ON IMPROVING THE INTELLIGIBILITY OF SYNCHRONIZED OVER-LAP-AND-ADD (SOLA) AT LOW TSM FACTOR

Peter H. W. Wong*, Oscar C. Au**, Justy W.C. Wong***, William H. B. Lau

Department of Electrical and Electronic Engineering
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong.
Email: ee peter@ee.ust.hk*, ecau@ee.ust.hk**, ejusty@ee.ust.hk***, eewillau@ee.ust.hk
Tel.: +852 2358-7053**

ABSTRACT
In this paper, we propose an algorithm to modify the Synchronized Overlap-and-Add (SOLA) technique. SOLA is a popular technique for modifying speech and audio signal. It changes the time scale of signal while maintaining the pitch information. However, when the time scale of the signal is compressed more, the synthesized speech of SOLA has high articulations rate that it becomes almost impossible for human ear to comprehend. The proposed algorithm can enhance the quality if the time scaled signal by using time varying time scale factor which is locally adapted to the input signal. It is found that the intelligibility of the synthesized signal using the proposed algorithm is considerably better than the one synthesized from simple SOLA with.

1. INTRODUCTION
Time scale modification (TSM) [1-8] is a class of algorithms to change the time scale of a signal. By changing the apparent rate of articulation, TSM can be useful to make degraded speech more intelligible. It can also be useful in variable speed playback of audio clips. There is an associated parameter, $\alpha$, called TSM factor. When $\alpha$ is one, the signal is unchanged. When $\alpha$ is greater than one, the signal is time expanded (e.g. from 1 second to 2 second if $\alpha=2$). When $\alpha$ is less than one, the signal is time compressed (e.g. from 1 second to 0.5 second if $\alpha=0.5$). Some TSM algorithms are time domain techniques such as overlap-and-add (OLA) and synchronized OLA (SOLA)[8]. Some are frequency domain techniques such as Least Square Error Estimation from Modified Short Time Fourier Transform Magnitude (LSEE-MSTFTM) [6]. Some are based on sinusoidal model method [7], which adjusts the system amplitudes, phases, excitation amplitude and frequencies.

This paper is concerned about the popular synchronized overlap-and-add (SOLA) which is relatively simple to implement and has good audio quality. The SOLA is based on OLA which simply overlaps and adds adjacent frames. The analysis (input) frame of length $S_a$ and the synthesis (output) frame of length $S_s$ are related by $S_s = \alpha S_a$ where $\alpha$ is the TSM factor. With the simple overlap-and-add operation, OLA may cause undesirable reverberation and clicks. SOLA solves this problem by overlapping only at the points with highest similarity between the two overlapping frames. The synthesized speech quality is considerably improved in SOLA. However, when $\alpha$ is very small (e.g. 0.25 or 0.3), the synthesized speech of SOLA has such a high articulation rate that it becomes almost impossible for human ear to comprehend. In other words, the synthesized speech becomes increasingly unintelligible when decreases beyond 0.3, even though the speech quality remains high. In this paper, we propose to modify SOLA such that the synthesized speech can be more intelligible at very low $\alpha$.

2. REVIEW OF SYNCHRONIZED OVER-LAP-AND-ADD (SOLA)
The input (or analysis) signal $x[n]$ is segmented into overlapping frames of length $N$ that are a distance of $S_a$ apart. The first frame is directly copied to the output (or synthesis) signal $y[n]$. The $(m+1)$-th frame which starts at $m \times S_a$ slides along the synthesized signal $y[n]$ around the location $m \times S_s$ in the range of $[k_{min}, k_{max}]$ to find a location which maximize the normalized cross-correlation function defined in (1) for the overlapping region.

$$R[k] = \frac{\sum_{i=0}^{K-1} y[m \times S_s + k + i] \cdot x[m \times S_a + i]}{\left(\sum_{i=0}^{K-1} y^2[m \times S_s + k + i] \cdot \sum_{i=0}^{K-1} x^2[m \times S_a + i]\right)^{1/2}}$$  (1)

The $S_a$ and $S_s$ are called the analysis and synthesis frame period respectively. The relation between $S_a$ and $S_s$ is defined in (2)

$$S_s = S_a \times \alpha$$  (2)
\( \alpha \) is called the time scale factor. The signal is time scale expanded when \( \alpha \) is greater than one and time scale compressed when \( \alpha \) is smaller than one. \( L \) is the length of the overlapping region between the shifted analysis frame and synthesized signal. Usually the \( k_{\text{min}} \) and \( k_{\text{max}} \) are set to \(-N/2\) and \(N/2\) respectively. Once the location which maximizes the cross-correlated is determined, the overlapping region is cross-faded and the remaining of the analysis frame is directly copied.

3. MODIFIED SOLA FOR SMALL TSM FACTOR

When \( \alpha \) decreases, the degree of overlapping between adjacent analysis frames increases. This is fine for vowels because vowels have essentially periodic waveforms in the time domain as shown in figure 1 and adjacent analysis frames will be highly correlated at certain shift-point as shown in figure 2. The optimal overlapping point in SOLA guarantees that the adjacent frames are synchronized and thus yield minimal distortion. However, for consonants with transient rather than periodic characteristics as shown in figure 3, the overlap-and-add operation causes a lot of problems. In the transient portion of a signal, adjacent analysis frames can be very different with very low cross-correlation as shown in figure 4. The optimal lag would often be the original position of the analysis frame. But staying at the original position implies a TSM factor of 1 rather than the desired. Thus the optimal lag is often outside the allowable search window, especially when it is small. The forced overlap-and-add operation in the transient region will result in significant distortion. Although the fade-in and fade-out windowing can prevent an abrupt change, the blurring of temporal and frequency characteristics is inevitably introduced. When the degree of overlapping is small, the blurring distortion may not be obvious. But when becomes very small and the overlapping area becomes very large, this blurring distortion will be the dominating artifacts.

As a result, we propose to apply more overlapping to the vowels and less overlapping to the transient sound. In other words, we propose to use a time varying TSM factor \( \alpha(t) \), rather than a fixed constant TSM factor, \( \alpha \), as in SOLA. The time varying TSM factor should be

---

Figure 1: Vowel waveform

![Vowel waveform](image1)

Figure 2: Cross-correlation (vowel)

![Cross-correlation (vowel)](image2)

Figure 3: Transient waveform

![Transient waveform](image3)

Figure 4: Cross-correlation (transient signal)

![Cross-correlation (transient signal)](image4)
small when adjacent analysis frames are very similar, and high when they are not so similar. It should be such that the average TSM factor is close to the target overall TSM factor. In addition, we propose to remove the silent frames characterized by very little frame energy. This will free up the time of the corresponding synthesized frames for the transient frames so that they can use a slightly higher TSM factor. From figure 2 and figure 4, it can be seen that the vowel frames can have cross-correlation higher than 0.9, while the transient sound seldom have cross-correlation higher than 0.9 except possibly near the original position. We thus propose to use the cross-correlation as a check. Let the target overall TSM factor be. At first, all frames are tested for silent frames which are discarded. All non-silent frames are assumed to be vowel-like frames and are to use a smaller-than-target TSM factor. A search window is defined and the cross-correlation for each search point is computed. If the cross-correlation ever exceeds 0.9 within the search range, the frame is confirmed to be vowel-like and the first peak above 0.9 will be considered the optimal overlapping position. This first peak may not be the global maximum, but is good enough. Such a peak would result in smaller time scale distortion. If the cross-correlation does not exceed 0.9 throughout the search range, the frame is considered a transient frame and a larger TSM factor is used. The search range is extended to cover the range for and further searching is done. The optimal overlapping position is the point with maximum cross-correlation in the extended range.

4. SIMULATION AND RESULTS

The testing signals for the modified SOLA is spoken by a male-speaker called 'Au', which is recorded in room condition with the sampling frequency of 8192Hz. The proposed modified SOLA algorithm is simulated with different TSM factor and the simulation results for $\alpha = 0.333$ are shown. Figure 5 shows the spectrogram of the original speech signal 'Au', figure 6 and 7 show the spectrogram of the time compressed 'Au' with $\alpha = 0.333$ using SOLA and proposed algorithm respectively. The spectrograms are stretched to the same dimension for comparison. It can be observed from figure 7 that the TSM factor is time varying. Figure 8 shows the time scale factor against the frame number of 'Au', which is time varying with the mean of about 0.331. After a lot of subjective tests are carried out, it is found that intelligibility of the synthesized speech from the proposed modified SOLA is considerably better than the one synthesized from simple SOLA with. This suggests that proposed modified SOLA is effective in improving the intelligibility of SOLA when the TSM factor is very small.

5. CONCLUSION

An algorithm is proposed to modify the popular SOLA method for the time scale modification of speech and audio signals. The proposed algorithm improves the intelligibility if the SOLA when the TSM factor is very small by using time varying time scale factor, rather than a constant TSM factor. The silent frames are removed and slightly higher TSM factor can be obtained locally without changing the overall TSM factor. The increasing of the TSM factor make the synthesized signals more intelligibility and more suitable for fast browsing or searching.

REFERENCE

Figure 5: The stretched spectrogram for the time compressed 'Au' with $\alpha = 0.333$ using SOLA

Figure 6: The stretched spectrogram for the time compressed 'Au' with $\alpha = 0.333$ using modified SOLA

Figure 7: Time scale factor for different frames using proposed algorithm