High Quality and Low Complexity Pitch Modification of Acoustic Signals

Gang-Janp Lin, Sau-Gee Chen & Terry Wu
Department of Electronics Engineering & Institute of Electronics
National Chiao Tung University
Hsinchu, Taiwan, ROC

ABSTRACT *- A high-quality and low-complexity algorithm for pitch modification of acoustic signals is proposed. It is high quality, because time-domain waveform shape and phase synchrony of a synthesized signal closely resemble that of its original signal. Low complexity is attained by performing the algorithm entirely in time domain with fast algorithm, and without resorting to complicated frequency domain analysis and synthesis. The time domain synthesis mostly consists of fast algorithms of finding minimum absolute error (MAE) and a cross fading operation for high correlation gain and phase synchrony. The MAE-based algorithm is shown to yield smaller complexity, and better performance, than other well-known correlation cost functions. Pitch modification is performed by selecting and synthesizing apposite frames from original input signals in a way such that the original waveform shape and phase synchrony are preserved. Subjective tests showed a comparable performance to that of the best known algorithms, but at much reduced complexity.*

1. INTRODUCTION

High quality pitch modification of acoustic signals is popular for applications such as voice mail, multimedia audio signal processing, sampling synthesizer, special sound effects for karaoke and vocoder, just to name a few. The conventional algorithms are based on signal resampling and formant analysis/synthesis. These algorithms either need very huge amount of computation or cause severe distortion.

Lee [1] compressed or expanded the resampling signals to keep the length of the pitch-shifted signal constant, but it often introduces pops, clicks, and other artifacts. Besides, its change of formant envelope create noise-like chipmunk. The analysis/synthesis method involving linear prediction filtering [2] and Fourier transforms can retain formant shape but the amount of computation is also quite huge. Similar arguments can be applied to Quatieri and McAulay's approach [3].

Roucos and Wilgus [4] developed an iterative algorithm based on a cost function of minimum square error between Fourier transforms of the original and its synthesized signals. It is also computation intensive. Verhelst and Roelands [5] modified the overlap-add (OLA) and synchronized overlap-add (SOLA) algorithms, and introduced a waveform-similarity overlap-add algorithm (WSOLA). This algorithm was intended for time-scale modification of an input signal. However, it can be modified for pitch-modification purpose. The algorithm applies Hanning window and 50% overlap to comparing the similarity between two windowed speech frames. It could get good result for speech signal, because speech waveforms are rather regular. For acoustic signals, however, the window is too wide and too computation-intensive to yield a good performance. Besides, it is impractical for real-time purpose. Zhang [6] applied the concept of resonator and filter bank to modify pitch. It causes very large noise for both low frequency and high frequency signals. It also needs massive computation. Lent [7] just simply added or discarded signal segments to achieve pitch shifting. The method is simple but yielded poor results.

An ideal pitch-shifted signal should possess similar time-domain waveform as well as frequency spectrum shapes to those of its source signal. The existing frequency-domain approaches produce better results, but require huge computation overheads. Also, the synthesized signals normally have a much different waveform from the original one. On the other hand, the existing time-domain approaches are simpler, but yield poorer performances, both in waveform and frequency spectrum similarities. To achieve both time-domain and frequency-domain similarities by using the time-domain approach, we select apposite frames from original input signals and synthesize them in a way such that the original waveform shape and phase synchrony are preserved.

For further computation reduction, an MAE cost function [5] is introduced other than the popular correlation cost functions such as the minimum square error, minimum angle deviation, maximum correlation and maximum correlation coefficient functions, for the best correlation measure between two data frames to be synthesized. Surprisingly, the MAE measure produced the best averaged results among all cost functions. As a further step, we introduce a block binary search

 0-7803-2431-5/95 $4.00

algorithm and a subsample algorithm for finding the minimum MAE. In short, the combinations of time-domain phase-synchronous approach, simple MAE operations, the fast binary search and subsampled computation lead to a high-performance and low cost pitch-modification algorithm.

2. THE ALGORITHM

For real-time and performance consideration, a signal to be pitch modified is first divided into frames. Then the frames are synthesized as shown in Figure 1 and Figure 2, where the dilation and compression factors of the modified frames are equal to pitch shifting factor. Figure 1 shows pitch-up operation while Figure 2 depicts pitch-down operation, where A) is the original signal. It is divided into frames which may overlap like (1), (2), and (3). For pitch-up operation, after the frames are squeezed by a pitch modification factor, the lengths of frames become shorter like (1'), (2'), and (3'), where these frames' signal contents don't change but their playing time become shorter. Then we get the next frame starting at the position where the end of the last modified frame maps upright to the original signal. Cross fading operation is applied to smooth the overlapped frame junction as follows:

$$CF(m) = x_1(m)\cdot\left(1-\frac{m}{cs}\right)+x_2(m)\cdot\frac{m}{cs} \qquad 0\le m\le cs-1$$

where cs is the size of cross fading region. Similar operations are applied to pitch-down operation, except that frames are dilated. There always exist differences between two adjoined synthesized frames such as period, amplitude, phase and so on, where a cross-fading operation can not eliminate entirely. As such, pops, clicks, and other noises can be perceived.

One way of getting around this problem is finding a best correlation position between two adjoining frames in an allowable cross-fading range as shown in Figure 3, where the next frame can be appended right to a position with maximum correlation. Doing this way ensures a phase synchrony between synthesized frames. Therefore we can attain high fidelity to both original signal's waveform and phase continuity, and correspondingly a high quality pitch-modified signal.

There are many cost functions can be applied to the search of maximum correlation, such as the conventional correlation function, correlation coefficient, normalized correlation, mean square error and mean absolute error (MAE) functions. Among them, MAE is the simplest one as shown below.

$$MAE(\tau) = \frac{1}{cs}\sum_{m=0}^{cs-1}\left|x_1(m)-x_2(m+\tau)\right|$$

And interestingly, MAE is also shown to yield the best simulation results in average, among all the mentioned cost functions.

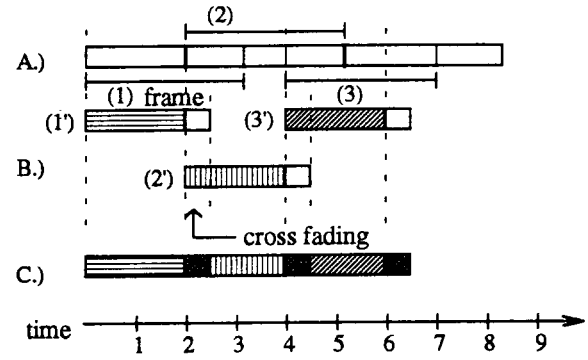

Figure 1. Time-domain synthesis for pitch-up operation

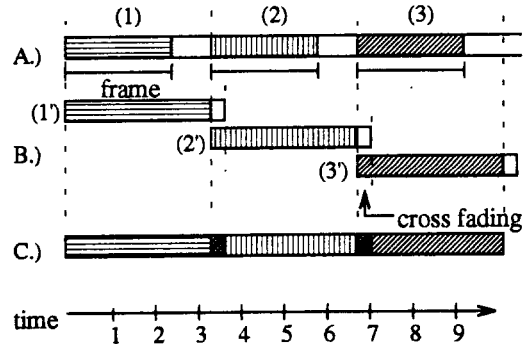

Figure 2. Time-domain synthesis for pitch-down operation.

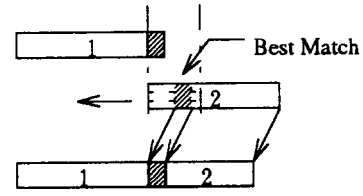

Figure 3. The region for cross fading and maximum frame correlation and phase synchronization.

For the success of a pitch-modification algorithm, frame size is a crucial factor. If a frame is too short, low frequency components will be distorted. On the other hand, a long frame normally introduces discontinuous short-time echoes especially when raising the pitch to a very high level. According to

experiences, for more pitch shift, a shorter frame is desirable.

After synthesis, duration of the synthesized signal are not the same as the input signal. The playing time of the synthesized signal can be made equal to that of the input signal by applying the the synthesized signal to a DAC with adjustable sampling rate. However, this approach is expensive. A practical approach is to interpolate or decimate the synthesized signal such that the final signal has the same amount of data as that of the input signal. Because audio waveforms are continuous, the difference between two samples is minor. So either zero-order or first-order interpolation/decimation schemes can be applied to the data resampling operation. Take the zero-order interpolation/decimation algorithm for instance, a data to be resampled can be obtained by simply averaging two known samples if it falls in between these two known samples. There is no multiplication involved in this resampling process. According to simulation, the distortions introduced are not conspicuous to a people's ears.

3. FAST MAE ALGORITHMS

There are two ways of reducing computation required in the search of cross-fading location which results in MAE. The first approach is that we can do MAE search with subsampled input signals, for instance, a two times subsampling as shown below.

$$MAE(t)=\frac{2}{CS}\sum_{m=0}^{CS-1}|x_1(m)-x_2(m+t)| \qquad \text{for } m=m+2$$

In this case, computation complexity is reduced to half that of the original. Subsampling factor can be increased to a tolerable maximum value based on simulation, which results in maximum complexity reduction.

The second approach is that we can include a fast MAE search algorithm in stead of full search, as shown in Figure 4. First, we choose 3 matching blocks within the search range to evaluate MAE's as shown in Figure 4-(1), which are at the positions of $(1/4)l$, $(2/4)l$, and $(3/4)l$ respectively. If for example, the position of $(3/4)l$ produces a smallest MAE, we narrow in the search region by comparing the MAE with two new MAE's corresponding to the locations of $(3/4+1/8)l$ and $(3/4-1/8)l$ as shown in Figure 4-(2). Then the process is repeated for the succeeding finer search locations, until no further iteration is possible. Finally, We can get a best conceived frame synthesis location with good phase synchronization.

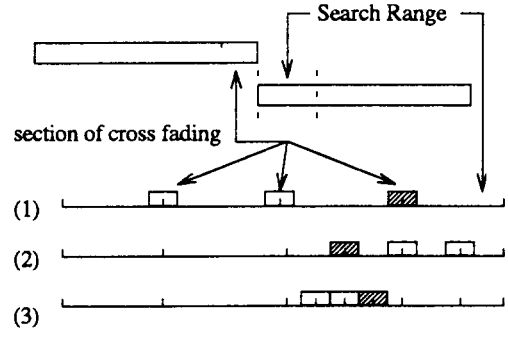

Figure 4. The fast binary search algorithm for minimum MAE.

The binary search may not produce good results if the search range is too wide. To remedy the situation, we can divide a search range into several sections. We call the improved method as Block Binary Search (BBS) algorithm. Within each section, binary search is applied to find its MAE. Then after comparing all the sections' best MAE's, the position corresponding to the global MAE is the best phase synchronization point for frame synthesis and cross fading. If the search range is divided into n blocks, the computation complexity is

$$n\cdot\left[3+2\cdot\left(\log_2 MS/n-2\right)\right]$$

where MS is search range. For example, if n = 4 and MS = 10ms * 22.05kHz, the amounts of computation are about 42 addition and absolution operations, which is about 20% that of originally required. We can reduce the complexity to 10% by combining BBS with subsampled algorithm. Simulation confirmed that this hybrid algorithm also produced good results.

Based on the fast search algorithm, the proposed pitch-modification algorithm is summarized by the flow chart as shown in Figure 5. In the chart, the search range is divided into 4 sections.

4. CONCLUSION

In summary, the combinations of time-domain in-phase synthesis approach, and the fast BBS subsampled MAE algorithm lead to a high-performance and low-complexity algorithm for pitch modification. Subjective simulation results showed that the proposed algorithms have better performances than the comparable known algorithms. As can be easily seen, this algorithm may also be applied to time-scale modification of an acoustic signal. This is trivially done when a synthesized (shortened or compressed) signal (according to the specified time-rate modification factor) is played with the same speed as the input signal.

REFERENCES

[1] F. Lee, "Time Compression and Extraction of Speech by the Sampling Method," Journal of the Audio Engineering Society, 20(9): 738-742.

[2] M. Dolson, "Phase Vocoder: A Tutorial," Computer Music Journal 10(4):14, 1986.

[3] Thomas F. Quatieri & Robert J. McAulay, "Shape Invariant Time-Scale and Pitch Modification of Speech," IEEE Trans. on Signal Processing. Vol. 40. No. 3. March 1992, pp. 497-510.

[4] S. Roucos & A. M. Wilgus, "High Quality Time-Scale Modification for Speech," Proc. ICASSP'85, pp. 493-496.

[5] Werner Verhelst & Marc Roelands, "An Overlap-Add Technique Based on Wave-form Similarity (WSOLA) for High Quality Time-Scale Modification of Speech," Proc. of IEEE ICASSP'93, pp.554-557.

[6] G. Zhang & W. F. McGee, "Windowing Techniques in the Design of Resonator Based Frequency Interpolation Filter Banks," Proc. of IEEE ICASSP'91, pp.1821-1824.

[7] K. Lent, "An Efficient Method for Pitch Shifting Digitally Sampled Sounds," Computer Music Journal, Vol. 13, No. 4, Winter 1989, pp. 65-71.

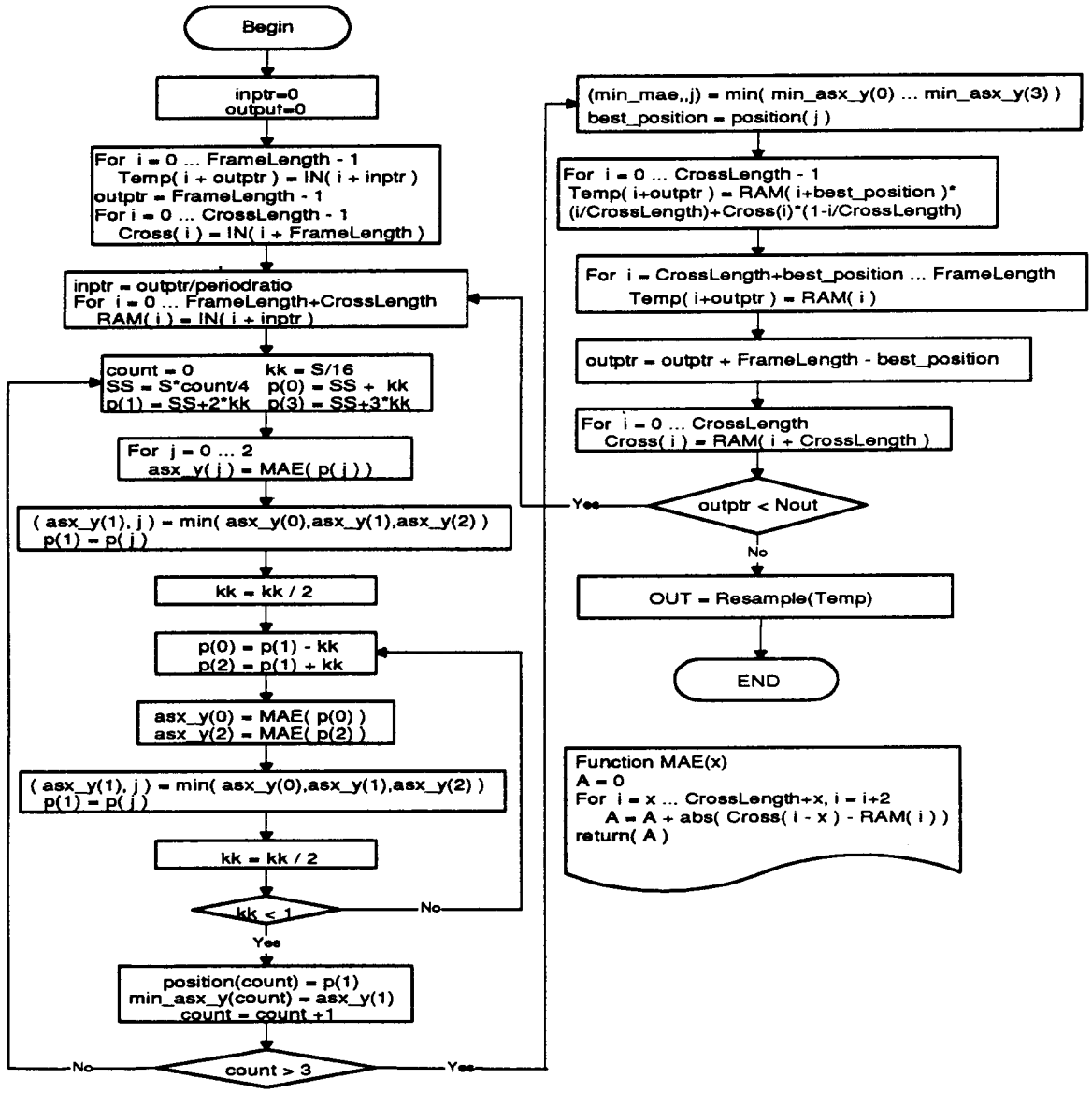

Figure 5. Flow chart of the proposed algorithm.