TIME-SCALE MODIFICATION USING FUZZY EPOCH-SYNCHRONOUS OVERLAP-ADD (FESOLA)

Timothy Roberts

Griffith University
Brisbane, Australia
timothy.roberts@griffithuni.edu.au

Kuldip K. Paliwal

Griffith University
Brisbane, Australia
k.paliwal@griffith.edu.au

ABSTRACT

A modification to the Epoch-Synchronous Overlap-Add (ESOLA) Time-Scale Modification (TSM) algorithm is proposed in this paper. The proposed method, Fuzzy Epoch-Synchronous Overlap-Add, improves on the previous ESOLA method through the use of cross-correlation to align time-smeared epochs before overlap-adding. This reduces distortion and artefacts while the speaker's fundamental frequency is stable, as well as reducing artefacts during pitch modulation. The proposed method is tested against well known TSM algorithms. It is preferred over ESOLA, and gives similar performance to other TSM algorithms for voice signals. It is also shown that this algorithm can work effectively with solo instrument signals containing strong fundamental frequencies. Full implementation of the proposed method and zero frequency resonator can be found at github.com/zygurt/TSM.

Index Terms— ESOLA, Time-Scale Modification, Time-Domain, FESOLA, Zero-Frequency Resonator, Epoch

1. INTRODUCTION

Time-Scale Modification (TSM) is the process of manipulating the temporal domain of a signal without changing the spectrum of the signal. It is usually achieved by adjusting the ratio between the analysis shift size and the synthesis shift size in an analysis, modification, synthesis framework. To remove discontinuities and retain phase coherence at the adjusted time-scale, a number of methods have been proposed. Recently, these methods include Harmonic-Percussive Time Scale Modification [1] in 2014 and Fuzzy Bin Classification Phase Vocoder [2], Mel-Scale Filterbanks [3] and Epoch-Synchronous Overlap-Add (ESOLA) [4] in 2018. ESOLA aims to improve the quality of time-scaled speech by extracting glottal pulses, also known as epochs, and using these markers to align the overlap process. By aligning these epoch markers, the primary structure of the signal is preserved in a computationally efficient manner. Additionally, as the epoch locations in the source file are constant, the positions of the epochs need only be generated once, and can then be used for future time-scaling.

2. BACKGROUND

The production of voiced speech is a well understood process with air provide by the lungs passing through vocal chords, which close against each other in an oscillatory motion at the glottis [5]. Significant excitation of the vocal system is generated at the moment of vocal chord closure [4]. These moments are known as epochs or glottal closure instants. Figure 1 shows the location of these epochs within male speech calculated using the Zero Frequency Resonator (ZFR) algorithm. Note that the periodicity of the epochs matches the fundamental period of the signal.

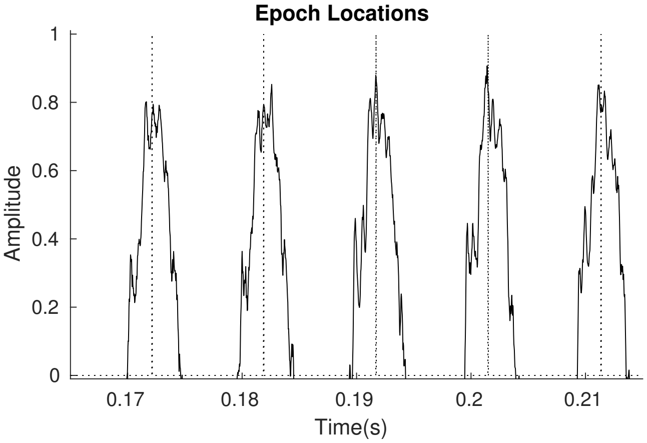

Figure 1: Epochs within male speech calculated using the Zero Frequency Resonator method.

ESOLA is motivated by the relatively small changes in fundamental frequency across a range of speaking rates [4]. As the fundamental frequency depends on the glottal closure instants, epochs make a logical candidate for re-aligning segments of the source file at a new time-scale. Multiple methods of producing epochs were considered [6], with the ZFR method used as it gives reliable estimates with a lower computational complexity. Modification of the source signal begins with pitch-blind windowing, with a 50% overlap. The synthesis overlap between frames is then increased or decreased for speed increase or decrease respectively. During the overlap-add process, the input analysis frame is compared to the previous output synthesis frame. The lag from the first epoch in the output synthesis frame to the next epoch in the input analysis frame is calculated. A new input analysis frame is then extracted at the lag offset before overlap-adding. This method is efficient, however it does not take changes in fundamental frequency into account and is prone to mis-alignment of epochs, shown in figure 2.

The proposed method was developed as an additional method for use in large scale subjective testing as part of a separate project. As such, the aim was to create a working implementation with artefacts distinct from previous methods. ESOLA was implemented initially, however results similar to those available online could not be achieved. Occasional loss of pitch information, and distortion dur-

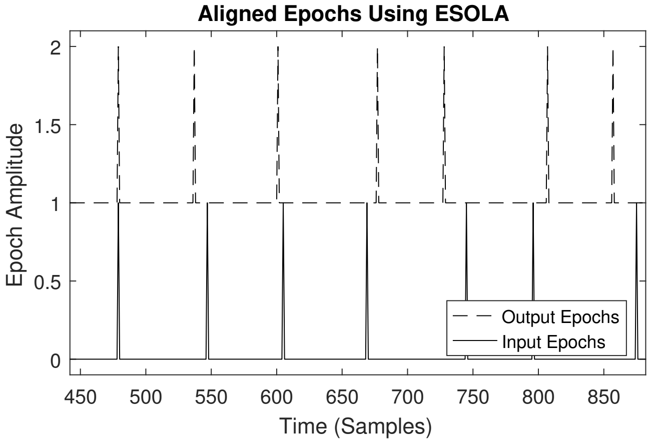

Figure 2: Epoch Synchronisation of the ESOLA method.

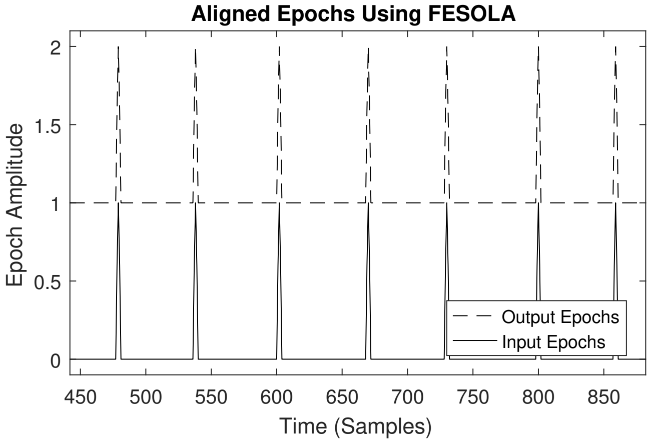

Figure 3: Epoch Synchronisation of the proposed FESOLA method.

ing changes in fundamental frequency were identified as primary problems to be addressed. As such, the changes described below were made to improve the quality of the algorithm.

3. METHOD

The proposed method, Fuzzy Epoch-Synchronous Overlap Add (FESOLA), extends the original ESOLA method through the use of cross-correlation when calculating the overlap offset. To enable the cross-correlation to work more effectively, the samples before and after each epoch are set to a magnitude of 0.6.

The ZFR method, as proposed by [4], is used for epoch extraction. The signal is first pre-processed by calculating the first difference (1), where $s[n]$ is the speech signal, to remove any low frequency bias present in the signal.

$$x[n] = s[n] - s[n-1] \tag{1}$$

$x[n]$ is then passed through two ideal Zero Frequency Resonators, (2) and (3).

$$y_1[n] = -\sum_{k=1}^{2} a_k y_1[n-k] + x[n] \tag{2}$$

$$y_2[n] = -\sum_{k=1}^{2} a_k y_2[n-k] + y_1[n] \tag{3}$$

The resulting trend in the filtered signal, $y_2[n]$ is removed through successive mean-subtraction operations, (4), where $2N+1$ is chosen to be 1 to 2 times the fundamental pitch period. An estimate of the fundamental pitch period is found through averaging magnitude spectrum frames across the entire signal and finding the maximum bin location. This bin location is then used to calculate an estimate of the fundamental pitch period. This allows the epoch extraction to be adaptive and suit both male and female voices as well as other non-voice signals. At a sampling frequency of 44.1 kHz, $N_{ZFR} = 217$ is appropriate for male and female speech.

$$y[n] = y_2[n] - \frac{1}{2N_{ZFR}+1} \sum_{m=-N_{ZFR}}^{N_{ZFR}} y_2[n+m] \tag{4}$$

Finally, zero crossings in $y[n]$ indicate epochs within the signal and are labeled with a magnitude of 1.

Once the epochs have been calculated they are spread in the time domain, made fuzzy, by setting the samples immediately before and after each epoch to a magnitude of 0.6, the value of which was determined experimentally. Manipulating additional samples around each epoch was explored, however no improvement in optimal frame positioning was found.

To achieve time-scaling, two frames of epochs are extracted according to [4]. Cross-correlation between these frames is calculated using (5), where L is the length of overlap and S_s is the synthesis shift size. The location of the maximum value within the cross-correlation determines the lead or lag to the start of the adjusted input frame. In the case of multiple maximums in the resultant cross-correlation, the lead or lag closest to the center of the cross-correlation is used.

$$R_{xy} = \sum_{m=0}^{L-1} x[m]y[m-k] \quad ; \quad \frac{-3S_s}{4} \leq k \leq \frac{3S_s}{4} \tag{5}$$

The adjusted next frame is extracted based on the required lead or lag, according to [4] and windowed using a Hann window before overlap adding to the output signal. Epochs of the adjusted frame are combined with an output epoch signal using overlap-adding for use in aligning the following frame. An output window signal is also created to allow for normalisation once processing is complete.

The improvement in epoch alignment for the proposed approach can be seen in figure 3, in comparison to ESOLA in figure 2. The examples are taken from the first input frame containing epochs for both methods using the same signal and same parameters.

The changes in algorithm also inherently account for changes in pitch of the speaker, removing distortion for small changes, and reducing distortion for large fast changes.

4. TESTING

Small-scale subjective preference testing and large-scale subjective quality testing was undertaken. Small-scale subjective testing compared ESOLA and FESOLA at two different time-scales (playback speeds of 53.58% and 78.21%) for 10 source files containing speech

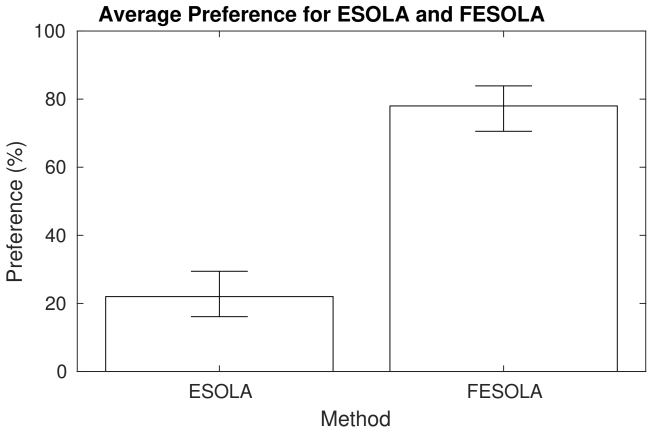

Figure 4: Mean preference comparison for ESOLA and FESOLA.

from 5 female and 5 male speakers. Testing was undertaken using the Web Audio Evaluation Toolkit (WAET) [7] in an AB format. 10 participants were involved with the testing, all with backgrounds in signal processing.

Large-scale subjective testing was undertaken as part of a larger project. Participants were presented with the source signal and modified versions processed using the proposed method, in addition to the Phase Vocoder [8], Identity Phase-Locking Phase Vocoder [9], Waveform Similarity Overlap-Add [10], Harmonic-Percussive Time-Scale Modification [1] and Mel-Scale Sub-band Time-Scale Modification [3] algorithms, and asked to rate the quality of the processing. 88 source files were scaled at 10 ratios, 38.38%, 44.27%, 53.83%, 65.24%, 78.21%, 82.58%, 99.61%, 138.1%, 166.7%, and 192.4%, resulting in 5280 files. Testing used the WAET with 6 pairs of files presented per page using horizontal sliders. The number of files in each session was refined during testing, and settled at 60 files per session for a testing time of between 10 and 20 minutes. Approximately 60% of participants were expert listeners, with an average age of 34 and standard deviation of 11 years.

5. RESULTS

The small-scale preference testing showed a clear preference (77%) towards the proposed algorithm, shown in figure 4. Signals in which the speaker greatly varies their pitch show a stronger preference for the proposed method, while source files with less variation show a more even preference between methods. This improvement is consistent with the changes made to the ESOLA algorithm.

The large-scale subjective testing results presented in this section are a selection of findings from a larger study, containing approximately 19000 signal ratings. The proposed method performs comparatively well for voice signals, figure 5 and solo instrument signals, figure 6. However it gives poor time-scaling for complex musical source material, shown in figure 7. This is due to the reliance on a strong fundamental frequency to allow for generation and successful alignment of epochs. The epoch alignment shown in figures 2 and 3 are from a flute recording, showing that this method is useful is situations beyond time-scaling of speech. Modifications to the length of analysis frames must be made if low frequency content is to be time-scaled however, to ensure that at least one pitch period or epoch falls within half a frame. The use of a 23ms frame window, in this implementation, results in a half frame pitch period of 11.61ms or 86 Hz.

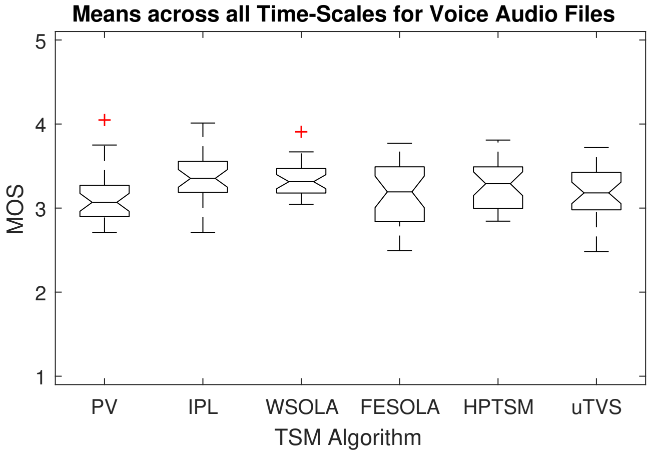

Figure 5: Comparison of mean opinion scores averaged across all voice processed files. MOS of 1-5 is Bad-Excellent.

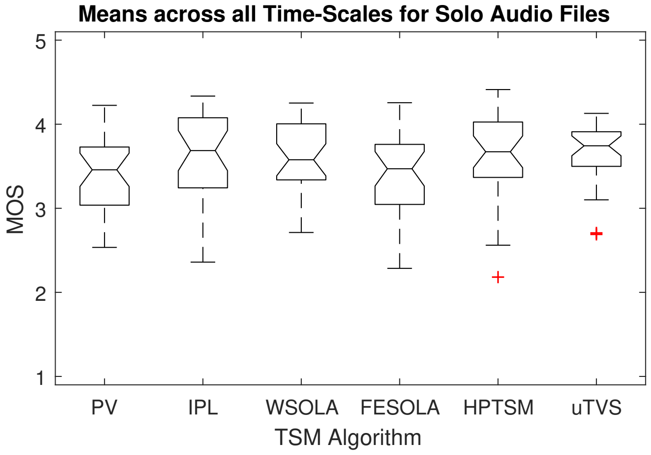

Figure 6: Comparison of mean opinion scores averaged across all solo processed files. MOS of 1-5 is Bad-Excellent.

Of interest is the relatively poor performance of all methods tested when comparing time-scaling of voice to time-scaling of musical material. This could be due to how often cadence changes within normal conversation. As talking faster or slower is part of general speech, perception may be more finely tuned for this type of modification. This may then result in a stronger reaction when artefacts corrupt speech signals.

When considering a large variety of source material, including music, solo instruments, sound effects and voice, the quality of the proposed method drops sharply as the time-scale ratio moves away from 100%, shown in figure 8. The falloff is accentuated by poor performance with harmonically complex signals, and while not limited to the proposed method, the drop in quality is more severe than the other methods tested. However, the proposed method has similar levels of falloff for voice files when compared to alternative methods. Figure 9 shows the mean MOS values for tested methods when processing voice signals across a range of time scales.

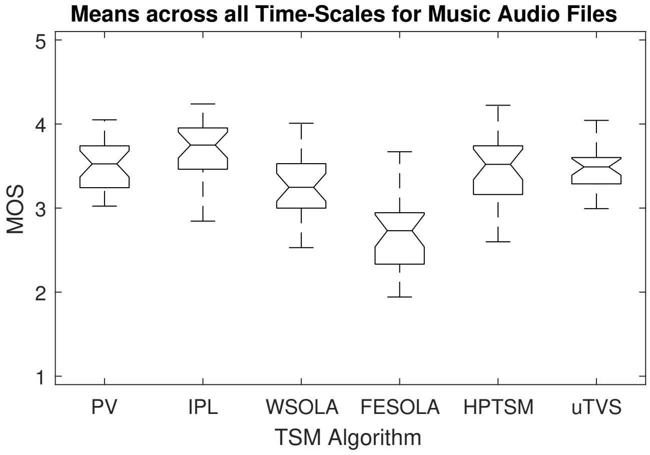

Figure 7: Comparison of mean opinion scores averaged across all music processed files. MOS of 1-5 is Bad-Excellent.

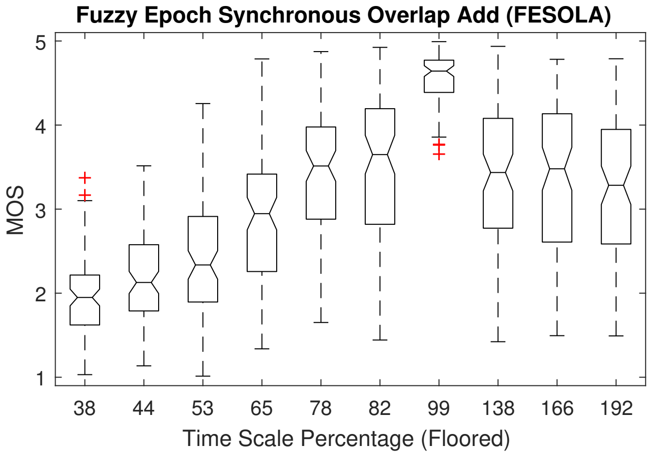

Figure 8: Box plots of Mean Opinion Scores for FESOLA across all source files. MOS of 1-5 is Bad-Excellent.

6. CONCLUSION

In this paper a modified TSM algorithm has been proposed. It extends the previous ESOLA method through the use of cross-correlation to align epochs when overlapping frames, and subsequently reduces distortion and artefacts. This change also reduces artefacts due to the speaker modifying their pitch. The proposed method has been tested against well known TSM algorithms and is found to be preferred over ESOLA, and give similar performance to other TSM algorithms. It was also shown that this algorithm can work effectively with solo instrument signals with strong fundamental frequencies. However, TSM of speech remains an open-problem particularly at slower time-scales.

7. REFERENCES

[1] J. Driedger, M. Muller, and S. Ewert, "Improving time-scale modification of music signals using harmonic-percussive separation," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 105–109, 2014.

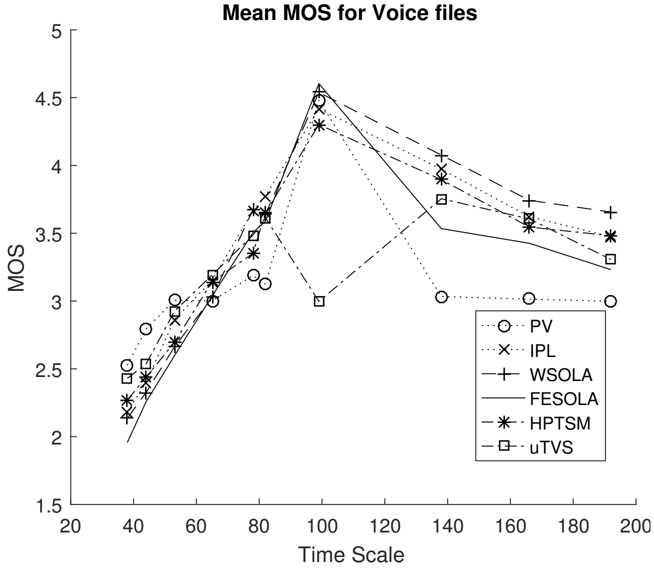

Figure 9: Mean MOS for all Voice signals. MOS of 1-5 is Bad-Excellent.

[2] E. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Applied Sciences*, vol. 7, no. 12, p. 1293, 2017.

[3] N. Sharma, S. Potadar, S. R. Chetupalli, and T. Sreenivas, "Mel-scale sub-band modelling for perceptually improved time-scale modification of speech and audio signals," in *Communications (NCC), 2017 Twenty-third National Conference on*. IEEE, 2017, pp. 1–5.

[4] S. Rudresh, A. Vasisht, K. Vijayan, and C. S. Seelamantula, "Epoch-synchronous overlap-add (esola) for time-and pitch-scale modification of speech signals," *arXiv preprint arXiv:1801.06492*, 2018, unpublished.

[5] X. Huang, A. Acero, H. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice Hall PTR Upper Saddle River, 2001, vol. 95.

[6] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 8, pp. 1602–1613, 2008.

[7] N. Jillings, D. Moffat, B. De Man, and J. D. Reiss, "Web Audio Evaluation Tool: A browser-based listening test environment," in *12th Sound and Music Computing Conference*, July 2015.

[8] M. Portnoff, "Implementation of the digital phase vocoder using the fast fourier transform," *IEEE Transactions on Acoustics, Speech, And Signal Processing*, vol. 24, no. 3, pp. 243–248, 1976.

[9] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[10] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," *Proceedings of ICASSP '93*, vol. 2, pp. 554–557, 1993.