | noise | periodic | no glitches | 0.00 | 1940 | 30 | 0 | Mod. | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| Distortion | noise | high bandwidth | periodic | some glitches | 0.00 | 1240000 | 44 | 0 | Rect. | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| Some distortion | noise | high bandwidth | periodic | some glitches | 0.00 | 331000 | 30 | 0 | Rect. | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| Many distortion | noise | high bandwidth | periodic | many glitches | 0.00 | 1230000 | 40 | 0 | Rect. | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| Few distortion | noise | high bandwidth | periodic | few glitches | 0.00 | 57300 | 100 | 0 | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| No distortion | noise | high bandwidth | periodic | no glitches | 0.00 | 185000 | 100 | 0 | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| No distortion | noise | high bandwidth | periodic | no glitches | 0.00 | 20500 | 100 | 0 | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| No distortion | no | noise | periodic | no glitches | 0.00 | 26700 | 100 | 0 | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |
| No distortion | no | noise | periodic | no glitches | 0.00 | 2485 | 100 | 0 | Ham. | Mod. | ISE-CA | ISE-CC | ISE-CW | ISE-NAWP | ISE-NAWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP | ISE-NWEP |

Table 7.2: Performance Comparison for 3564 Time-Step Expansion
Figure 7.8: (a) Original signal; (b) 350x expansion for (b); (c-e) LSE-MITTM.
Figure 7.9a - Desired MSTFTM

Figure 7.9b - STFTM of LSEE-MSTFTM estimate

Figure 7.9c - STFTM of OA-MSTFTM estimate

Figure 7.9d - STFTM of Simultaneous Conjugate Gradient estimate
Figure 7.10 - a) Original signal, 3554 expansion for b) OSCC, and c) OS-Newab
Figure 7.11a - Desired MSTFTM

Figure 7.11b - STFTM of OS-CG estimate

Figure 7.11c - STFTM of OS-Nawab estimate
Figure 7.13a - Desired MSTFTM

Figure 7.13b - STFTM of LSEE-CG estimate

Figure 7.13c - STFTM of LSEE-Nawab estimate
<table>
<thead>
<tr>
<th>No distortion</th>
<th>Noise</th>
<th>Periodic</th>
<th>No glitches</th>
<th>No wideband</th>
<th>No glitches</th>
<th>No wideband</th>
<th>No glitches</th>
<th>No wideband</th>
<th>No glitches</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Learning loss</td>
<td>Spectrogram</td>
<td>Waveform</td>
<td>D&lt;sup&gt;2&lt;/sup&gt;</td>
<td>D&lt;sup&gt;1&lt;/sup&gt;</td>
<td>D&lt;sup&gt;1&lt;/sup&gt;</td>
<td>D&lt;sup&gt;1&lt;/sup&gt;</td>
<td>D&lt;sup&gt;1&lt;/sup&gt;</td>
<td>D&lt;sup&gt;1&lt;/sup&gt;</td>
<td>D&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

Table 7.3: Performance Comparison for 2:1 Compression Followed by 1:2 Expansion
Figure 7.14 - (a) Original signal, (b) compression followed by L2, (c) L2E-MSTIM, (d) OA-MSTIM, and (e) L2E-Nawab expansion for (b) L2E-MSTIM, (c) OA-MSTIM, and (d) L2E-Nawab.
Figure 7.15a - STFTM of Original signal

Figure 7.15b - Desired MSTFTM (1:2 expansion of Figure 7.3b)

Figure 7.15c - STFTM of LSEE-MSTFTM estimate
Figure 7.16a - STFTM of Original signal

Figure 7.16b - Desired MSTFTM (1.2 expansion of Figure 7.3c)

Figure 7.16c - STFTM of OA-MSTFTM estimate
Figure 7.17a - STFTM of Original signal

Figure 7.17b - Desired MSTFTM (1:2 expansion of Figure 7.7c)

Figure 7.17c - STFTM of LSEE-Nawab estimate
Chapter 8

Conclusions

8.1. Summary

In this thesis, we have developed a number of methods for estimating a signal from modified STFT (MSTFT) or modified STFT magnitude (MSTFTM). A performance comparison for these methods is given in Table 8.1. In the next section, we will summarize the results for signal estimation from MSTFT. Then, in the following section, the results for signal estimation from MSTFTM will be summarized.

8.1.1. Summary of Signal Estimation from MSTFT

In Chapter 2, we developed the LSEE-MSTFT method which estimates a signal with STFT closest to a given MSTFT in terms of a least-squares distance measure. This algorithm is very similar to the overlap-add algorithm except that the inverse FT for each frame is windowed before overlap-adding. In Chapter 5, we showed that LSEE-MSTFT reconstructs the original signal exactly from its STFT. In Chapter 6, we compared LSEE-MSTFT with the overlap-add algorithm for speech enhancement by spectral subtraction and found that the overlap-add algorithm introduces slightly more smearing in time than LSEE-MSTFT since the effective synthesis window length is longer. However, the quality of the signal estimates produced by these two methods was comparable in informal listening tests. The LSEE-MSTFT algorithm was also used in Chapter 3 to develop new algorithms for signal estimation from MSTFTM.
<table>
<thead>
<tr>
<th>Method</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSTFT Methods</td>
<td>Theoretically minimizes least-squares distance measure. Performance approximately the same as overlap-add for spectral subtraction. Can be used to develop an iterative algorithm to estimate a signal from MSTFTM.</td>
</tr>
<tr>
<td>LSEE-MSTFT</td>
<td>Ad hoc method does not minimize least-squares distance measure. Performance approximately the same as LSEE-MSTFT.</td>
</tr>
<tr>
<td>MSTFTM Simultaneous Methods</td>
<td></td>
</tr>
<tr>
<td>LSEE-MSTFT</td>
<td>Estimates a signal from MSTFTM iteratively using LSEE-MSTFT. Guaranteed to converge to the critical points of a magnitude distance measure. Produces a good signal estimate with no glitches. Can be implemented in real-time using cascaded processors, one for each iteration.</td>
</tr>
<tr>
<td>OA-MSTFT</td>
<td>Estimates a signal from MSTFTM iteratively using overlap-add. Is not guaranteed to converge to the critical points of any magnitude distance measure. Produces a good signal estimate with no glitches. Can be implemented in real-time using cascaded processors, one for each iteration.</td>
</tr>
<tr>
<td>CG</td>
<td>Estimates a signal from MSTFTM using the conjugate gradient method. Converges extremely slowly due to the high dimension of the minimization problem. Produces a signal estimate with some glitches due to convergence to non-global minima.</td>
</tr>
<tr>
<td>MSTFTM Sequential Methods</td>
<td></td>
</tr>
<tr>
<td>OS-CG</td>
<td>Overlap-save conjugate gradient method for estimating a signal from MSTFTM frame by frame. Converges more slowly than Nawab's method and obtains a poor signal estimate.</td>
</tr>
<tr>
<td>OS-Nawab</td>
<td>Overlap-save combined with Nawab's method. Converges faster than conjugate gradient and obtains a better signal estimate. However, the signal estimate still tends to contain glitches resulting in some distortion.</td>
</tr>
<tr>
<td>LSEE-CG</td>
<td>Uses conjugate gradient to estimate the phase to combine with the MSTFTM, then estimates a signal from the MSTFT using LSEE-MSTFT. Produces a somewhat better signal estimate than OS-CG, but due to the poor estimate produced by the conjugate gradient algorithm, the signal estimate is still of poor quality.</td>
</tr>
<tr>
<td>LSEE-Nawab</td>
<td>Uses Nawab's method to estimate the phase to combine with the MSTFTM, then estimates a signal from the MSTFT using LSEE-MSTFT. Produces the best estimate of all of the sequential algorithms and has performance comparable to LSEE-MSTFTM and OA-MSTFTM and better than simultaneous conjugate gradient.</td>
</tr>
</tbody>
</table>
8.1.2. Summary of Signal Estimation from MSTFTM

In Chapter 3, we developed "simultaneous" and "sequential" methods for estimating a signal from MSTFTM. The "simultaneous" methods estimate the signal by decreasing a distance function over all windows simultaneously. Of these methods, LSEE-MSTFTM and OA-MSTFTM produce a better estimate in fewer iterations for the applications of signal reconstruction from unmodified STFTM, speech enhancement by spectral subtraction, and time-scale modification of speech in the examples considered. In addition, these two algorithms can be implemented in real time by cascading processors, one for each iteration.

The "sequential" methods estimate the signal by combining estimates produced by decreasing distance functions of individual windows of the MSTFTM. In Chapters 5, 6, and 7, we showed that Nawab's method combined with LSEE-MSTFT produces a higher quality speech estimate than the other sequential methods for the applications of signal reconstruction from unmodified STFTM, speech enhancement by spectral subtraction, and time-scale modification of speech in the examples considered.

Comparisons between LSEE-MSTFTM, OA-MSTFTM, and LSEE-Nawab indicate that these algorithms produce signal estimates with comparable quality for the applications considered. The LSEE-Nawab algorithm requires less computation and less I/O and so is better suited for implementation on general purpose machines. However, LSEE-MSTFTM and OA-MSTFTM are much better suited for real time applications since they can be implemented with cascaded processors.

8.2. Suggestions for Further Research

A number of issues which require further research are raised by the MSTFTM algorithms. These include investigating methods for increasing the convergence rate of these algorithms and obtaining a better initial estimate. In addition, these algorithms allow development
of new methods that only need to process the STFT magnitude of signals. For example, for applications such as helium speech enhancement the STFTM of the original signal can be modified in any way desired without considering what modifications of the STFT phase are required. Another possible application is bit rate reduction for speech communication systems. For this application, the magnitude could either be directly encoded or the linear prediction coefficients and the magnitude of the residual could be encoded. One criterion for the usefulness of MSTFTM algorithms as opposed to more conventional algorithms is that the STFTM of the signal estimates produced by the conventional algorithms must differ significantly from the desired MSTFTM. This criterion is met for time-scale modification of speech where such algorithms as Fairbanks' method produce an estimate with STFTM far from the desired MSTFTM. For speech enhancement by spectral subtraction, the overlap-add method produces an estimate with STFTM very close to the desired MSTFTM and consequently, very little improvement is possible using methods which estimate the signal directly from the MSTFTM.
Appendix A

Convergence Proof for LSEE-MSTFTM

In this appendix, we show that the LSEE-MSTFTM algorithm converges to a set consisting of the critical points of Equation (3.1). This can be shown using the following Global Convergence Theorem [13].

Let \( A \) be an algorithm on \( \mathbb{R}^N \), and suppose that, given \( x^0(n) \) the sequence \( \{x^i(n)\}_{i=0}^\infty \) is generated satisfying

\[
x^{i+1}(n) = A[x^i(n)]
\]

Let a solution set \( \Gamma \subset \mathbb{R}^N \) be given, and suppose

i) all signal estimates \( x^i(n) \) are contained in a compact set \( X \subset \mathbb{R}^N \).

ii) there is a continuous distance measure \( D \) on \( \mathbb{R}^N \) such that

a) if \( x^i(n) \) is not an element of \( \Gamma \), then \( D[x^{i+1}(n)] < D[x^i(n)] \)

b) if \( x^i(n) \) is an element of \( \Gamma \), then \( D[x^{i+1}(n)] \leq D[x^i(n)] \)

iii) the mapping \( A \) is closed at points outside \( \Gamma \).

Then the limit of any convergent subsequence of \( \{x^i(n)\} \) is a solution.
The first requirement of the Global Convergence Theorem is that all estimates are contained in the compact set \( X \). Defining

\[
X = \frac{\sum_{m=-\infty}^{\infty} w(mS-n) \frac{1}{2\pi} \int_{-\pi}^{\pi} |Y_w(mS,\omega)| e^{j(\Theta_y(mS,\omega)+\omega n)} d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}
\]  

(A1)

where \( \Theta_y(mS,\omega) \in [-\pi, \pi] \) for all \( m, \omega \)

We will show that \( X \) is compact since it is both closed and bounded. In order to ensure that \( x(n) \) is a finite length sequence, the given MISTFTM \( |Y_w(mS,\omega)| \) will be assumed to be zero outside of a given range of \( m \). In Equation (A1), \( X \) has been expressed as a continuous function of the closed set consisting of the phase angles \( \Theta_y(mS,\omega) \) which indicates that \( X \) is closed.

We can further show that \( X \) is bounded as follows:

\[
|x(n)| \leq \frac{\sum_{m=-\infty}^{\infty} w(mS-n) \frac{1}{2\pi} \int_{-\pi}^{\pi} |Y_w(mS,\omega)| e^{j(\Theta_y(mS,\omega)+\omega n)} d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}
\]  

(A2)

Equation (A2) leads to

\[
|x(n)| \leq \frac{\sum_{m=-\infty}^{\infty} w(mS-n) \frac{1}{2\pi} \int_{-\pi}^{\pi} |Y_w(mS,\omega)| d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}
\]  

(A3)

where

\[ x(n) \in X \]

Therefore, since \( \frac{1}{2\pi} \int_{-\pi}^{\pi} |Y_w(mS,\omega)| d\omega \) is bounded and the sum over \( m \) reduces to a finite sum for any single value of \( n \), then \( x(n) \) is bounded and so is the set \( X \).

The second requirement is the existence of a distance measure \( D \) for a solution set \( \Gamma \) and the algorithm \( A \) that satisfies ii) of the Global Convergence Theorem. Using the distance measure of Equation (2.4), \( \hat{x}_w^i(mS,\omega) \) of Equation (3.3) minimizes \( D[x^i(n), \hat{x}_w(mS,\omega)] \) for
$x^i(n)$ fixed and $\hat{x}_w^i(mS,\omega)$ constrained to have magnitude $|Y_w(mS,\omega)|$. Thus, we must have

$$D[x^i(n), \hat{x}_w^i(mS,\omega)] \leq D[x^i(n), \hat{x}_w^{i-1}(mS,\omega)]$$

(A4)

and

$$D[x^{i+1}(n), \hat{x}_w^{i+1}(mS,\omega)] \leq D[x^{i+1}(n), \hat{x}_w^i(mS,\omega)]$$

(A5)

Now, the modified STFT $\hat{x}_w^i(mS,\omega)$ is available which allows estimation of the next signal $x^{i+1}(n)$ using LSEE-MSTFT. Since this procedure minimizes $D[x(n), \hat{x}_w^i(mS,\omega)]$ for $\hat{x}_w^i(mS,\omega)$ fixed, we have

$$D[x^{i+1}(n), \hat{x}_w^i(mS,\omega)] \leq D[x^i(n), \hat{x}_w^i(mS,\omega)]$$

(A6)

and equality holds if and only if $x^{i+1}(n)=x^i(n)$. Combining Equations (A5) and (A6), we obtain

$$D[x^{i+1}(n), \hat{x}_w^{i+1}(mS,\omega)] \leq D[x^i(n), \hat{x}_w^i(mS,\omega)]$$

(A7)

![Figure A1 - LSEE-MSTFTM Iteration](image)

and equality holds if and only if $x^{i+1}(n)=x^i(n)$. Figure A1 shows $D[x^i(n), \hat{x}_w^i(mS,\omega)]$ as segment a, $D[x^{i+1}(n), \hat{x}_w^i(mS,\omega)]$ as segment b, and $D[x^{i+1}(n), \hat{x}_w^{i+1}(mS,\omega)]$ as segment c. Since $X^i_w(mS,\omega)$ and $\hat{x}_w^i(mS,\omega)$ have the same phase, the distance represented by segment a is
equivalent to the distance between $|X_w^i(mS,\omega)|$ and $|Y_w(mS,\omega)|$. This can be shown by writing $D[x^i(n), \hat{X}_w^i(mS,\omega)]$ explicitly

$$D[x^i(n), \hat{X}_w^i(mS,\omega)] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} |X_w^i(mS,\omega) - |Y_w(mS,\omega)|| \frac{X_w^i(mS,\omega)}{|X_w^i(mS,\omega)|} |^2 d\omega$$  \hspace{1cm} (A8)

which reduces to

$$D[x^i(n), \hat{X}_w^i(mS,\omega)] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{-\pi}^{\pi} [|X_w^i(mS,\omega)| - |Y_w(mS,\omega)|]^2 d\omega$$  \hspace{1cm} (A9)

This leads to the definition of the distance function based on STFT magnitude given by Equation (3.1). Since $D_1[x^i(n), |Y_w(mS,\omega)|] = D[x^i(n), \hat{X}_w^i(mS,\omega)]$, Equation (A7) can be written as

$$D_1[x^{i+1}(n), |Y_w(mS,\omega)|] \leq D_1[x^i(n), |Y_w(mS,\omega)|]$$  \hspace{1cm} (A10)

and equality holds if and only if $x^{i+1}(n) = x^i(n)$. Taking the gradient of $D_1$ with respect to $x(n)$ yields

$$\nabla D_1[x(n), |Y_w(mS,\omega)|] = 2[x^i(n) - x^{i+1}(n)] \sum_{m=-\infty}^{\infty} w^2(mS - n)$$  \hspace{1cm} (A11)

Since the gradient is a constant times the difference between successive estimates, the solution set $\Gamma$ corresponds to the zeros of the gradient of $D_1$. So, if $x^i(n)$ is not an element of $\Gamma$, then $x^{i+1}(n) \neq x^i(n)$ and $D_1[x^{i+1}(n), |Y_w(mS,\omega)|] < D_1[x^i(n), |Y_w(mS,\omega)|]$. If $x^i(n)$ is an element of $\Gamma$, then $x^{i+1}(n) = x^i(n)$ and $D_1[x^{i+1}(n), |Y_w(mS,\omega)|] \leq D_1[x^i(n), |Y_w(mS,\omega)|]$.

The final requirement for convergence is that the mapping $A$ be closed. Since $A$ is a continuous function of $x(n)$, it must be a closed mapping which satisfies iii) of the Global Convergence Theorem. Thus, LSEE-MSTFTM converges to a solution set consisting of the critical points of the STFT magnitude distance measure $D_1$. 

Appendix B

Convergence Proof for Nawab's Algorithm

To prove convergence of this algorithm, we use the following Global Convergence Theorem [13].

Let $A$ be an algorithm on $R^N$, and suppose that, given $x^0(n)$ the sequence $\{x^i(n)\}_{i=0}^{\infty}$ is generated satisfying

$$x^{i+1}(n) = A[x^i(n)]$$

Let a solution set $\Gamma \subset R^N$ be given, and suppose

i) all signal estimates $x^i(n)$ are contained in a compact set $X \subset R^N$.

ii) there is a continuous distance measure $D$ on $R^N$ such that

   a) if $x^i(n) \notin \Gamma$, then $D[x^{i+1}(n)] < D[x^i(n)]$

   b) if $x^i(n) \in \Gamma$, then $D[x^{i+1}(n)] \leq D[x^i(n)]$

iii) the mapping $A$ is closed at points outside $\Gamma$.

Then the limit of any convergent subsequence of $\{x^i(n)\}$ is a solution.

The first requirement of the Global Convergence Theorem is that all estimates are con-
tained in the compact set $X$. Defining

$$X = \begin{cases} \frac{1}{2\pi} \int_{\omega = -\pi}^{\pi} |Y_w(m_0S,\omega)| e^{j(\Theta_y(m_0S,\omega) + \omega n)} d\omega, & m_0S - L + M < n \leq m_0S \\ c(n), & \text{otherwise} \end{cases}$$

where $\Theta_y(m_0S,\omega) \in [-\pi, \pi]$ for all $\omega$.

We will show that $X$ is compact since it is both closed and bounded. In Equation (B1), $X$ has been expressed as a continuous function of the closed set consisting of the phase angles $\Theta_y(m_0S,\omega)$ which indicates that $X$ is closed. We can further show that $X$ is bounded as follows:

$$|x(n)| \leq \begin{cases} \frac{1}{2\pi} \int_{\omega = -\pi}^{\pi} |Y_w(m_0S,\omega)| d\omega, & m_0S - L + M < n \leq m_0S \\ |c(n)|, & \text{otherwise} \end{cases}$$

where

$$x(n) \in X$$

Therefore, since $\frac{1}{2\pi} \int_{\omega = -\pi}^{\pi} |Y_w(m_0S,\omega)| d\omega$ and $|c(n)|$ are bounded, $x(n)$ is bounded and so is the set $X$.

The second requirement is the existence of a distance measure $D$ for a solution set $\Gamma$ and the algorithm $A$ that satisfies ii) of the Global Convergence Theorem. Defining this distance measure as:

$$D[x_w(m_0S,n), \hat{x}_w(m_0S,\omega)] = \frac{1}{2\pi} \int_{\omega = -\pi}^{\pi} |X_w(m_0S,\omega) - \hat{X}_w(m_0S,\omega)|^2 d\omega$$

$\hat{x}_w(m_0S,\omega)$ of Equation (3.32) is the minimum of $D[x_w(m_0S,n), \hat{x}_w(m_0S,\omega)]$ for $x_w(m_0S,n)$ fixed and $\hat{x}_w(m_0S,\omega)$ constrained to have magnitude $|Y_w(m_0S,\omega)|$. Thus, we must have:

$$D[x_w(m_0S,n), \hat{x}_w(m_0S,\omega)] \leq D[x_w(m_0S,n), \hat{x}_w^{-1}(m_0S,\omega)]$$

and

$$D[x_w(m_0S,n), \hat{x}_w^l(m_0S,\omega)] \leq D[x_w(m_0S,n), \hat{x}_w^{l-1}(m_0S,\omega)]$$
Now, \( x_w^{i+1}(m_0S,n) \) is calculated from \( \hat{X}_w^i(m_0S,\omega) \) using Equation (3.31). Since this procedure minimizes \( D[x_w(m_0S,n), \hat{X}_w^i(m_0S,\omega)] \) for \( \hat{X}_w^i(m_0S,\omega) \) fixed and \( x_w(m_0S,n) \) constrained to be \( c(n) \) for \( n > m_0S \) or \( n = m_0S - L + M \), we have

\[
D[x_w^{i+1}(m_0S,n), \hat{X}_w^i(m_0S,\omega)] \leq D[x_w^i(m_0S,n), \hat{X}_w^i(m_0S,\omega)] \tag{B6}
\]

and equality holds if and only if \( x_w^{i+1}(m_0S,n) = x_w^i(m_0S,n) \). Combining Equations (B5) and (B6), we obtain

\[
D[x_w^{i+1}(m_0S,n), \hat{X}_w^{i+1}(m_0S,\omega)] \leq D[x_w^i(m_0S,n), \hat{X}_w^i(m_0S,\omega)] \tag{B7}
\]

![Figure B1 - Iteration of Nawab's Algorithm](image)

and equality holds if and only if \( x_w^{i+1}(m_0S,n) = x_w^i(m_0S,n) \). Figure B1 shows \( D[x_w^i(m_0S,n), \hat{X}_w^i(m_0S,\omega)] \) as segment a, \( D[x_w^{i+1}(m_0S,n), \hat{X}_w^i(m_0S,\omega)] \) as segment b, and \( D[x_w^{i+1}(m_0S,n), \hat{X}_w^{i+1}(m_0S,\omega)] \) as segment c. Since \( X_w^i(m_0S,\omega) \) and \( \hat{X}_w^i(m_0S,\omega) \) have the same phase, the distance represented by segment a is equivalent to the distance between \( |X_w^i(m_0S,\omega)| \) and \( |Y_w(m_0S,\omega)| \). This can be shown by writing \( D[x_w^i(m_0S,n), \hat{X}_w^i(m_0S,\omega)] \) explicitly.
\[ D[x_w^i(m_0S,n), \hat{x}_w^i(m_0S,\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left| X_w^i(m_0S,\omega) - Y_w(m_0S,\omega) \right| \left| \frac{X_w^i(m_0S,\omega)}{|X_w^i(m_0S,\omega)|} \right|^2 d\omega \]

which reduces to

\[ D[x_w^i(m_0S,n), \hat{x}_w^i(m_0S,\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \left[ |X_w^i(m_0S,\omega)| - |Y_w(m_0S,\omega)| \right]^2 d\omega \] (B9)

This leads to the definition of the distance function based on STFT magnitude given by Equation (3.18) for \( p = 1 \). Since \( D_1[x_w^i(m_0S,n), |Y_w(m_0S,\omega)|] = D[x_w^i(m_0S,n), \hat{x}_w^i(m_0S,\omega)] \), Equation (B7) can be written as

\[ D_1[x_w^{i+1}(m_0S,n), |Y_w(m_0S,\omega)|] \leq D_1[x_w^i(m_0S,n), |Y_w(m_0S,\omega)|] \] (B10)

and equality holds if and only if \( x_w^{i+1}(m_0S,n) = x_w^i(m_0S,n) \). Taking the gradient of \( D_1 \) with respect to \( x_w(m_0S,n) \) yields

\[ \nabla D_1[x_w^i(m_0S,n), |Y_w(m_0S,\omega)|] = 2[x_w^i(m_0S,n) - x_w^{i+1}(m_0S,n)] \] (B11)

Since the gradient is a constant times the difference between successive estimates, the solution set \( \Gamma \) corresponds to the zeros of the gradient of \( D_1 \). So, if \( x_w^i(m_0S,n) \notin \Gamma \), then \( x_w^{i+1}(m_0S,n) \neq x_w^i(m_0S,n) \) and

\[ D_1[x_w^{i+1}(m_0S,n), |Y_w(m_0S,\omega)|] < D_1[x_w^i(m_0S,n), |Y_w(m_0S,\omega)|] \]. If \( x_w^i(m_0S,n) \in \Gamma \), then \( x_w^{i+1}(m_0S,n) = x_w^i(m_0S,n) \) and

\[ D_1[x_w^{i+1}(m_0S,n), |Y_w(m_0S,\omega)|] \leq D_1[x_w^i(m_0S,n), |Y_w(m_0S,\omega)|]. \]

The final requirement for convergence is that the mapping \( A \) be closed. Since \( A \) is a continuous function of \( x(n) \), it must be a closed mapping which satisfies iii) of the Global Convergence Theorem. Thus, Nawab's algorithm converges to a solution set consisting of the critical points of \( D_1[x_w(m_0S,n), |Y_w(m_0S,\omega)|] \) for the constraints \( c(n) \).
References


