Stereo Time-Scale Modification Using Sum and Difference Transformation

Timothy Roberts
Signal Processing Laboratory
Griffith University
Brisbane, Australia
timothy.roberts@griffithuni.edu.au

Kuldip K. Paliwal
Signal Processing Laboratory
Griffith University
Brisbane, Australia
k.paliwal@griffith.edu.au

***Abstract*—The phase relationship between channels should be maintained when processing multiple channel signals with Time-Scale Modification (TSM). This paper proposes a method and additional variant for maintaining the phase relationship between channels, and retaining the presence in the centre of the stereo signal as a result. The method involves pre- and post-processing the file with the variant processing each frame for real-time suitability. Sum and difference transforms of the stereo signal are used for time-scale modification and results in a large improvement in stereo phase coherence as well as maintaining the stereo field. The proposed method produces a high quality stereo output and greatly improves quality over the independent channel processing method. It also allows for simple implementation, and can be implemented around existing TSM frameworks. The proposed method and variant are suitable for both frequency and time domain TSM methods.**

Availability: All source code, figures, and source audio files can be found at github.com/zygurt/TSM/.

***Index Terms*—time-scale modification, stereo, sum, difference, phase vocoder, mid-side, coherence, balance, TSM**

I. Introduction

Time Scale Modification (TSM) is a well-researched area, with the main processing methods making use of the frequency domain or time domain [1]–[7]. Most published methods ignore application of TSM in multi-channel environments. Exceptions to this are [3], [8], [9], however the stereo field is considered an after-thought, with no published results on the improvement made through the proposed algorithm. Bonada [8] presents that the phase relationship between each of the signals must be considered when processing multi-channel signals and proposes post-TSM phase adjustment to maintain stereo channel phase relationship before and after processing. This method is effective; however, it suffers at slow timescales and when processing independent channels. Ravelli [3] uses cross correlation at transient onsets to align transients in each channel, increasing channel phase coherence. This stereo method is only applicable to the TSM method described in the paper, and does not consider the phase relationship for non-transient content. Instead it aims to increase the phase coherence rather than maintain the original relationship. Finally, Altoe [9] attempts to improve on [8] by processing the sum of the channels and maintaining the phase relationship between the sum and independent channels. The methods presented by Bonada and Altoe are however constrained to frequency domain methods due to adjusting the phase of each frame during TSM.

Blauert [10] states that partially coherent stereo signals produce larger and less sharply located stereo fields compared to a perfectly coherent signal. If the channels of a multi-channel source are processed independently, the phase relationship between the signals can be lost resulting in a distorted stereo field. Small time delays between channels for time domain methods and changes in phase spectra for frequency domain methods cause the change in phase relationship between channels. This change in phase relationship is perceived as the sound source moving to the outside of the stereo field. This effect only occurs when there are differences between the channels, such as a stereo recording with natural reverberation. Sum and difference is a useful stereo signal representation that finds use in applications from stereo FM radio transmission [11] to the Mid-Side microphone technique [11]. The representation is mono compatible with the sum containing in-phase signal information, while the difference contains the out-of-phase signal information.

The method proposed in this paper utilises the sum and difference representation of a two channel signal to improve the quality of the time-scaled signal. Two methods are presented with the first transforming the entire signal prior to and after TSM, and the second transforming the signal after framing and then prior to overlap-add reconstruction. Objective and subjective testing is undertaken comparing the proposed method with [8], [9], and the naive method of processing each channel independently.

It should be noted that while the sum signal is generally of a greater energy level (4-6 dB), compared to the difference signal, this is of little importance to the proposed method as the signal is transformed to and from the sum and difference representation.

The proposed method is presented in Section II, testing methodology is presented in Section III, results are presented in Section IV, and the conclusion is presented in Section V.

II. Method

The proposed method uses sum and difference signals during TSM, which allows for implementation within or

external to existing TSM methods, illustrated in Figure 1. Pre-processing creates the sum and difference signals, while post-processing converts back to the left and right representation.

During pre-processing the stereo signal is transformed using equation 1, where S_1 is the element-wise sum of left and right samples, and D_1 is the element-wise difference between left and right samples. This process is reversible without loss using equation 2. Post-processing transforms the modified sum and difference signals back to the standard left and right signals, using equation 2. In its simplest form this transform is applied to the entire audio file, but can also be applied to each frame in a real-time application. Due to the orthogonal nature of the sum and difference representation, any change moves the resultant phase relationship towards phase coherence, resulting in a processed signal that does not lose presence in the centre of the stereo field.

$$S_1 = L + R \quad ; \quad D_1 = L - R \tag{1}$$

$$L_1 = \frac{S_1' + D_1'}{2} \quad ; \quad R_1 = \frac{S_1' - D_1'}{2} \tag{2}$$

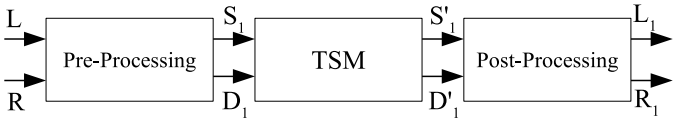

Fig. 1. *Block diagram for the proposed file method. Pre-processing transforms the signal to Sum and Difference while Post-processsing transforms the scaled signal back to Left and Right*

It was found through experimentation that the system is improved in some cases if equation 1 is used for signals biased to the left and equation 5 is used for signals biased to the right. As a result, an additional frame based method was developed. The method first calculates the balance of the stereo field, using equations 3 and 4, to determine the appropriate set of equations for calculating the sum and difference signals.

$$\hat{B}(n) = \frac{|x_L(n)|}{max[|x|]} - \frac{|x_R(n)|}{max[|x|]} \tag{3}$$

$$B = \frac{1}{N}\sum_{n=0}^{N-1} \hat{B}(n) \tag{4}$$

Equations 1 and 2 are used for left biased signals, while equations 5 and 6 are used for right biased signals. This method allows for the signal balance to shift between channels. This modification also preserves the balance when processing signals with silence in either channel. However, due to the use of overlapping frames in TSM this method must be implemented within the TSM algorithm. Alternatively, if the entire signal is framed before processing, the bias may be calculated before time scale processing.

$$S_2 = L + R \quad ; \quad D_2 = R - L \tag{5}$$

$$L_2 = \frac{(S_2' - D_2')}{2} \quad ; \quad R_2 = \frac{(S_2' + D_2')}{2} \tag{6}$$

III. Testing

Objective and subjective methods were used during testing. To facilitate objective testing two features were developed, similar to that found in [3], [12]. The two features developed (Stereo Phase Coherence (SPC) and Balance) give an indication of important features within the stereo field. As the name suggests, SPC is used to measure the phase coherence between each channel and is calculated in the time domain. SPC also gives an indication of the perceived width of the stereo field. Balance is used to measure the mid-point, or pan of the signal. These features were chosen as they are used extensively in music production to give visual feedback about the signal under examination, in addition to their use in [3].

The SPC feature, C, shows the average time domain phase coherence of the frame under investigation and ranges from 1 (completely coherent) to -1 (completely incoherent). It is calculated by framing the signal and performing element wise multiplication between each of the channels, shown in equation 7. Each value is subsequently bounded such that positive values are set to 1 (in-phase) and negative values are set to -1 (out-of-phase), shown in equation 8. The phase coherence for each frame is finally calculated through computing the mean of each frame, shown in equation 9. The frames are subsequently concatenated to form the feature. By using this method, many cross-correlations are removed from the method proposed in [3].

$$\hat{C}(n) = \frac{x_L(n)x_R(n)}{max[|x|]} \tag{7}$$

$$\hat{C}_{SIGN}(n) = \begin{cases} 1 & \hat{C}(n) > 0 \\ -1 & \hat{C}(n) < 0 \\ 0 & \hat{C}(n) = 0 \end{cases} \tag{8}$$

$$C = \frac{1}{N}\sum_{n=0}^{N-1} \hat{C}_{SIGN}(n) \tag{9}$$

The Balance feature, B, shows the midpoint of the stereo field for the frame under investigation and ranges from 1 (Left) to -1 (Right). It is calculated as the difference between the normalised absolute values of the left and right channels, shown in equation 3. The mean of each frame is then calculated, as in equation 4, with the result for each frame concatenated to form the feature. A DC signal in one of the channels results in Balance at its full range.

An examples of these features for synthetic test files can be seen in Figure 2. White SD Fade is 2 independent channels of white noise fading between sum and difference representations, while Sine panning is a sine tone panning from right to left. As can be seen, a signal fading from sum to difference maintains a balance in the centre of the stereo field, while the SPC moves from in-phase to out-of-phase. Similarly, a coherent signal moving from right to left maintains a SPC of 1, while the balance moves from negative to positive.

During testing, each of the features were calculated from both the original and processed signals. The original features

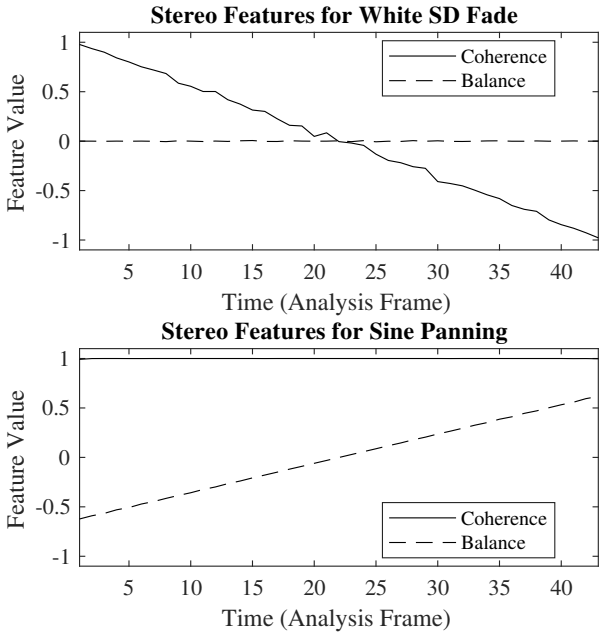

Fig. 2. *Stereo features for white noise crossfading in the sum and difference representation and a sine tone panning right to left.*

are linearly interpolated to match the length of the processed signal features, before L2 norms, equations 10 and 11, were used to calculate the distance between the processed and interpolated signals. $\hat{B}_D$ and $\hat{C}_D$ are the feature distances, m is the frame number, $B_x(m)$ and $B_y(m)$ are Balance features for the original and processed signals, and $C_x(m)$ and $C_y(m)$ are SPC features for the original and processed signals. The resulting distances were then averaged, using equations 12 and 13, for each feature resulting in a dissimilarity for each of the TSM methods under testing. B_D and C_D are the mean feature distances and M is the total number of frames.

$$\hat{B}_D = \sqrt{(B_x(m) - B_y(m))^2} \tag{10}$$

$$\hat{C}_D = \sqrt{(C_x(m) - C_y(m))^2} \tag{11}$$

$$B_D = \frac{1}{M} \sum_{m=0}^{M-1} \hat{B}_D(m) \tag{12}$$

$$C_D = \frac{1}{M} \sum_{m=0}^{M-1} \hat{C}_D(m) \tag{13}$$

Double-blind subjective testing was undertaken with twelve participants testing 4 sets of files processed with Naive, Proposed, Bonada [8] and Altoe [9] Phase Vocoder implementations at a TSM ratio of 82.58%. TSM ratio is $\frac{1}{\alpha}$ and denotes the speed of playback, where α is the oft-used time scaling parameter describing the length of the resulting file. Participants, with backgrounds in signal processing and music technology, were first trained using an additional set of files that portray a change in balance, a comparison of stereo width and the loss of phase coherence. The participants were then played the source material for familiarisation before testing. Each test consisted of the playback of the original file followed by a pair of processed files. Participants were asked to select the file that had the highest similarity to the stereo field of the original file and were asked to pay specific attention to the location of sound sources within the stereo field. One point was given to the chosen method, with points split evenly between methods if the processed files were judged to sound the same. All permutations for each set were presented in random order resulting in 48 tests for each participant, resulting in 576 total tests. Sound reproduction was through Sennheisser HD280 headphones, in a quiet office, with files normalised before playback.

For objective testing, eleven TSM ratios, 38.38%, 44.27%, 53.83%, 65.24%, 78.21%, 82.58%, 96.12%, 125.70%, 146.92%, 169.61% and 184.12%, were applied to the source audio files listed in Table I. Naive, Altoe, Bonada, Proposed File and Proposed Frame Phase Vocoder implementations, along with Naive and Proposed File Waveform Similarity Overlap Add (WSOLA) and Naive and Proposed Harmonic Percussive Time Scale Modification (HP) TSM methods were used, resulting in 396 processed files. Features for each of these files were calculated followed by the mean dissimilarity calculation. Audio files all had a sample rate of 44.1kHz and a bit depth of 16 bits. Features were extracted using a frame size of approximately 50ms. Non-integer TSM ratios were not used to ensure a loss of phase coherency due to phase unwrapping errors [2]. WSOLA [13] and HP [4] TSM methods were implemented around the MATLAB TSM Toolbox [14], with new MATLAB implementations for all other methods. The frame-based sum and difference method was implemented in a traditional vocoder such that both channels were processed simultaneously, giving the ability to produce sum and difference signals for each frame. The proposed method was also implemented within the Extempore programming environment [15].

TABLE I
Source audio files with description

Test File	Comments
Choral	Choir in a small reverberant church.
Electropop	Synthetic polyphonic music with specific panning of sounds, to test reproduction of stereo features and moving sound sources.
Jazz	Big Band with vocalist. Strong central voice and percussion elements, with wind instruments panned through the stereo field.
Saxophone Quartet	Wide stereo field, low reverberation.

To further examine the resulting features, particularly the large variance in dissimilarity, a larger dataset of 88 files was processed, resulting in 8712 processed files. The overall SPC and Balance for processed and original files was graphed. These source files, contain single channel, stereo recordings of a single source and complex stereo fields. This allowed for

the impact of the original stereo field, the TSM ratio and the sound source to be evaluated. A smaller subset of 12 files was used to generate plots for this paper to increase clarity.

IV. Results

Subjective testing, Figure 3, shows a clear preference for Altoe, Bonada and the Proposed method over the naive method when considering the representation of the stereo field. The proposed method was judged to be comparable to Altoe and approaching Bonada in the files tested. Expert listeners were able to more accurately detect changes in the stereo field with results further in favour of the proposed method, with all participants reporting that it was often difficult to discriminate between the test material. Participants also reported that the centre presence of the signal was maintained, however there was a slight narrowing of the stereo field in some test cases for the proposed method.

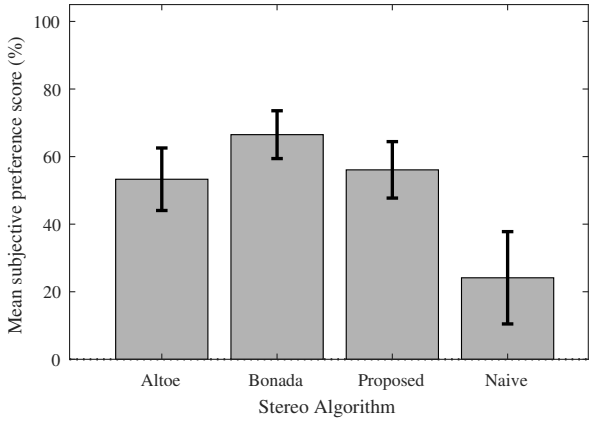

Fig. 3. *Mean Subjective Preference score for Altoe, Bonada, Proposed and Naive Stereo TSM methods. Error bars show ±1 standard deviation from the mean.*

The proposed whole-file and frame methods show a large improvement over the naive approach of single channel implementations for the SPC feature, as per Figure 4. The frame method results in a slightly higher dissimilarity when compared to the whole file method but makes improvements when sound sources move within the stereo field. If the whole file method is used, a sine wave panning from right to left will pan into the centre and back to the right. This problem is rectified using the frame method at the expense of the computation time for calculating the mid-point and a small increase in dissimilarity for the SPC feature. A small increase in dissimilarity for the Balance feature is observed for the proposed methods, Figure 5, however this change is negligible. The small dissimilarity for all methods confirmed informal listening tests that the balance of the signal is minimally impacted by stereo TSM. The general nature of the proposed method can also be seen with reduced dissimilarity for frequency domain and time domain methods.

Smaller feature extraction frame sizes were tested, and resulted in increased, but minimal, dissimilarity for all feature measurements. Synthetic signals were also tested and showed similar, yet less significant improvements.

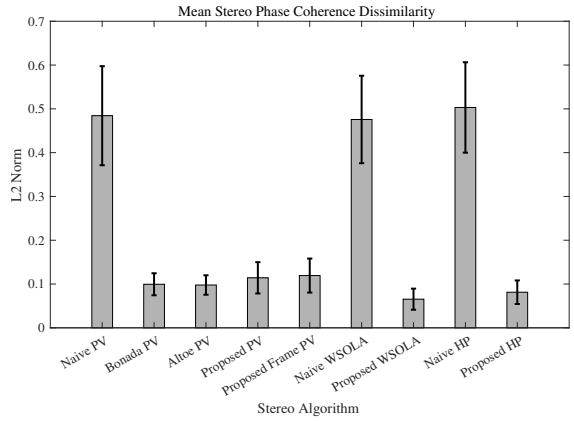

Fig. 4. *Mean Stereo Phase Coherence Dissimilarity for multiple TSM methods and stereo implementation. Error bars show ±1 standard deviation from the mean. Lower is better.*

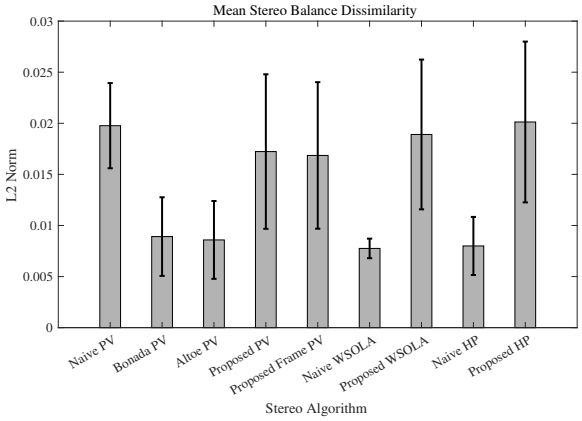

Fig. 5. *Mean Stereo Balance Dissimilarity for multiple TSM methods and stereo implementation. Error bars show ±1 standard deviation from the mean. Lower is better.*

The mean SPC and Balance values for the subset of files at multiple TSM ratios, can be seen in Figures 6 and 7. The SPC of a mono signal is effectively unchanged when doubled across both channels before TSM, Male 12 for example. However, if the recording has been made using multiple microphones, any differences in phase between channels are increased. This is particularly noticeable with the reverberant Brass and Percussion recording. Also of note is the large variability when processing similar sound sources with different recording technique, suggesting that factors other than the source impact on the stereo field. When considering the impact of the TSM ratio, there is a slight downward trend in very coherent source material, however this does not hold true in other cases where there is no discerable trend. A trend which is discernable, for source material with a complex sound field, is an increase in SPC for the Altoe, Bonada and Proposed methods, resulting

in a more coherent phase relationship. This gives reasoning to the slight narrowing of the stereo field.

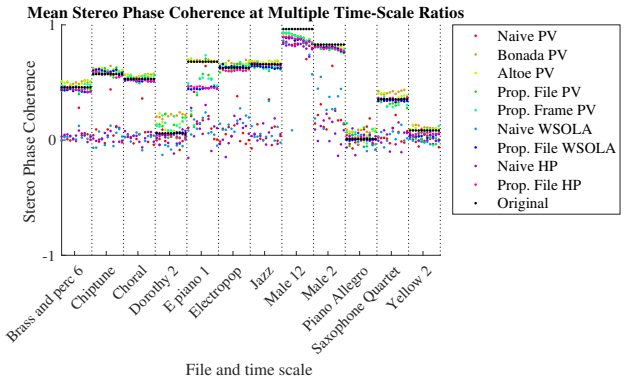

Fig. 6. *[In Colour] Mean Stereo Phase Coherence for multiple TSM methods at multiple TSM ratios.*

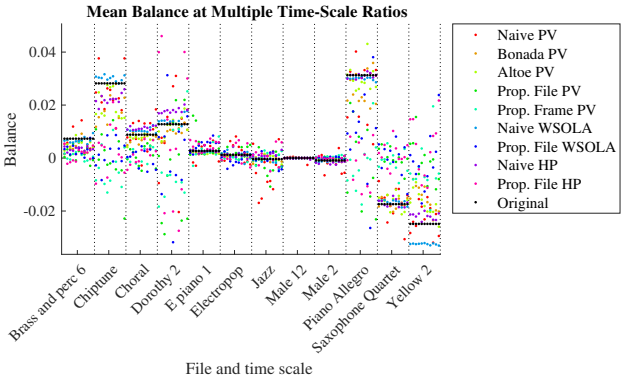

Fig. 7. *[In Colour] Mean Balance for multiple TSM methods at multiple TSM ratios.*

When considering the Balance feature, the change after processing is most pronounced when there are a large variety of sound positions in the stereo field, such as Chiptune, Dorothy 2, Piano Allegro and Yellow 2. These files contain hard panned sounds, many sound effects filling the sound field, a wide recording of a piano and hard panned violin with flute respectively. Reverberant sounds contribute to the next level of change in Balance with mono signals maintaining the balance in the centre of the stereo field. It can also be seen that any change in Balance tends towards the centre for all methods, with the naive WSOLA method tending further away in rare cases.

The proposed methods show improvement in maintaining the channel phase relationship, particularly in complex signals and signals containing reverberation. As a result, the presence in the middle of the stereo field is maintained. Due to not explicitly forcing phase relationships to be maintained, and rather causing any drifting out of phase to result in a drift into phase, a narrowing of the stereo field can be heard in some instances. In complex signals such as the Electropop file, the central instruments with high spectral energy cause the low spectral energy percussion elements to converge to the centre of the stereo image. While this causes a difference in the width of the signal after time-stretching, the end result is more pleasing to the ear, than the loss of channel phase coherence.

V. Conclusion

In this paper two methods of maintaining stereo phase coherence were proposed. These methods used either pre- and post-processing or processing each frame to give real-time suitability. The sum and difference transform of the stereo signal was calculated before processing and then transformed back after TSM processing. This resulted in a large improvement in stereo phase coherence and maintained the stereo field. The proposed methods produced a high quality stereo output and greatly improved quality over the independent channel processing method, and matched previously published methods. It also allowed for simple implementation around previous TSM implementations. The proposed methods were also suitable for frequency and time domain TSM methods.

References

[1] 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.

[2] 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.

[3] E. Ravelli, M. Sandler, and J. P. Bello, "Fast implementation for non-linear time-scaling of stereo signals," in *Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx-05)*, Madrid, Spain, 2005, pp. 182–185.

[4] 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.

[5] N. Sharma, S. Potadar, S. R. Chetupalli, and TV 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.

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

[7] 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.

[8] J. Bonada, "Automatic technique in frequency domain for near-lossless time-scale modification of audio," in *International Computer Music Conference*, 2000.

[9] A. Altoe, "A transient-preserving audio time-stretching algorithm and a real-time realization for a commercial music product," M.S. thesis, Faculty of Engineering, University of Padova, Padua, Italy, 12 2012.

[10] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*, MIT Press, 1997.

[11] M. Ryan and M. Frater, *Communications and Information Systems*, Argos Press P/L, Yarralumla, AUS, 2002.

[12] C. Avendano and J.M. Jot, "Frequency domain techniques for stereo to multichannel upmix," in *AES 22nd International Conference on Virtual, Synthetic and Entertainment Audio*, Espoo, Finland, 2002, pp. 1–10.

[13] 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.

[14] J. Driedger and M. Muller, "Tsm toolbox: Matlab implementations of time-scale modification algorithms," in *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, 2014, pp. 1–8.

[15] A. Sorensen and H. Gardner, "Systems level liveness with extempore," in *Proceedings of the 2017 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, New York, NY, USA, 2017, pp. 214–228, ACM.