Signal Estimation from Modified Short-Time Fourier Transform

DANIEL W. GRIFFIN AND JAE S. LIM, SENIOR MEMBER, IEEE

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 have been mostly heuristic [6]–[8], and have been limited to estimating a signal from the MSTFT [6], [7]. 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 conceptually and appears to have 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. The resulting algorithm is quite simple computationally. In Section III, the algorithm in Section II is used to develop an iterative algorithm that estimates a signal from the MSTFTM. The iterative algorithm is shown to decrease, in each iteration, the mean squared error between the STFTM of the estimated signal and the MSTFTM. In Section IV, we present an example of the successful application of our theoretical results. Specifically, we develop a time-scale speech modification system by modifying the STFTM first and then estimating a signal from the MSTFTM using the algorithm developed in Section III. The resulting system has been demonstrated to generate very high quality, time-scale modified speech.

Manuscript received December 27, 1982; revised May 12, 1983, and September 26, 1983. This work was supported in part by the Advanced Research Projects Agency monitored by ONR under Contract N00014-81-K-0742 NR-049-509 and the National Science Foundation under Grant ECS80-07102.

The authors are with the Research Laboratory of Electronics, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139.

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 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 nonzero for $0 \leq n \leq L - 1$. From the definition of the STFT

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

where

$$x_w(mS, l) = w(mS - l)\, x(l) \tag{2}$$

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

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

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 $Y_w(mS, \omega)$:

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

0096-3518/84/0400-0236$01.00 © 1984 IEEE

The distance measure in (4) is the squared error between $X_w(mS, \omega)$ and $Y_w(mS, \omega)$ integrated over all ω and summed over all m. It 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, (4) can be written as

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

Since (5) is in the quadratic form of $x(n)$, minimization of $D[x(n), Y_w(mS, \omega)]$ can be accomplished by setting the gradient with respect to $x(n)$ to zero and solving for $x(n)$ which leads to the following result:

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

This solution is similar in form to the standard overlap-add procedure [6], [7], or the weighted overlap-add procedure [9], [10]. The overlap-add procedure can be expressed as

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

The weighted overlap-add procedure can be expressed as

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

for some "synthesis" filter $f(n)$. The major difference between (6) and (7) is that (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 (6) and (8) is that (6) explicitly specifies what $f(n)$ is and has the normalization constant. In addition, the major difference between (6), and (7) and (8), is that (6) was theoretically derived explicitly for the purpose of estimating a signal from the MSTFT based on the least squares error criterion of (4). Equations (7) and (8), however, were derived to reconstruct a signal from its exact STFT or to estimate a signal from the MSTFT for a very restricted class of modifications, and were sometimes used as ad hoc methods 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, (6) can be implemented with little on-line storage and delay, in the same manner [10] as the standard-overlap procedure of (7) or the weighted overlap-add procedure of (8). Since the algorithm represented by (6) minimizes the distance measure of (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 $\Sigma_{m=-\infty}^{\infty} w(mS - n)$ is unity for all n in order to reduce computation. As in the overlap-add method, the window in (6) can be normalized so that $\Sigma_{m=-\infty}^{\infty} w^2(mS - n)$ is unity for all n. Any nonzero window can be normalized in this manner for maximum window overlap ($S = 1$). 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

$$w_r(n) = \begin{cases} \frac{\sqrt{S}}{\sqrt{L}}, & 0 \leq n < L \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

has the property

$$\sum_{m=-\infty}^{\infty} w_r^2(mS - n) = \sum_{m=0}^{(L/S)-1} \frac{S}{L} = 1. \quad (10)$$

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

$$w_s(n) = \frac{2w_r(n)}{\sqrt{4a^2 + 2b^2}} \left[a + b \cos\left(\frac{2\pi n}{L} + \phi\right)\right] \quad (11)$$

has the property given by (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 $\phi = \pi/L$. 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 time-scale modification of speech in Section IV will be defined as (11) for $a = 0.54$, $b = -0.46$, and $\phi = \pi/L$. 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 (11) for $a = 0.5$, $b = -0.5$, and $\phi = \pi/L$. Use of these modified windows eliminates the need for normalizing by $\Sigma_{m=-\infty}^{\infty} w^2(mS - n)$ in (6), which reduces computation and/or storage requirements for partial window overlap.

Estimating $x(n)$ based on (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, (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 based on the LSEE-MSTFT algorithm which is similar in style to several other iterative algorithms [11], [12]. In this algorithm, 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 ith iteration. The $i + 1$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 (6). The

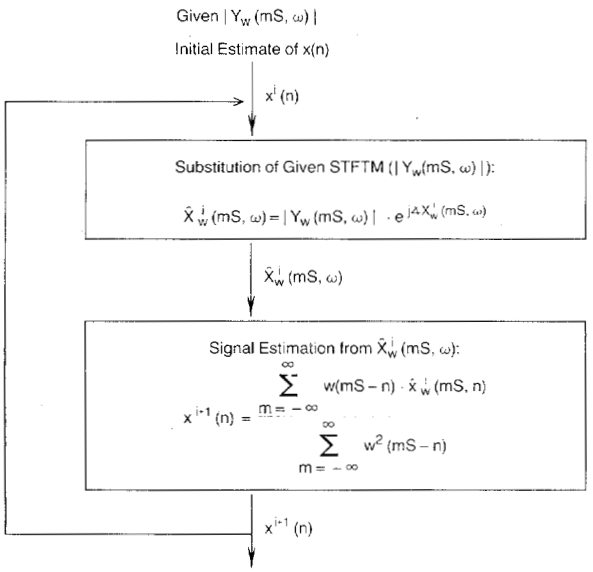


Fig. 1. LSEE-MSTFTM algorithm.

iterative algorithm, which is illustrated in Fig. 1, results in the following update equation:

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS - n) \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \hat{X}_w^i(mS, \omega)\, e^{j\omega n}\, d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS - n)} \tag{12}$$

where

$$\hat{X}_w^i(mS, \omega) = \left|Y_w(mS, \omega)\right| \frac{X_w^i(mS, \omega)}{\left|X_w^i(mS, \omega)\right|}. \tag{13}$$

In (13), if $\left|X_w^i(mS, \omega)\right| = 0$, then $\hat{X}_w^i(mS, \omega)$ is set to $\left|Y_w(mS, \omega)\right|$. It can be shown (see Appendix) that the algorithm in Fig. 1 decreases in each iteration the following distance measure:

$$D_M[x(n), \left|Y_w(mS, \omega)\right|] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} [\left|X_w(mS, \omega)\right| - \left|Y_w(mS, \omega)\right|]^2\, d\omega. \tag{14}$$

It can also be shown (see Appendix) 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 Fig. 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:

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \hat{X}_w^i(mS, \omega)\, e^{j\omega n}\, d\omega}{\sum_{m=-\infty}^{\infty} w(mS - n)} \tag{15}$$

where $\hat{X}_w^i(mS, \omega)$ is given by (13). 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_M. As will be shown in Section IV, however, 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 $\left|Y_w(mS, \omega)\right|$ can be written as the product of a component due to the excitation function $\left|P_w(mS, \omega)\right|$ and a component due to the vocal tract impulse response $\left|H_w(mS, \omega)\right|$. 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 $\left|P_w(mS, \omega)\right|$ will correspond to the rapidly varying portion of $\left|Y_w(mS, \omega)\right|$ with ω, taking on an harmonic structure for voiced speech or noise for unvoiced speech. The function $\left|H_w(mS, \omega)\right|$ will correspond to the slowly varying portion of $\left|Y_w(mS, \omega)\right|$ with ω, and will include the formant information of the speech signal. Since the speech signal is assumed to be approximately stationary over the window length, $\left|P_w(mS, \omega)\right|$ and $\left|H_w(mS, \omega)\right|$ 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 $\left|P_w(mS, \omega)\right|$ and $\left|H_w(mS, \omega)\right|$ vary with time, and hence the rate at which $\left|Y_w(mS, \omega)\right|$ 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 $\left|Y_w(mS, \omega)\right|$. A time-scale modification of $S_1 : S_2$ can be performed by calculating $\left|Y_w(mS_1, \omega)\right|$ 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, $\left|Y_w(mS_1, \omega)\right|$ for the sentence "line up at the screen door." sampled at 10 kHz is shown in Fig. 2 for a 256 point modified Hamming window and a window shift S_1 of 128. Fig. 3(a) shows a 128:64 time-scale modified version of $\left|Y_w(mS_1, \omega)\right|$ produced by displaying these samples of $\left|Y_w(n, \omega)\right|$ 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 FFT's. Fig. 3(b) shows $\left|X_w^i(mS_2, \omega)\right|$ for $S_2 = 64$ after 100 iterations. Similarly, Fig. 3(c) shows $\left|X_w^i(mS_2, \omega)\right|$ after 100 iterations of the OA-MSTFTM algorithm using the same initial estimate. Comparisons of Fig. 3(b) and 3(c) with Fig. 3(a) 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 Fig. 4, the distance measure D_M is shown as a function of the number of iterations for LSEE-

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

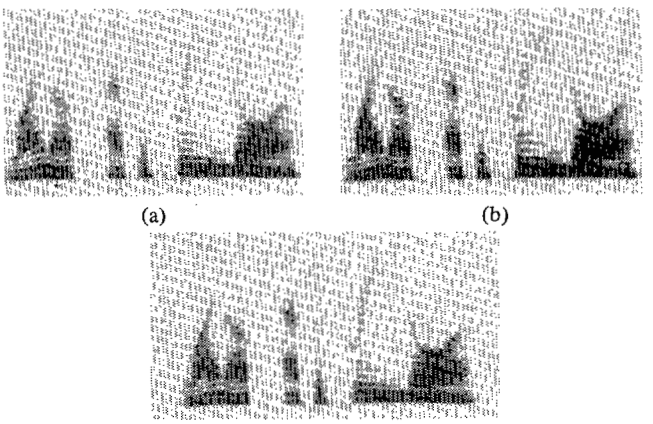


Fig. 3. (a) 128:64 time-scale compressed STFTM of original speech. (b) STFTM of LSEE-MSTFTM estimate. (c) STFTM of OA-MSTFTM estimate.

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.

To show that these methods perform as well for noninteger compression or expansion factors, the second example shows a 35:64 expansion. Fig. 5(a) shows a 35:64 time-scale modified version of $|Y_w(mS_1, \omega)|$ calculated from the original speech signal. As in the first example, the initial estimate was a white Gaussian noise sequence. Fig. 5(b) and 5(c) show the STFTM of the signal estimate after 100 iterations using a 256 point modified Hamming window for LSEE-MSTFTM and OA-MSTFTM, respectively. In both of these examples, the resultant signal estimate was clean high quality speech and the estimates produced by LSEE-MSTFTM and OA-MSTFTM were indistinguishable in listening tests.

The final example consists of a 1:2 time-scale expansion of the 2:1 time-scale compressed speech generated in the first example. The STFTM of the signal estimates produced are then compared with the STFTM of the original speech signal. Fig. 6(a) and 6(b) show the STFTM of the signal estimates after 100 iterations of LSEE-MSTFTM and OA-MSTFTM, respectively. Comparisons of Fig. 6(a) and 6(b) with Fig. 2 show that both LSEE-MSTFTM and OA-MSTFTM produce a signal estimate with STFTM close to the STFTM of the original speech signal. The primary difference between these signal estimates and the original speech signal is that a small amount

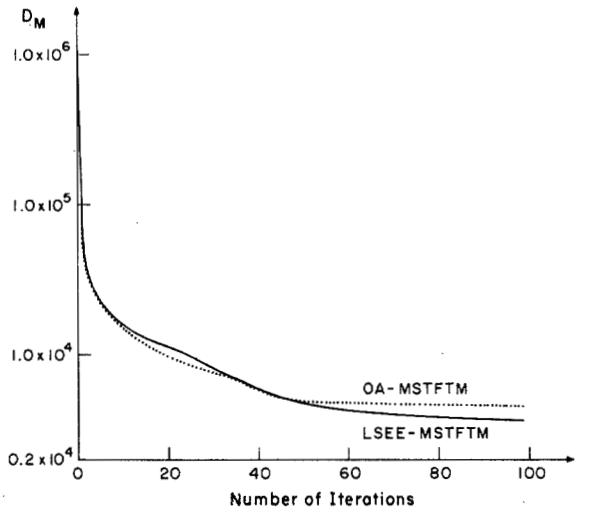


Fig. 4. D_M versus iteration number of LSEE-MSTFTM and OA-MSTFTM.

of reverberation is detectable in the signal estimate due to the nonstationarity of the 2:1 time-scale compressed speech over the window length.

In addition to the above three examples, other speech material including noisy speech has been processed by the two methods at various compression and expansion ratios. Informal listening appears to indicate that the performance of these methods is superior to that of the system by Portnoff [1]. It should be noted that this 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 we used a large number of iterations (100) for the examples illustrated in this paper, we have observed that essentially the same results in terms of speech quality can be obtained after 25 to 100 iterations. In addition, 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 Fig. 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 certain performance.

Despite the large number of iterations[1] required, real time[2]

[1]Due to iterations, the total number of computations is considerably larger than Portnoff's method [1]. In a multiprocessor environment, however, the computational requirement of each processor is comparable or perhaps less than that of Portnoff's method.

[2]The definition of "real time" for time-scale modification depends on the application. In applications where the input to the algorithm is from some storage device and the output is converted to an analog signal which the user listens to, the algorithm must produce one output sample in an average time less than T_1 where T_1 is the sampling period associated with the digital to analog converter used to generate output speech. In applications where the input to the algorithm is digitized directly from the user's speech and the output is placed on some storage device, the algorithm must process an input data sample in an average time less than T_2 where T_2 is the sampling period associated with the analog to digital converter used to digitize the input speech.

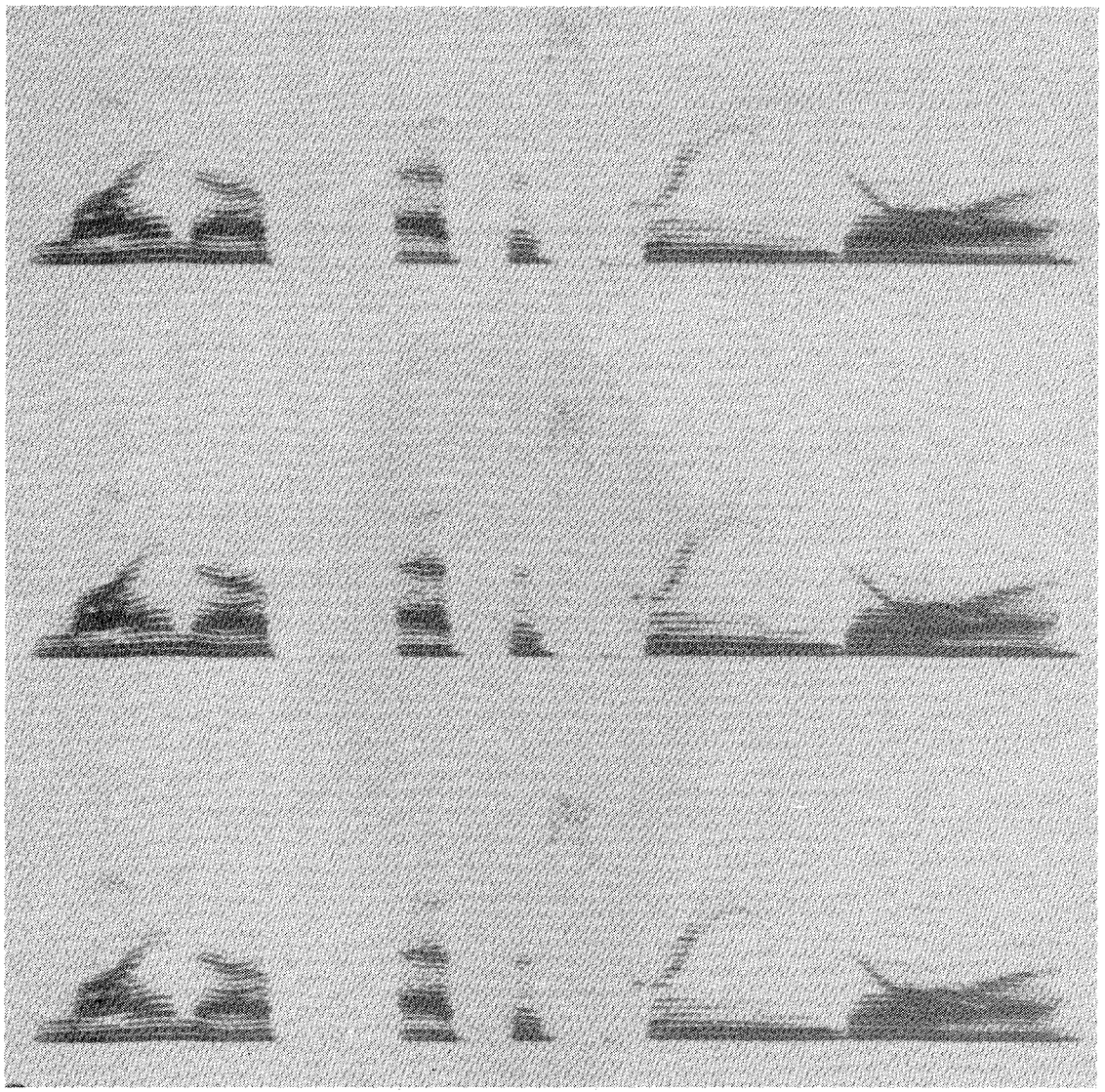


Fig. 5. (a) 35:64 time-scale expanded STFTM of original speech. (b) STFTM of LSEE-MSTFTM estimate. (c) STFTM of OA-MSTFTM estimate.

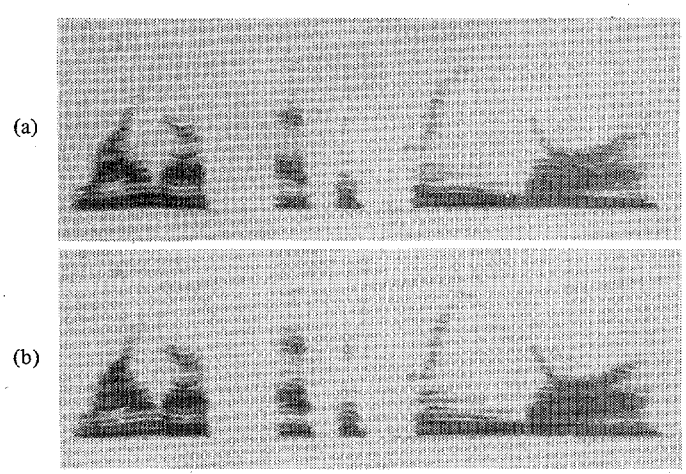


Fig. 6. 1:2 expansion of 2:1 compressed speech for (a) LSEE-MSTFTM and (b) OA-MSTFTM.

implementation appears possible if enough processors are used in series. Specifically, as input data are received, the ith processor can perform the ith iteration and the $i+1$st processor which follows the ith 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 [10], 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 IDFT, if each processor can compute two 512-point DFT's once every 6.4 ms, then the iterative algorithm can be implemented in real time with a delay of about 1.3 s. 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 (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 the theoretical results in this paper 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 algorithm, LSEE-MSTFT, estimates a signal with STFT closest to a MSTFT and is similar to the overlap-add method. The second algorithm, 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 that appear to be superior to the method developed by Portnoff [1].

Appendix

In this Appendix, we show that the iterative algorithm (LSEE-MSTFTM) in Fig. 1 decreases in each iteration the distance measure of (14) and always converges to a critical point where the gradient of the distance measure of (14) with respect to $x(n)$ is zero. It should be noted that convergence to a critical point does not necessarily imply convergence to the global minimum.

To show the above, 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 $D[x^{i+1}(n)] \leqslant D[x^i(n)]$

and iii) the mapping A is closed at points outside Γ. Then the limit of any convergent subsequence of $\{x^i(n)\}$ is in the solution set.

The first requirement of the global convergence theorem is that all estimates are contained in the compact set X. Define

$$X = \frac{\sum_{m=-\infty}^{\infty} w(mS - n) \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} |Y_w(mS, \omega)| \, e^{j(\Theta_y(mS,\omega)+\omega n)} \, d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS - n)} \tag{A1}$$

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

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 MSTFTM $|Y_w(mS, \omega)|$ will be assumed to be zero outside of a given range of m. In (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:

$$|x(n)| \leqslant \frac{\sum_{m=-\infty}^{\infty} w(mS - n) \left| \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} |Y_w(mS, \omega)| \, e^{j(\Theta_y(mS,\omega)+\omega n)} \, d\omega \right|}{\sum_{m=-\infty}^{\infty} w^2(mS - n)}. \tag{A2}$$

Equation (A2) leads to

$$|x(n)| \leqslant \frac{\sum_{m=-\infty}^{\infty} w(mS - n) \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} |Y_w(mS, \omega)| \, d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS - n)} \tag{A3}$$

where

$$x(n) \in X.$$

Therefore, since $(1/2\pi) \int_{\omega=-\pi}^{\pi} |Y_w(mS, \omega)| \, d\omega$ 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 Γ and the algorithm A that satisfies ii) of the global convergence theorem. Using the distance measure of (4), $\hat{X}_w^i(mS, \omega)$ of (13) minimizes $D[x^i(n), \hat{X}_w(mS, \omega)]$ for $x^i(n)$ fixed and $\hat{X}_w(mS, \omega)$ constrained to have magnitude $|Y_w(mS, \omega)|$. Thus, we must have

$$D[x^i(n), \hat{X}_w^i(mS, \omega)] \leqslant D[x^i(n), \hat{X}_w^{i-1}(mS, \omega)] \tag{A4}$$

and

$$D[x^{i+1}(n), \hat{X}_w^{i+1}(mS, \omega)] \leqslant D[x^{i+1}(n), \hat{X}_w^i(mS, \omega)]. \tag{A5}$$

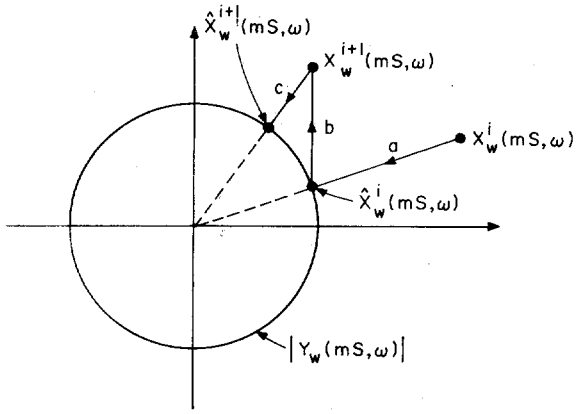


Fig. 7. Successive iterations of LSEE-MSTFTM.

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

$$D[x^{i+1}(n), \hat{X}_w^i(mS, \omega)] \leqslant D[x^i(n), \hat{X}_w^i(mS, \omega)] \tag{A6}$$

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

$$D[x^{i+1}(n), \hat{X}_w^{i+1}(mS, \omega)] \leqslant D[x^i(n), \hat{X}_w^i(mS, \omega)] \tag{A7}$$

and equality holds if and only if $x^{i+1}(n) = x^i(n)$. Fig. 7 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 $X_w^i(mS, \omega)$ 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:

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

which reduces to

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

This leads to the definition of the distance function based on STFT magnitude given by (14). Since $D_M[x^i(n), |Y_w(mS, \omega)|] = D[x^i(n), \hat{X}_w^i(mS, \omega)]$, (A7) can be written as

$$D_M[x^{i+1}(n), |Y_w(mS, \omega)|] \leqslant D_M[x^i(n), |Y_w(mS, \omega)|] \tag{A10}$$

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

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

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

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_M.

References

[1] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, 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 enhancement using the short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, 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. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 235–238, June 1977.

[8] S. H. Nawab, T. F. Quatieri, and J. S. Lim, "Signal reconstruction from short-time Fourier transform magnitude," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 986–998, Aug. 1983.

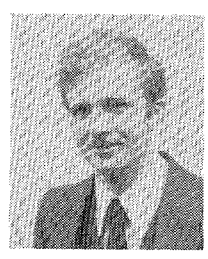
[9] M. R. Portnoff, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 55–69, Feb. 1980.

[10] R. E. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 99–102, Feb. 1980.

[11] R. W. Gerchberg and W. O. Saxton, "A practical algorithm for the determination of phase from image and diffraction plane pictures," *Optik*, vol. 35, pp. 237–246, 1972.

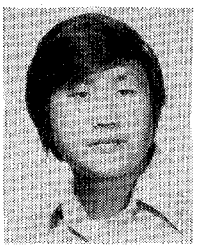
[12] A. V. Oppenheim, M. H. Hayes, and J. S. Lim, "Iterative procedures for signal reconstruction from phase," in *Proc. 1980 Int. Opt. Comp. Conf.*, Apr. 1980, vol. SPIE-231, pp. 121–129.

[13] D. G. Luenberger, *Introduction to Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley, 1973.



Daniel W. Griffin was born in Detroit, MI, on December 18, 1960. He received the B.S. degree in computer engineering from the University of Michigan, Ann Arbor, in April 1981.

He is currently a Research Assistant in the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, Cambridge. His research interests include digital signal processing and its application to speech processing.



Jae S. Lim (S'76-M'78-SM'83) was born on December 2, 1950. He received the S.B., S.M., E.E., and Sc.D. degrees in electrical engineering and computer science from the Massachusetts Institute of Technology, Cambridge, in 1974, 1975, 1978, and 1978, respectively.

He joined the MIT faculty in 1978 as an Assistant Professor, and is currently Associate Professor in the Department of Electrical Engineering and Computer Science. His research interests include digital signal processing and its applications to image and speech processing. He has contributed more than sixty articles to journals and conference proceedings, and is the editor of a reprint book, *Speech Enhancement* (Prentice-Hall, 1982).

Dr. Lim is the winner of two prize paper awards, one from the Boston chapter of the Acoustical Society of America in December 1976, and one from the IEEE ASSP Society in April 1979. He is a member of Eta Kappa Nu, Sigma Xi, and is the chairman of the IEEE ASSP Society's Technical Committee on Digital Signal Processing.

Maximum Likelihood Spectral Estimation and Its Application to Narrow-Band Speech Coding

ROBERT J. McAULAY, MEMBER, IEEE

Abstract—**Itakura and Saito [1] used the maximum likelihood (ML) method to derive a spectral matching criterion for autoregressive (i.e., all-pole) random processes. In this paper, their results are generalized to periodic processes having arbitrary model spectra. For the all-pole model, Kay's [2] covariance domain solution to the recursive ML (RML) problem is cast into the spectral domain and used to obtain the RML solution for periodic processes. When applied to speech, this leads to a method for solving the joint pitch and spectrum envelope estimation problems. It is shown that if the number of frequency power measurements greatly exceeds the model order, then the RML algorithm reduces to a pitch-directed, frequency domain version of linear predictive (LP) spectral analysis. Experiments on a real-time vocoder reveals that the RML synthetic speech has the quality of being heavily smoothed.**

I. Introduction

ITAKURA and Saito [1] have shown that spectral envelope estimation using linear predictive coding techniques (LPC) has a fundamental theoretical basis in maximum likelihood (ML) estimation. Furthermore, they have used this theory to develop a spectral matching interpretation in terms of the Itakura-Saito criterion. Their basic mathematical model dealt with speech waveforms that were sample functions of an autoregressive (AR) random process. While this is an appropriate model for the class of unvoiced sounds, one wonders if perhaps the criterion is valid for voiced speech sounds as well, since in this case the waveforms are periodic. This is the problem addressed in this paper.

In setting up the formalism for applying the ML method for periodic processes, it was not necessary to impose the all-pole constraint on the model spectrum. The analysis in Section II leads to a spectral matching criterion identical to that obtained by Itakura and Saito, which shows that the criterion is a fundamental property of the maximum likelihood method. Furthermore, it is shown that in the periodic case, the model spectrum is fitted to the power measurements at the harmonic frequencies.

In Section III extensive use is made of results obtained by Kay [2] to specialize the ML criterion to the case in which the spectral envelope is all-pole. Then in Section IV, lattice methods are used to derive a recursive maximum likelihood (RML) algorithm for estimating the AR parameters. In Section V the application of the RML technique to speech coding is described, and the results of a perceptual evaluation using a real-time analysis/synthesis system are discussed. A brief discussion is presented in Section VI on the application of the ML criterion to the joint estimation of the pitch and vocal tract spectral parameters. Finally, in Section VII, some general conclusions regarding the usefulness of the application of ML techniques to speech analysis are discussed.

II. Theoretical Formulation

By definition a real, stationary random process $s(n)$ is periodic with period N if its autocorrelation function $R(m) = E[s(n)s(n+m)]$ is periodic with period N [4]. Then $R(m)$ can be expanded by Fourier series as

$$R(m) = \sum_{k=0}^{N-1} P_k \exp(jm\omega_k) \tag{1}$$

where $\omega_k = 2\pi k/N$, and where

$$P_k = \frac{1}{N} \sum_{m=0}^{N-1} R(m) \exp(-jm\omega_k) \tag{2}$$

Manuscript received March 23, 1982; revised October 25, 1982, April 29, 1983, and August 26, 1983. This work was sponsored by the Department of the Air Force. The U.S. Government assumes no responsibility for the information presented.

The author is with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02173.

0096-3518/84/0400-0243S01.00