High Quality Time-Scale Modification for Speech

Salim Roucos and Alexander M. Wilgus

Bolt Beranek and Newman Inc.
Cambridge, MA 02238

ABSTRACT

We present a new and simple method for speech rate modification that yields high quality rate-modified speech. Earlier algorithms either required a significant amount of computation for good quality output speech or resulted in poor quality rate-modified speech. The algorithm we describe allows arbitrary linear or nonlinear scaling of the time axis. The algorithm operates in the time domain using a modified overlap-and-add (OLA) procedure on the waveform. It requires moderate computation and could be easily implemented in real time on currently available hardware. The algorithm works equally well on single voice speech, multiple-voice speech, and speech in noise. In this paper, we discuss an earlier algorithm for time-scale modification (TSM), and present both objective and informal subjective results for the new and previous TSM methods.

1. INTRODUCTION

The ability to modify the apparent rate of speech is desirable in a number of applications. For example, one can reduce the bit rate required for mediumband speech coding by time-scale compression of the input speech, followed by coding and transmission, followed by time-scale expansion to the original time scale at the receiver. Also, in voice mail systems, speech rate speedup is useful for quicker playback of received voice messages.

In time-scale modification, we wish to modify the perceived rate of speech while preserving the formant structure (for intelligibility) and the perceived pitch (for naturalness). A mathematical model for such a process is to measure the spectral envelope and the pitch of speech at a set of discrete time points $\{t_i;\ i=1,\ n\}$ and then to synthesize speech which will have approximately the same spectral envelope and pitch when measured at the warped set of time points $\{f(t_i);\ i=1,\ n\}$.

Systems for speech rate modification [1, 2, 3] differ in the representation of the spectral envelope and pitch information, the distance measure used to determine what approximate equality is, and the corresponding analysis/synthesis methods used either to extract the parameters of the representation from speech or to synthesize speech from these parameters. In this paper, we will describe the work of Griffin and Lim [4] since their algorithm, the least-squares error estimation from the modified short-time Fourier transform magnitude (LSEE-MSTFTM), is expected to have the best quality of earlier systems and because it was a basis for our research.

The LSEE-MSTFTM TSM algorithm is designed to enforce equality of the short-time Fourier transform magnitudes (STFTM) of the original and rate-modified signal, provided that those magnitudes are calculated at the corresponding time points. The STFTM contains both the spectral envelope and pitch information. Through an iterative process, the LSEE-MSTFTM algorithm produces successive signal estimates whose STFTMs are monotonically closer (using a Euclidean distance on the STFTM) to the required STFTMs [4]. The LSEE-MSTFTM algorithm, described in more detail in Section 2, produces high quality speech, but requires large computational resources.

We also report on a study of the convergence behavior of the LSEE-MSTFTM algorithm for various initial signal estimates. In an attempt to reduce the computational load by choosing good initial estimates that would require few iterations, we derived the synchronized overlap-and-add (SOLA) algorithm. The SOLA algorithm yields high quality rate-modified speech without any iterative application of the LSEE-MSTFTM algorithm. Our new algorithm will be described in Section 3. We present our conclusions in Section 4.

2. THE LSEE-MSTFTM ALGORITHM

In this section, we describe the LSEE-MSTFTM algorithm and show some results on the convergence behavior of the algorithm for various initial estimates. For convenience, we will use a notation similar to that of Griffin and Lim [4]. We will define the short-time Fourier transform (STFT) of a signal y(n) to be:

$$Y_w(mS_a,\omega)=\sum_{n=-\infty}^{\infty} w(mS_a-n)\ y(n)e^{-j\omega n} \qquad (1)$$

where w(.) is a window function and S_a is the sample shift between successive STFT computations. Suppose that the speech rate of the signal y(n) is to be changed by a rational factor $\alpha = S_s/S_a$ to yield the rate-modified speech signal x(n) ($\alpha > 1$ corresponds to slowing the speech rate and $\alpha < 1$ to increasing the speech rate). The LSEE-MSTFTM

CH2118-8/85/0000-0493 $1.00 © 1985 IEEE

algorithm will iteratively derive a signal $x^i(n)$ at the i^{th} iteration whose STFTM measured every S_s samples is monotonically closer to the STFTM of the original signal y(n) measured every S_a samples. Using an initial estimate $x^0(n)$, the LSEE-MSTFTM algorithm iteratively applies the following two steps to obtain the $i+1^{st}$ signal estimate, $x^{i+1}(n)$, from the i^{th} signal estimate, $x^i(n)$:

1. Magnitude Constraint:

$$\hat{X}_w^i(mS_s,\omega) = |Y_w(mS_a,\omega)| \frac{X_w^i(mS_s,\omega)}{|X_w^i(mS_s,\omega)|} \qquad (2)$$

where $X_w^i(mS_s,\omega)$ is the STFT of $x^i(n)$ at time mS_s. Note that the STFT of $x^i(n)$, computed every S_s points, is modified to obtain a MSTFT $\hat{X}_w^i$ that has the same magnitude as Y_w and the same phase as X_w^i. Because $\hat{X}_w^i$ is not in general a valid STFT, we must estimate in the second step of the algorithm a real signal that has the STFT closest to $\hat{X}_w^i$.

2. Least-Squares Error Estimation:

$$x^{i+1}(n) = \frac{\sum_{m=-\infty}^{\infty} w(mS_s - n) \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} \hat{X}_w^i(mS_s,\omega) e^{-j\omega n} d\omega}{\sum_{m=-\infty}^{\infty} w^2(mS_s - n)} \qquad (3)$$

The $i+1^{st}$ signal estimate is simply the least squares error estimate of the sequence of complex MSTFTs obtained in Step 1. Eq. 3 is the standard weighted mean value solution of least squares. Since each inverse transform of an MSTFT is not necessarily time-limited, the mean computation is a weighted overlap-and-add (OLA) procedure on the windowed inverse transforms of the successive MSTFTs. As it is an estimate, the LSEE step yields STFTMs that are not equal to the desired STFTMs; however, Eqs. 2 and 3 ensure the convergence of the successive estimates to the critical points of the following magnitude distance function [4]:

$$D_m[x^i(n), |Y_w(mS_a,\omega)|] = \sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} [|X_w^i(mS_s,\omega)| - |Y_w(mS_a,\omega)|]^2 d\omega \qquad (4)$$

To study the convergence of the LSEE-MSTFTM algorithm, Griffin and Lim normalized D_m to account for signal energy and called the new measure of the difference between two magnitude spectra the Signal to Error Ratio (SER):

$$SER[x(n), |Y_w(mS_a,\omega)|] = 10 \log_{10} \frac{\sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} |Y_w(mS_a,\omega)|^2 d\omega}{\sum_{m=-\infty}^{\infty} \frac{1}{2\pi} \int_{\omega=-\pi}^{\pi} [|X_w(mS_s,\omega)| - |Y_w(mS_a,\omega)|]^2 d\omega} \qquad (5)$$

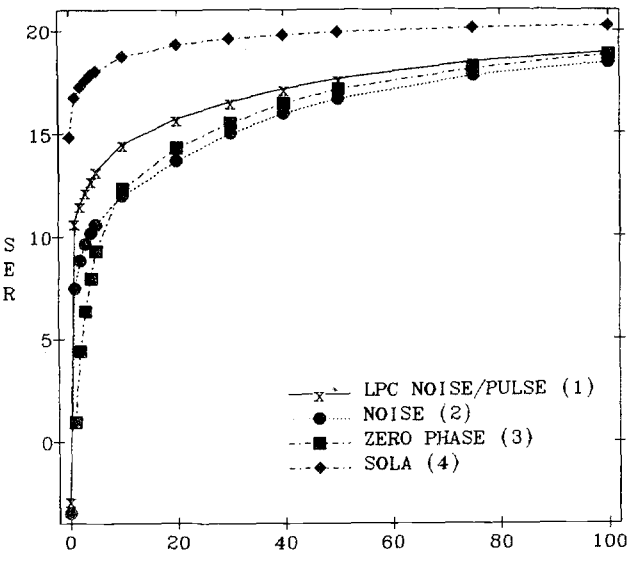

FIGURE 1. SER (dB) for various initial estimates

We will also use this parameter, though we must note that the SER by definition accounts only for the magnitudes of the successive spectra, and does not directly concern their phases.

To evaluate the LSEE-MSTFTM algorithm, we studied the convergence of the SER for various initial estimates. The initial estimate of the rate-modified signal can be made either in the time domain or in the frequency domain. In our study of the algorithm, we used three initial estimates of interest, two in the time domain and one in the frequency domain. The first time-domain initial estimate is Gaussian white noise. The second time-domain initial estimate is the noise/pulse excitation typical of LPC synthesis obtained through application of pitch extraction to the original speech. The pitch is computed every S_a samples from the original waveform and is updated every S_s samples in contsructing the initial estimate. Both initial estimates were studied by Griffin and Lim [4]. For the frequency-domain estimate, we used as an initial estimate the sequence of STFTMs $|Y_w(mS_a,\omega)|$, the STFTMs of our original input signal y(n) taken at shift S_a, thereby forcing the initial estimate to have zero phase. Our motivation for using this initial estimate will be discussed shortly.

The results of applying the MSTFTM-LSEE algorithm to each of these three estimates are shown in curves 1-3 of Fig. 1. The representative results shown are for the sentence 'Which utterance contains the fewest sonorants.' spoken by a male speaker and expanded in time by a factor of 2.0. For the objective SER measure, all three initial estimates have a poor SER (around 0 dB), and the SER is still increasing after 100 iterations. In informal listening tests, we compared rate-modified signals estimated after 10, 50, and 100 iterations using all three initial estimates. All initial

estimates still yield poor quality speech after 10 iterations. The LPC excitation initial estimate produces much better speech than the other two estimates, whose output speech sounds severely reverberant. After 50 iterations, the LPC excitation initial estimate is quite acceptable (exhibiting slight buzziness), the zero-phase estimate is still slightly reverberant, and the noise estimate is moderately reverberant. After 100 iterations, all three initial estimates do yield high quality speech; however, an extremely slight hollow-tube sound effect can be heard.

3. SYNCHRONIZED OVERLAP-ADD OF TIME SIGNALS

While the high quality rate-modified speech obtained by the LSEE-MSTFTM algorithm with the LPC excitation initial estimate is quite impressive, the associated intensive computational requirements are not generally acceptable. To reduce the computational load, we wanted to find an initial estimate that would require a very small number of iterations (e. g., less than 5) instead of the 100 iterations typically required. The search for a better initial estimate was based on an intuitive description of the behavior of the LSEE-MSTFTM algorithm.

Suppose that the chosen initial estimate is a frequency-domain estimate consisting of the STFTs of the original input signal calculated every S_a samples. The first step of the LSEE-MSTFTM algorithm will not change the STFTs, since this first magnitude normalization step consists only of multiplication by unity. The LSEE step will consist of an OLA of windows as shown in Fig. 2. To illustrate the point, we show a pulse train that corresponds to a periodic region of the original waveform. We also show three successive windows (Figs. 2-b, 2-c, and 2-d), collected at intervals of S_a samples, that have been aligned at intervals of S_s samples in preparation for the OLA step. The result given by averaging these windows is shown in Fig. 2-e. The presence of the extraneous pulses explains the tendency of the LSEE-MSTFTM algorithm to sound reverberant after only a small number of iterations. Note that the initial frequency-domain estimate starts out with the correct magnitude, but that the OLA step modifies the magnitude drastically, yielding a very low SER, a consequence of the large linear phase inconsistency between successive windows. Our proposal of the zero-phase initial estimate was an attempt to remove the phase structure of the original STFTs taken from y(n) every S_a samples. By removing this phase structure, we hoped to reduce the number of iterations required to overcome the original phase and to establish a new phase structure compatible with the new OLA shift, S_s. However, zero phase implies a certain structure in itself, so this modification was not helpful.

In the new TSM algorithm, we propose to time-align the successive windows with respect to signal similarity (magnitude and phase) before the OLA step by maximizing the time-domain crosscorrelation between successive windows. This new initial estimate is given by:

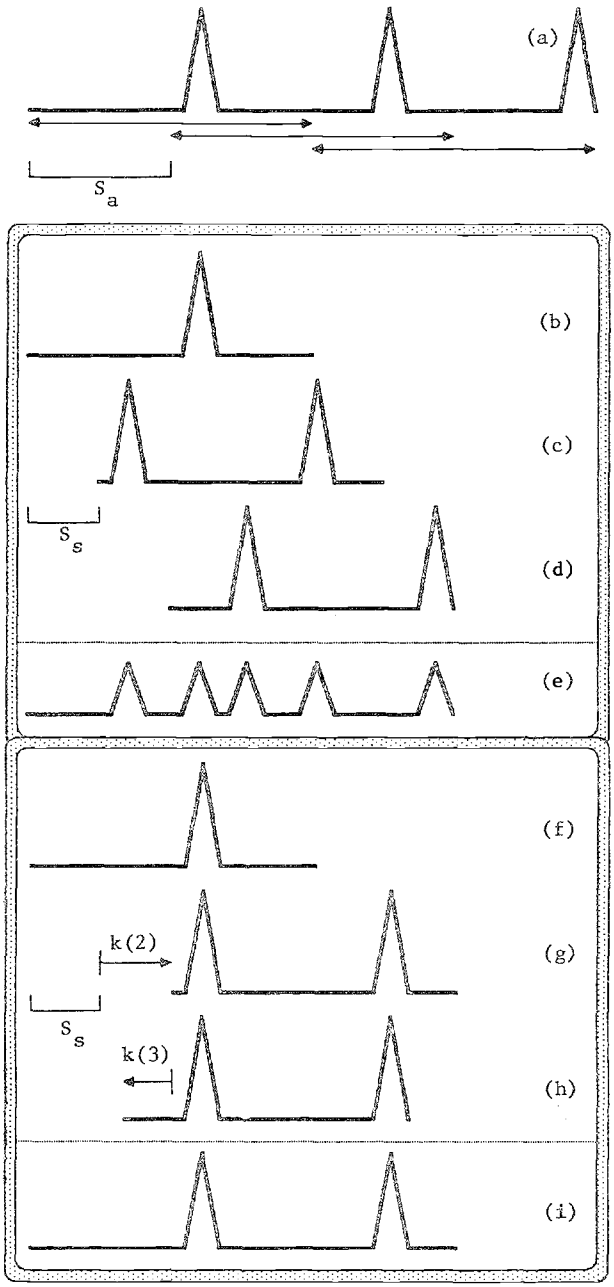

FIGURE 2. Illustrations of OLA and SOLA for TSM

$$x^o(n) = \frac{\sum_{m=-\infty}^{\infty} w^2(mS_s - n)\, y[n - m(S_s - S_a) - k(m)]}{\sum_{m=-\infty}^{\infty} w^2(mS_s - n)} \qquad (6)$$

If k(m)=0, this equation is the same as Step 2 (Eq. 3) of the LSEE described above, and we will obtain the same result shown in Fig. 2-e. However, if k(m) is chosen to maximize a crosscorrelation function, we obtain the synchronized overlap-and-

add algorithm (SOLA). The choice of k(m) is that value of k that maximizes the normalized crosscorrelation between the m^{th} window of the waveform and the rate-modified signal computed up to the $m\text{-}1^{st}$ window. Figs. 2-f, 2-g, and 2-h show the successive windows, each shifted by the appropriate k(m) to maximize the crosscorrelation. Fig. 2-i shows the average waveform that does not exhibit the extraneous pitch pulses appearing in the signal estimate of Fig. 2-e.

The algorithm is formally specified below (the symbol ':=' in the specification means the usual assignment of typical computer languages). Define a 2-D signal consisting of the successive windows (taken every S_a samples) of the original signal, denoted by $x_w(mS_s,n)$ and given by:

$$x_w(mS_s,n) := w(mS_s-n)\; y[n-m(S_s - S_a)]$$

1. Initialization:
   $x(n) := w(n)\; x_w(0,n)$
   $c(n) := w^2(n)$

2. Do steps a) and b) for m=1 to total number of frames.
   a) Maximize crosscorrelation: find k that maximizes

$$R_{xx_w}(k) = \frac{\sum_{n=mS_s}^{mS_s+\ell} x(n)\,x_w(mS_s,\, n+k)}{\left[\sum_{n=mS_s}^{mS_s+\ell} x^2(n) \sum_{n=mS_s}^{mS_s+\ell} x_w^2(mS_s,\, n+k)\right]^{1/2}}$$

   for k from -130 to -20

   b) Extend estimate by incorporating the m^{th} window:
   $x(n) := x(n) + w(mS_s+k-n)\,x_w(mS_s,n+k)$
   $c(n) := c(n) + w^2(mS_s+k-n)$

3. Normalize waveform for all n
   $x(n) := x(n)/c(n)$

The maximization of the crosscorrelation ensures that the OLA procedure will be averaging the 'next' window of the waveform with the most similar region of the reconstructed signal as it exists at that point. The variability allowed by such a scheme means that the time-scale factor of the estimate will not be exact; however, the reconstructed signal will always be within k_{max} (the range of delays allowed in crosscorrelation maximization, 110 samples in our case) samples of the ideal rate-modified signal.

The quality of this initial estimate is such that no iteration under the MSTFTM-LSEE algorithm is necessary. In informal listening tests, the SOLA initial estimate, with no iterations, was found to be at least as high in quality as the LPC excitation initial estimate after 100 iterations. It is also effective on speech in noise and on speech passages including more than one speaker. The SER of the SOLA initial estimate is shown in curve 4 of Fig. 1. While the SER is increased by roughly 5 dB, the initial and final signal estimates are subjectively indistinguishable. Also, the SOLA initial estimate, with no iterations, is far superior to the LPC excitation initial estimate after 50 iterations, despite comparable SER figures, an indication that the SER is only a gross measure of quality. The objective and subjective results for the speech of two simultaneous speakers are the same as for the single-voice utterance. We therefore conclude that the correction of the linear phase has the significant effect of reducing the distortion produced in the STFTM by the invariant OLA step.

4. CONCLUSION

We have presented a method for very high quality time-scale modification of digital speech signals that requires little computational power. The algorithm works well for both male and female speech, and for any desired time-scaling factor (including nonlinear time-scaling). It also works well for speech in noise and for multiple-voice speech passages. The method was compared with the TSM technique of Griffin and Lim, and was found to result in quality at least as high; however, SOLA TSM requires a much smaller fraction of the computations.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. John Makhoul for many stimulating discussions in the course of this research effort. The work was performed under a contract from the Department of Defense.

REFERENCES

1. J.L. Flanagan and R.M. Golden, ``Phase Vocoder'', Bell System Tech. J., Vol. 45, November 1966, pp. 1493-1509.

2. M.R. Portnoff, ``Time-Scale Modification of Speech Based on Short-Time Fourier Analysis'', IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-29, No. 3, June 1981, pp. 374-390.

3. S. Seneff, ``System to Independently Modify Excitation and/or Spectrum of Speech Waveform Without Explicit Pitch Extraction'', IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-30, No. 4, August 1982, pp. 566-578.

4. D.W. Griffin and J.S. Lim, ``Signal Estimation from Modified Short-Time Fourier Transform'', IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-32, No. 2, April 1984, pp. 236-243.