SIGNAL ESTIMATION FROM MODIFIED SHORT-TIME FOURIER TRANSFORM*

Daniel W. Griffin
Jae S. Lim

Research Laboratory of Electronics
Department of Electrical Engineering and Computer Science, 36-653
Massachusetts Institute of Technology
Cambridge, Massachusetts

ABSTRACT

In this paper, we present an algorithm to estimate a signal from its modified short-time Fourier transform (STFT). This algorithm is computationally simple and is obtained by minimizing the mean squared error between the STFT of the estimated signal and the modified STFT. Using this algorithm, we also develop an iterative algorithm to estimate a signal from its modified STFT magnitude. The iterative algorithm is shown to decrease, in each iteration, the mean squared error between the STFT magnitude of the estimated signal and the modified STFT magnitude. The major computation involved in the iterative algorithm is the discrete Fourier transform (DFT) computation, and the algorithm appears to be real-time implementable with current hardware technology. The algorithm developed in this paper has been applied to the time-scale modification of speech. The resulting system generates very high-quality speech, and appears to be better in performance than any existing method.

I. 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 the time-scale modification of speech, one approach is to modify the STFTM and then to 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 whose STFT or STFTM is close in some sense to the MSTFT or MSTFTM. Previous approaches to this problem, such as the overlap-add method [6,7], have been mostly heuristic and have been limited to estimating a signal from the MSTFT. In this paper, we develop new algorithms based on theoretical grounds to estimate a signal from the MSTFT or the MSTFTM. In addition, the new algorithm is applied to the problem of time-scale modification of speech. The resulting system is considerably simpler computationally and has significantly better performance than the system described by Portnoff [1].

The paper is organized as follows. In Section II, we develop an algorithm to estimate a signal from the MSTFT by minimizing the mean squared error between the STFT of the estimated signal and the MSTFT. In Secton III, the algorithm in Section II is used to develop an iterative algorithm that estimates a signal from the MSTFTM. In Section IV, we present the application of our theoretical results to the problem of time-scale modification of speech.

II. SIGNAL ESTIMATION FROM MODIFIED SHORT-TIME FOURIER TRANSFORM

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 rate 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 $0 \le n \le L-1$. From the definition of the STFT,

$$X_w(mS,\omega) = F_l[x_w(mS,l)] = \sum_{l=-\infty}^{\infty} x_w(mS,l)e^{-j\omega l} \quad (1)$$

where

$$x_w(mS,l) = w(mS-l)x(l) \quad (2)$$

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

$$y_w(mS,l) = \frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} Y_w(mS,\omega)e^{j\omega l}\,d\omega \quad (3)$$

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 $Y_w(mS,\omega)$:

$$D[x(n),Y_w(mS,\omega)] = \quad (4)$$
$$\sum_{m=-\infty}^{\infty} \frac{1}{2\pi}\int_{\omega=-\pi}^{\pi} |X_w(mS,\omega) - Y_w(mS,\omega)|^2 d\omega$$

The distance measure in Equation (4) 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 (4) can be written as

*This work has been supported in part by the National Science Foundation under Grant ECS80-07102 and in part by the Advanced Research Projects Agency monitored by ONR under Contract N00014-81-K-0742 NR-049-506.

CH1841-6/83/0000-0804 $1.00 © 1983 IEEE

$$D[x(n),Y_w(mS,\omega)] = \sum_{m=-\infty}^{\infty}\sum_{l=-\infty}^{\infty}[x_w(mS,l) - y_w(mS,l)]^2 \tag{5}$$

Since Equation (5) is in the quadratic form of $x(n)$, minimizing $D[x(n),Y_w(mS,\omega)]$ with respect to $x(n)$ leads to the following result:

$$x(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n)y_w(mS,n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)} \tag{6}$$

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

$$x(n) = \frac{\sum_{m=-\infty}^{\infty} y_w(mS,n)}{\sum_{m=-\infty}^{\infty} w(mS-n)} \tag{7}$$

The weighted overlap-add procedure can be expressed as

$$x(n) = \sum_{m=-\infty}^{\infty} f(mS-n)y_w(mS,n) \tag{8}$$

for some "synthesis" filter $f(n)$. The major difference between Equations (6) and (7) is that Equation (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 (6) and (8) is that Equation (6) explicitly specifies what $f(n)$ is and has the normalization constant. In addition, the major difference between Equation (6) and Equations (7) and (8) is that Equation (6) was theoretically derived explicitly for the purpose of estimating a signal from the MSTFT while Equations (7) and (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 computatonal 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 (6) can be implemented with little on-line storage and delay, in the same manner [9] as the standard overlap-add procedure of Equation (7) or the weighted overlap-add procedure of Equation (8). Since the algorithm represented by Equation (6) minimizes the distance measure of Equation (4), it will be referred to as LSEE-MSTFT, the least squares error estimation from the MSTFT.

In the standard overlap-add method, the window is usually normalized so that $\sum_{m=-\infty}^{\infty} w(mS-n)$ is unity for all n in order to reduce computation. As in the overlap-add method, the window in Equation (6) can be normalized so that $\sum_{m=-\infty}^{\infty} w^2(mS-n)$ is unity for all n. Any non-zero window can be normalized in this manner for maximum window overlap ($S=1$). For partial window overlap, however, the window is more restricted. One particular class of window useful for the partial overlap case is the sinusoidal window. Specifically, if the window length (L) is a multiple of 4 times the window shift (S), then the sinusoidal window defined by

$$w_s(n) = \frac{2w_r(n)}{\sqrt{4a^2+2b^2}}\left[a + b\cos\left[\frac{2\pi n}{L} + \phi\right]\right] \tag{9}$$

where $w_r(n)$ is a rectangular window that has amplitude $\sqrt{S}/\sqrt{L}$ and is zero outside the range $0 \leq n < L$, has the desired property. By choosing values for a, b, and ϕ, windows similar to the Hamming window and the Hanning window can be obtained. Thus, the modified Hamming window used for time-scale modification of speech in Section IV will be defined as Equation (9) for $a=.54$, $b=-.46$, and $\phi=\frac{\pi}{L}$. Further discussion of the class of windows that do not require normalization càn be found in Reference [10].

Estimating $x(n)$ based on Equation (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 section, Equation (6) can also be used to develop an iterative algorithm that estimates a signal from the MSTFTM.

III. SIGNAL ESTIMATION FROM MODIFIED STFT MAGNITUDE

In this section, we consider the problem of estimating $x(n)$ from the modified STFT magnitude $|Y_w(mS,\omega)|$. The algorithm we develop is an iterative procedure in which the squared error between $|X_w(mS,\omega)|$ and $|Y_w(mS,\omega)|$ is decreased in each iteration. Let $x^i(n)$ denote the estimated $x(n)$ after the i^{th} iteration. The $i+1^{\text{st}}$ 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 Equation (6). This iterative algorithm is illustrated in Figure 1. It can be shown [10] that the algorithm in

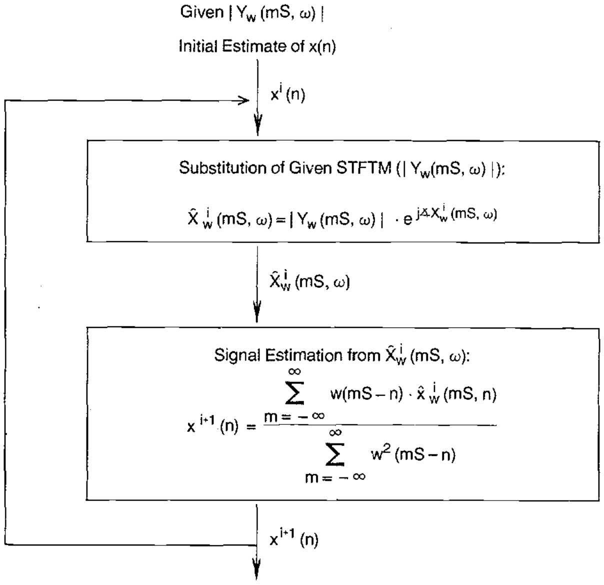
Given $|Y_w(mS,\omega)|$
Initial Estimate of x(n)
$x^i(n)$
Substitution of Given STFTM ($|Y_w(mS,\omega)|$):
$\hat{X}_w^i(mS,\omega) = |Y_w(mS,\omega)| \cdot e^{j\measuredangle X_w^i(mS,\omega)}$
$\hat{X}_w^i(mS,\omega)$
Signal Estimation from $\hat{X}_w^i(mS,\omega)$:
$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS-n)\cdot\hat{x}_w^i(mS,n)}{\sum_{m=-\infty}^{\infty} w^2(mS-n)}$
$x^{i+1}(n)$

Fig. 1. LSEE-STFTM algorithm.

Figure 1 decreases in each iteration the following distance measure:

$$D_M[x(n),|Y_w(mS,\omega)|] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} [|X_w(mS,\omega)| - |Y_w(mS,\omega)|]^2 d\omega \tag{10}$$

It can also be shown [10] that the algorithm always converges to a set consisting of the critical points of the distance measure D_M as a function of $x(n)$. This algorithm will be referred to as LSEE-MSTFTM.

It is possible to develop ad-hoc methods to estimate $x(n)$ from the MSTFTM by modifying the iterative algorithm in Figure 1. For example, we could use in the signal estimation step of the iterative procedure the standard overlap-add method of Equation (7) rather than the LSEE-MSTFT method in obtaining the next estimate $x^{i+1}(n)$ from the MSTFT $\hat{X}_w^i(mS,\omega)$. With this modification, the 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_M. However, as will be discussed in Section IV, OA-MSTFTM does appear to reduce D_M enough to produce a reasonable signal estimate for the purposes of time-scale modification of speech.

IV. Time-Scale Modification of Speech

One method of decomposing a speech signal $y(n)$ is to represent it as the convolution of an excitation function with the 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.

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, and hence the rate at which $|Y_w(mS,\omega)|$ varies with time, without affecting the spectral characteristics. This can be accomplished by estimating a signal with STFT magnitude close to a time-scale modified version of $|Y_w(mS,\omega)|$. A time-scale modification of $S_1 : S_2$ can be performed by calculating $|Y_w(mS_1,\omega)|$ at the window shift S_1 and $X_w^i(mS_2,\omega)$ at the window shift S_2 in the LSEE-MSTFTM or OA-MSTFTM algorithms. For example, $|Y_w(mS_1,\omega)|$ for the sentence "Line up at the screen door," sampled at 10 kHz is shown in Figure 2 for a 256 point modified Hamming window and a window shift S_1 of 128. Figure 3a shows a 128:64 time-scale modified version of $|Y_w(mS_1,\omega)|$ produced by displaying these samples of $|Y_w(n,\omega)|$ with a spacing of 64 samples instead of 128 samples. A signal with STFTM close to this MSTFTM was estimated by starting with an initial white Gaussian noise sequence and then iterating with LSEE-MSTFTM until the distance measure D_M was decreased to the desired level. The Fourier transforms in the algorithm were implemented with 512 point FFTs. Figure 3b shows $|X_w^i(mS_2,\omega)|$ for $S_2 = 64$ after 100 iterations. Similarly, Figure 3c shows $|X_w^i(mS_2,\omega)|$ after 100 iterations of the OA-MSTFTM algorithm using the same initial estimate. Comparisons of Figures 3b and 3c with Figure 3a indicate that the STFTM of the signal estimate is very close to the desired MSTFTM and that the performance of LSEE-MSTFTM and OA-MSTFTM is similar. In Figure 4, the distance measure D_M is shown as a function of the iteration for LSEE-MSTFTM and OA-MSTFTM. Although OA-MSTFTM performs somewhat better during the initial iterations, LSEE-MSTFTM eventually surpasses it. This same performance difference was noted in all of the examples where these two methods were compared. In addition, LSEE-MSTFTM was observed to always decrease D_M, whereas OA-MSTFTM usually stopped decreasing D_M after about 100 iterations, and in some cases increased D_M as more iterations were performed.

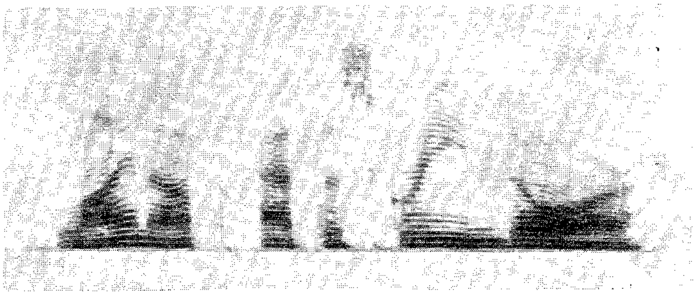

Fig. 2. STFTM of "Line up at the screen door".

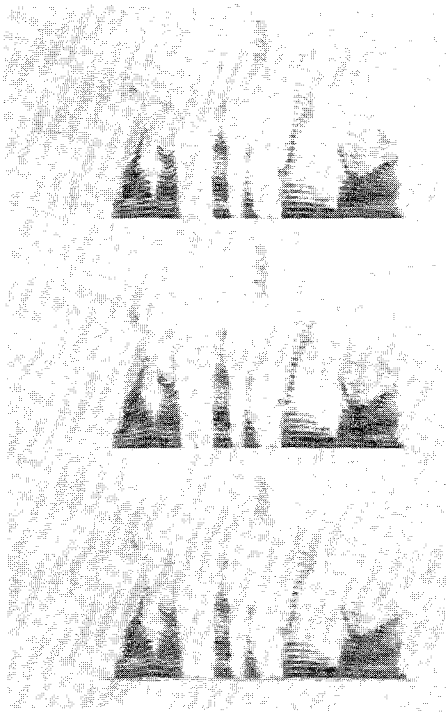

Fig. 3(a) 128 : 64 time-scale compressed STFTM of original speech

(b) STFTM of LSEE-MSTFTM estimate

(c) STFTM of OA-MSTFTM estimate

In addition to the above example, other speech material, including noisy speech, was processed by the two methods at various compression and expansion ratios. Informal listening tests clearly indicated that the performance of these methods is superior to that of Portnoff's system [1]. Note that this

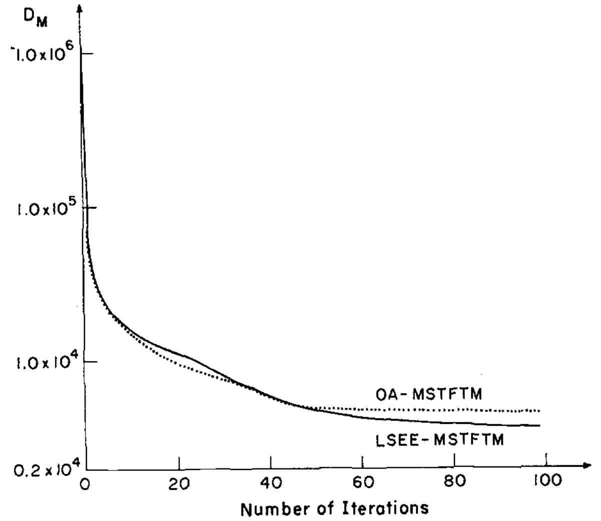

Fig. 4. D_M as a function of the number of iterations.

approach to time-scale modification of speech differs considerably from that of Portnoff. In Portnoff's method, the phase of $Y_w(mS, \omega)$ is explicitly obtained by phase unwrapping, which is undesirable due to various considerations, including the computational aspect. In the LSEE-MSTFTM or OA-MSTFTM algorithms, the phase of $Y_w(mS, \omega)$ is implicitly estimated in the process of estimating a signal with STFTM close to $|Y_w(mS, \omega)|$, and no phase unwrapping is performed.

Even though a large number of iterations (50 to 100) are typically needed to obtain the best possible performance from the algorithm, we have observed that speech quality improves rapidly initially and then more slowly as the number of iterations increases. This is evidenced, to some extent, in Figure 4, where D_M decreases rapidly initially but more slowly as the number of iterations increases. With a better choice of the initial estimate of $x(n)$ than a Gaussian noise sequence, it may be possible to reduce the number of iterations required to achieve a specified performance level.

Despite the large number of iterations required, real time implementation appears possible if enough processors are used in series. Specifically, as input data is received, the i^{th} processor can perform the i^{th} iteration and the $i+1^{st}$ processor which follows the i^{th} processor can perform the $i+1^{st}$ iteration. The inherent delay associated with each iteration is only the length of the analysis window, L data points. This is due to the fact that the computational aspect of each iteration of the algorithm is essentially the same as the weighted overlap-add method [9], in which the delay between the input and output data is L points assuming the required computation for each windowed data segment can be performed during the time corresponding to the window shift, S data points. As an example that illustrates the computational requirements and delay involved, suppose $S_1 = S_2 = 64$, $L = 256$, the size of the DFT used is 512, the number of iterations required and the number of processors available is 50, and speech is sampled at a 10 kHz rate. Since the major computations involved in the algorithm are due to the DFT and the IDFT, if each processor can compute two 512-point DFTs once every 6.4 msec, then the iterative algorithm can be implemented in real time with a delay of about 1.3 seconds. Current hardware technology is more than capable of handling such computational requirements, and a delay of a few seconds is not a serious problem in most applications of time scale modification of speech.

Even though LSEE-MSTFTM and OA-MSTFTM had similar performance in the context of time-scale modification of speech, it should be pointed out that LSEE-MSTFTM decreases the distance measure D_M of Equation (14) in each iteration until it converges to a critical point, while OA-MSTFTM sometimes increases D_M. In all cases we considered so far, LSEE-MSTFTM always produced a smaller D_M than OA-MSTFTM after a sufficiently large number of iterations. This difference may be significant in other applications.

In this paper, we considered the application of our theoretical results only to the problem of time-scale modification of speech. The application of these results to other problems, such as enhancement of speech degraded by helium, is currently under study, and these results will be reported in a later paper.

V. SUMMARY

Three new algorithms have been presented in this paper. The first, LSEE-MSTFT, estimates a signal with STFT closest to a MSTFT and is similar to the overlap-add method. The second, LSEE-MSTFTM, is an iterative algorithm based on LSEE-MSTFT which was shown to converge to a solution set consisting of the critical points of a magnitude only distance measure. A third algorithm, OA-MSTFTM, is heuristically developed based on the overlap-add method. LSEE-MSTFTM and OA-MSTFTM were applied to time-scale modification of speech with results superior to the method developed by Portnoff [1].

REFERENCES

[1] M.R. Portnoff, "Time-Scale Modification of Speech Based on Short-Time Fourier Analysis," IEEE TRANS. ASSP. Vol. ASSP-29, pp. 374-390:June 1981.

[2] J.S. Lim, "Enhancement and Bandwidth Compression of Noisy Speech," PROC. IEEE. Vol. 67, pp. 1586-1604:Dec. 1979.

[3] J.S. Lim (ed.), SPEECH ENHANCEMENT, Englewood Cliffs, NJ: Prentice-Hall, 1983.

[4] J.B. Allen, D.A. Berkley, and J. Blauert, "Multimicrophone Signal-Processing Techniques to Remove Room Reverberation from Speech Signals," J. ACOUST. SOC. AMER. Vol. 62, pp. 912-915:Oct. 1977.

[5] M.A. Richards, "Helium Speech Enhancment Using the Short-Time Fourier Transform," IEEE TRANS. ASSP. Vol. ASSP-30, pp. 841-853:Dec. 1982.

[6] L.R. Rabiner and R.W. Schafer, DIGITAL PROCESSING OF SPEECH SIGNALS, Englewood Cliffs, NJ: Prentice-Hall, 1978.

[7] J.B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE TRANS. ASSP. Vol. ASSP-25, pp. 235-238:June 1977.

[8] M.R. Portnoff, "Time-Frequency Representation of Digital Signals and Systems Based on Short-Time Fourier Analysis," IEEE TRANS. ASSP. Vol. ASSP-28, pp. 55-69:Feb. 1980.

[9] R.E. Crochiere, "A Weighted Overlap-Add Method of Short-Time Fourier Analysis/Synthesis," IEEE TRANS. ASSP. Vol. ASSP-28, pp. 99-102, Feb. 1980.

[10] D.W. Griffin and J.S. Lim, "Signal Estimation from Modified Short-Time Fourier Transform," submitted to IEEE TRANS. ASSP., 1982.