TIME-SCALE MODIFICATION OF AUDIO SIGNALS WITH COMBINED HARMONIC AND WAVELET REPRESENTATIONS *

Khaled N. Hamdy[1] *Ahmed H. Tewfik*[1] *Ting Chen*[2] *Satoshi Takagi*[3]

[1]Department of Electrical Engineering, University of Minnesota, Minneapolis, MN, USA
[2]Department of Electrical Engineering, Stanford University, Palo Alto, CA, USA
[3]Digitronics Development Department, Sony Corporation, Kanagawa,Japan

ABSTRACT

We propose a new time-scale modification method for high quality audio signals. Our approach strives to preserve pitch and timbre. In our method, the signal is represented as the sum of sinusoidal components and a residual (edges and noise). The decomposition is computed via a combined harmonic and wavelet representation. Time-scaling is performed on the harmonic components and residual components separately. The harmonic portion is time-scaled by demodulating each harmonic component to DC, interpolating and decimating the DC signal, and remodulating each component back to its original frequency. The residual portion is time-scaled by preserving edges and relative distances between the edges while time-scaling the stationary (noise) components between the edges.

1. INTRODUCTION

Time-scale modification of audio is important in music synthesis, audio/video synchronization, and commercial broadcast applications. For example, a radio station may have a three minute time slot in which to air a three minute and twenty second program without losing any content or distorting the perceptual qualities of the signal. In this paper, we present a novel method for time-scale modification of complex audio signals with little or no distortion of the pitch and timbre of the music signal. In our method, the signal is represented as the sum of sinusoidal components and a residual. The decomposition is evaluated using a combined harmonic and wavelet representation. In the wavelet domain, the residual is further separated into transient (edges) and noise components. Time-scaling is performed on the harmonic components and residual components separately. The harmonic portion is time-scaled by demodulating each harmonic component to DC, interpolating and decimating the DC signal, and remodulating each component back to its original frequency. The residual portion is time-scaled by preserving the edges and the relative distances between the edges while time-scaling the stationary (noise) components between the edges.

Several methods have been developed for time-scale modification of speech signals. Frequency based methods such as the phase vocoder [1] do not preserve the temporal characteristics of the signals. Several time domain waveform based overlap-add (OLA) methods have been proposed for the time-scale modification of speech and audio signals [2, 3]. OLA methods modify the time-scale by adding successive signal segments in a manner that maintains maximal local similarity to the original waveform. In [4], Laroche, et. al. propose to model speech signals as harmonic components and noise. The sinusoidal parameters are estimated in the time domain using a least squares approach and the residual is modeled as an AR process. McAulay and Quatieri [5] present a sinusoidal analysis method for the time-scale modification of speech. The amplitude, frequency, and phase information is extracted from the signal. The signal is scaled by resynthesizing the signal over a different length time interval. The phase parameters are adjusted using estimates of the pitch period within the framework of a vocal tract and vocal cord excitation model.

*THIS WORK WAS SUPPORTED IN PART BY TEXAS INSTRUMENTS.

2. STRUCTURE OF TIME SCALING ALGORITHM

The structure of the time-scale modification algorithm is shown in Fig. 1. The idea is to scale the signal in time by a constant scaling factor α, (where $\alpha > 1$ for expansion, and $\alpha < 1$ for compression), without modifying the pitch or timbre of the signal. The harmonic analysis/synthesis of the signal is performed as in [6]. Thomson's harmonic analysis technique is used to estimate the frequency, amplitude, and phase of each tonal component located in the signal. This information is used in the time-scale modification of the sinusoidal portion of the signal. Also, the harmonic signal is resynthesized at the original time-scale and the result is subtracted from the original signal in order to generate the residual portion of the signal, $r(n) = x(n) - s(n)$, where $x(n)$ is the original signal and $s(n)$ is the harmonic portion. This allows us to deal with the harmonic and the residual portions of the signal separately. The procedure for time-scaling the harmonic and residual portions of the signal are described in the next two sections.

Note that our approach differs from that of [4] in several respects. The pitch synchronous approaches are not as useful for general audio because of the ambiguity of the pitch for complex signals. Also, music signals tend to be very complex, containing sharp attacks or transitions. We decompose the residual into *edge-like* and *noise-like* features. In our method, we analyze the residual with a wavelet transform and perform the time-scale modifications in the wavelet domain. We treat edges in the residual signal sep-

arate from the rest of the residual signal. In particular, we do not rely on a stochastic model of the residual that does not preserve important temporal features of the signal. It is known that Fourier based or sinusoidal techniques are more efficient and accurate for the representation of tonal signals, where as wavelet based methods are more suitable for representing *edge-like* and *noise-like* signals [6]. Second, our harmonic analysis procedure is more accurate because we use Thomson's harmonic analysis as in [6] which provides a test as to whether a harmonic is statistically significant. Further, perceptually irrelevant harmonics are discarded using a frequency domain masking model. Because we model the harmonic portion of the signal as pure tones, we can use more accurate masking models for pure tones and tone complexes [7]. Moreover, we utilize information about just-noticeable differences (JND) to set a tolerance for the modification of each frequency present in the signal [7]. Finally, we rely on time-scale modification of the in-phase and quadrature components corresponding to each harmonic in the signal to modify the harmonic time-scale information.

3. TIME-SCALING OF HARMONIC COMPONENT

The pitch of an audio signal is a subjective attribute which can't be measured directly. The pitch corresponds to the frequency of a pure tone, and to the fundamental frequency of more complex audio signals such as speech. There is, however, a large class of complex music signals for which this does not hold true. In [8], it is shown that the pitch is not only a function of frequency, but also a function of level. Below 2 kHz pitch decreases as the signal level increases, and above 4 kHz pitch increases as the signal level increases.

Although the *spectral pitch* of the fundamental frequency may be heard, the *true* pitch of the signal is determined by harmonics other than the fundamental, which may or may not be present in the signal. This is known as the *residue, or virtual pitch* [9]. The residue pitch is a subharmonic of a dominant (resolvable) frequency rather than the lowest frequency. The residue pitch is heard only when at least one partial is resolvable from the tone complex. *Based on this definition of pitch it is not only important to maintain the absolute frequencies of the harmonics, but also the relative spacings between them.*

3.1. Harmonic analysis

The first part of our algorithm identifies the harmonics present in the signal. We begin by performing a harmonic analysis on the input signal, segment by segment. This yields the frequency content of the signal in each analysis segment.

The harmonic analysis of signals is carried out using a technique developed by Thomson [10] in order to detect the presence of sinusoids in a signal. In this technique, the discrete prolate sequences are used as a set of orthogonal data windows, which provide a weighted set of estimates of the frequency spectrum. Using point regression, we may then find an estimate of the harmonic mean at a given frequency. Once the estimate of the harmonic mean is obtained, an "F-test" is performed for the "goodness of fit" of the regression analysis [10]. Finally, masking is evaluated to discard irrelevant harmonics.

The output of this harmonic analysis are the frequencies, amplitudes, and phases of the sinusoids present in the signal. This information is used to compute the residual signal by direct subtraction of the non-time scaled harmonic part. It is also used to time-scale the harmonic component of the signal as explained below.

3.2. Demodulation/Time-Scale/Modulation

The time-scaling of the harmonics is performed using the structure shown in Fig. 2. The first step in this procedure consists of demodulating each harmonic detected by the harmonic analysis step down to DC. We have elected to use a coherent demodulation approach. After multiplication by sine and cosine functions of appropriate frequencies, the signals are passed through lowpass filters to extract the in-phase and quadrature components at the frequencies of interest. In our initial implementation, the bandwidths of these lowpass filters were selected based on psychoacoustic considerations. For frequencies below 2.75 kHz, the bandwidth of the lowpass filter was chosen to be 11 Hz. For frequencies between 2.75 KHz and 5.5 KHz, the bandwidth of the lowpass filter was 22 Hz. For frequencies between 5.5 KHz and 11 KHz, the bandwidth of the lowpass filter was 44 Hz. Finally, for frequencies above 11 KHz, the bandwidth of the lowpass filter was 88 Hz [11]. For all but the lowest band, these figures correspond to lower limits on the just noticeable frequency difference that our ear can perceive.

The in-phase and quadrature signals are then time-scaled appropriately using interpolation and decimation operations. Finally, the time-scaled in-phase and quadrature components are modulated back to their original frequency. Note that since interpolation and decimation do not affect DC (zero frequency), the harmonic content of the signal is preserved. Note however, that the shape of the main lobe around each dominant harmonic is affected. Our experiments so far, indicate that this is not a problem.

Note also, that this approach to time-scaling the harmonics avoids the phase discontinuities associated with earlier time-scale approaches that relied on a harmonic decomposition of the signal. Phase discontinuities occur typically at segment boundaries from interpolation of phase from segment to segment. This is a more severe problem for complex audio/music signals. In [5], the excitation phase is modified by the pitch period estimate in each frame. In speech, the pitch period is related to the fundamental frequency. This, however, is not necessarily the case with audio signals. The pitch information in music requires a more accurate model.

3.3. Frequency Tracking

As in [12], we need to track the harmonic components of each audio segment. Two types of tracking are required. We need to start and stop harmonic components as they appear and disappear in a smooth manner. This entails proper windowing of the in-phase and quadrature signals close to the time instants where harmonics must appear or disappear.

We also need to track slow variations in the frequency of the harmonic components. We disregard variations that are smaller than the just noticeable frequency difference corre-

sponding to each of the frequencies that we are tracking. Larger changes in frequencies are used to slowly change the frequencies of the cosine and sine demodulation and modulation signals used in our coherent demodulators and modulators. In particular, we use cubic phase interpolation to slowly change the frequencies of the cosine and sine demodulating and modulating signals. The interpolation guarantees that the proper frequency is used in the middle of each analysis segment.

4. TIME-SCALING OF RESIDUAL COMPONENT

The residual component, r, contains the transient (edge) and noise information about the signal. The edges are related to attacks of musical notes, transitions, or non-harmonic instruments such as castanets, drums and other percussive instruments. Such information may be related to temporal aspects of a music signal such as tempo and timbre.

Special care must be taken when manipulating the time-scale of the residual component. First, it is important to preserve the shape or slope of the attacks (edges). If the slope is not preserved, the instruments tend to sound *dull* because the high frequency information is lost. Second, it is important to preserve the relative distances between the edges while maintaining synchronization with the harmonic component. This contains the information relative to tempo. In the next section, we provide an example that illustrates the difficulties that can occur when edges are not treated separately.

To extract the edges we proceed as follows. The signal is analyzed using a wavelet packet decomposition whose frequency bands are designed to correspond with the critical band structure of the human auditory system. We determine the locations of the edges in each band by subdividing each analysis segment and comparing the relative energy within each of the subsegments. The high energy regions are labeled as edge regions. A binary representation of the residual is then computed. In that representation, ones correspond to edge regions and zeros to non edge regions. The resulting signal is analyzed using an approach similar to the one discussed in Section 3 to detect periodicities in the edge information. The time-scaled signal is finally obtained by reproducing the edge information at the locations specified by a coherent time-scale modification of the binary waveform that represents edge locations. The coherent time-scale modification of the binary waveform is similar to the approach of Section 3 except that it relies on Hadamard transforms and generally involves longer time analysis windows.

The *noise* coefficients (non-edge residual wavelet transform coefficients) are time-scaled using simple interpolation and decimation. The total residual is then resynthesized with the inverse wavelet packet transform.

5. RESULTS

Figures 3 and 4 shows a segment of a signal with 4 sinusoids present. After time-scaling there is a slight alteration in the shape of the spectrum but the center frequencies remain constant. There is also little distortion in the temporal characteristics of the sinusoids. Informal listening tests showed there was no difference in pitch.

Fig. 5 illustrates the difficulties with a straightforward approach that does not separate the residual into edges and noise. In this figure, the residual corresponding to a square pulse is scaled in time by first performing a subband decomposition of the residual using a wavelet transform. Compression was by a factor of 10 percent, i.e., $\alpha = 0.9$. Each band is time-scaled by a fractional sampling alteration scheme with factor $\alpha = M/N$, where M is the interpolation factor and N is the decimation factor. The residual is then resynthesized using the corresponding synthesis filter bank. Note that in Fig. 5 the slope of the edge is affected by the fractional sampling scheme. The degradation of the edge increases for more extreme scaling factors. For real music signals characterized by sharp transients such as castanets or drums, the effect is a dulling of the perceived sound. Thus, when scaling the residual it is important to separate the *edge* component from the *noise* component. This shows that we must deal with the edge and noise components separately.

REFERENCES

[1] M. Portnoff, "Time-Scale Modifications of Speech Based on Short-Time Fourier Analysis," *IEEE Trans. on ASSP,* Vol. ASSP-29, No. 3, June 1981.

[2] S. Yim and B. Pawate, "Computationally Efficient Algorithm for Time Scale Modification (GLS-TSM)", *ICASSP-96, Atlanta, GA.*

[3] W. Verhelst and M. Roelands, "An Overlap-Add Technique Based on Waveform Similarity(WSOLA) for High Quality Time-Scale Modification of Speech," *ICASSP-93, Minneapolis, MN*, Vol. II, p. 554-557.

[4] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech Modification Based on a Harmonic + Noise Model," *ICASSP-93, Minneapolis, MN*, Vol. II, p. 550-553.

[5] T. Quatieri and R. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech," *IEEE Trans. on Sig. Proc.,* Vol. 40, No. 3, March 1992, p. 497-510.

[6] K. Hamdy, M. Ali, and A. Tewfik, "High Quality Audio Coding of Audio Signals with a Combined Harmonic and Wavelet Representation," *ICASSP-96, Atlanta, GA.*

[7] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models,* Springer Verlag, 1990.

[8] B. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, 1989.

[9] E. Terhardt, G. Stoll, and M. Seewann, "Algorithm for Extraction of Pitch and Pitch Salience from Complex Tonal Signals," JASA, Vol. 71, No. 3, March 1982, p. 679-688.

[10] D. Thomson, "Spectrum Estimation and Harmonic Analysis," *Proceedings of the IEEE,* Vol. 70, No. 9, Sept. 1982, p. 1055-1096.

[11] M. Ali, *Adaptive Signal Representation with Application in Audio Coding*, Ph.D. Thesis, University of Minnesota, 1995.

[12] R. McAulay and T. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. on ASSP,* Vol. ASSP-34, No. 4, Aug. 1986, p. 744-754.

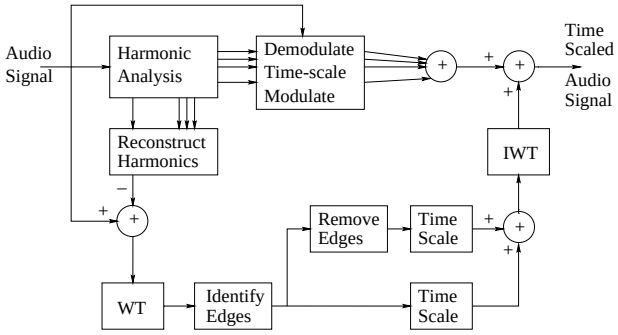

Figure 1. Block Diagram of time-scaling algorithm

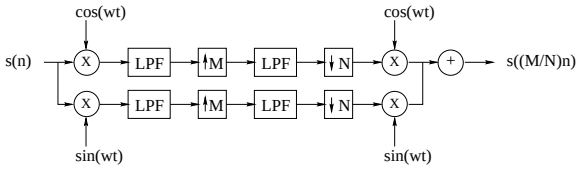

Figure 2. Block diagram of harmonic component scaling procedure.

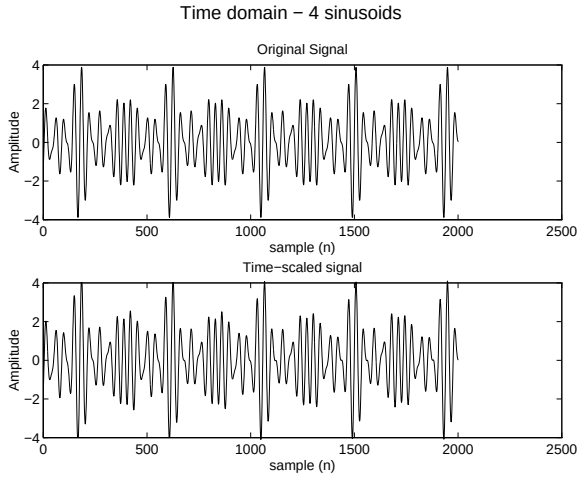

Figure 3. Time Domain - 4 sinusoids (1, 1.1, 1.3, 1.5 kHz)

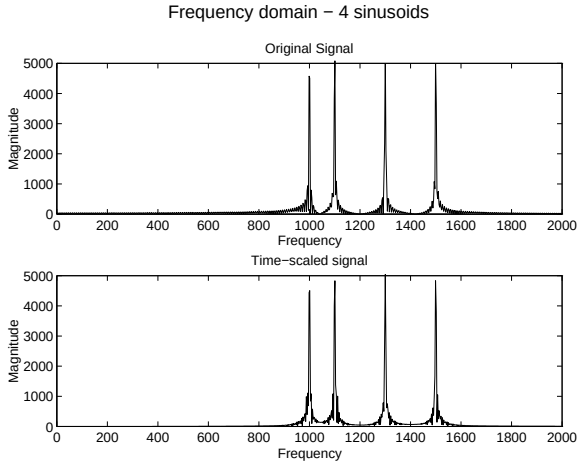

Figure 4. Frequency Domain - 4 sinusoids (1, 1.1, 1.3, 1.5 kHz)

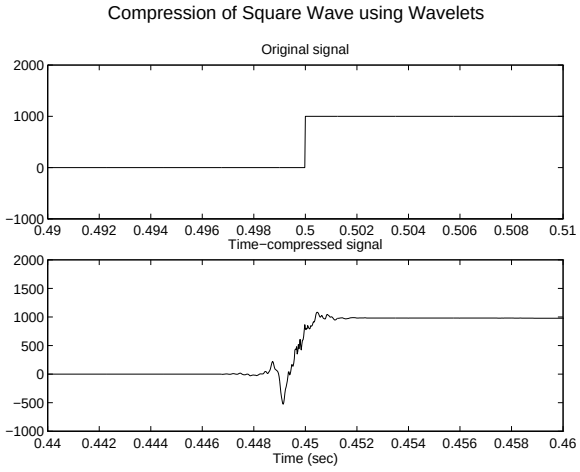

Figure 5. Time compression of square wave – zoomed in

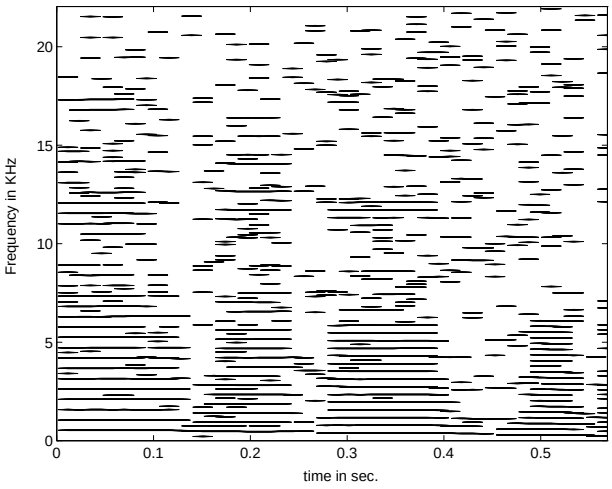

Figure 6. Frequency tracking of about 0.6 seconds from clarinet