JASA ARTICLE

A time-scale modification dataset with subjective quality labels

Timothy Roberts[a] and Kuldip K. Paliwal[b]
Signal Processing Laboratory, Griffith University, 170 Kessels Road, Nathan, Queensland 4111, Australia

ABSTRACT:
Time Scale Modification (TSM) is a well-researched field; however, no effective objective measure of quality exists. This paper details the creation, subjective evaluation, and analysis of a dataset for use in the development of an objective measure of quality for TSM. Comprised of two parts, the training component contains 88 source files processed using six TSM methods at 10 time scales, while the testing component contains 20 source files processed using three additional methods at four time scales. The source material contains speech, solo harmonic and percussive instruments, sound effects, and a range of music genres. Ratings (42 529) were collected from 633 sessions using laboratory and remote collection methods. Analysis of results shows no correlation between age and quality of rating; expert and non-expert listeners to be equivalent; minor differences between participants with and without hearing issues; and minimal differences between testing modalities. A comparison of published objective measures and subjective scores shows the objective measures to be poor indicators of subjective quality. Initial results for a retrained objective measure of quality are presented with results approaching average root mean squared error loss and Pearson correlation values of subjective sessions. The labeled dataset is available at http://ieee-dataport.org/1987.
 https://doi.org/10.1121/10.0001567

(Received 16 February 2020; revised 21 June 2020; accepted 23 June 2020; published online 14 July 2020)

[Editor: Paavo Alku] Pages: 201–210

I. INTRODUCTION

Time Scale Modification (TSM) is the process of modifying the duration of a signal without modifying timbre and pitch. It has found use in areas including music production, language learning, and speech recognition systems. Despite being a well-researched field, an effective objective measure of quality has not yet been published, limiting comparisons between TSM algorithms. When subjective evaluation has been used, each paper has used a unique set of source material and methods, further reducing comparison to only the methods involved in the evaluation. In order to develop an effective objective measure, a dataset with subjective quality labels is required. This work details the creation, subjective evaluation, and analysis of the first dataset for this purpose, and gives preliminary results for a neural-network-based objective measure of quality.

TSM algorithms most commonly modify the temporal domain by varying the ratio between analysis (S_a) and synthesis (S_s) shift sizes within an Analysis Modification Synthesis framework. This ratio, given by

$$\beta = \frac{1}{\alpha} = \frac{S_a}{S_s}, \tag{1}$$

shows α to be the change in signal duration (Roucos and Wilgus, 1985), while β is the playback speed (Sylvestre and Kabal, 1992) and will be used within this paper.

Algorithms for TSM can be classified into three main categories: frequency domain, time domain, and hybrid methods. In general, frequency-domain methods excel in scaling harmonically complex material but struggle to produce high quality results with highly transient signals. Time-domain methods are more effective at scaling transient signals but give poor results for polyphonic signals. Hybrid methods leverage the strengths of frequency and time domain methods to produce higher quality results (Driedger *et al.*, 2014).

Common artefacts produced during TSM include "phasiness" and reverberation (Laroche and Dolson, 1997; Portnoff, 1981), musical and metallic noise or undesirable roughness (Laroche and Dolson, 1999), a buzzy quality (Laroche, 2002), and transient smearing (Laroche and Dolson, 1999). Phasiness and reverberation are heard as a loss of spectral definition and are most commonly associated with frequency domain methods. Laroche and Dolson (1999) suggest that this is due to a change in relationship between the phases of bins in the spectral domain. Musical noise, also known as musical artefacts or musical tones, is due to isolated holes and/or peaks within the power spectrum (Torcoli, 2019). Within TSM, these artefacts are caused by periodicity introduced to noise bins during phase progression, due to the sum of sines model of the Short Time Fourier Transform (STFT). Depending on the frequency relationships between these periodic signals the noise will be perceived as musical for simple harmonic relationships and metallic for complex harmonic relationships. Transient smearing occurs due to the trade-off between STFT spectral and temporal resolution in frequency domain

[a]Electronic mail: timothy.roberts@griffithuni.edu.au, ORCID: 0000-0002-8937-0643.
[b]ORCID: 0000-0002-3553-3662.

algorithms. As the frame size increases to improve spectral resolution, temporal resolution decreases leading to smearing of transients in time. The buzzy quality, also known as transient skipping or duplication, is an artefact of time-domain methods in which transients may be skipped for $\beta > 1$ or duplicated for $\beta < 1$.

The aim of TSM is often noted; however, an exploration of ideal TSM has not been published. For the purpose of subjective evaluation, we describe ideal TSM as indistinguishable from a change by the sound source, that is, the processing should be transparent. A musician changing tempo or a speaker changing cadence would therefore be ideal and should be the goal for TSM algorithms. Consequently, ideal TSM should be determined by the sound source being scaled. For example, a dry recording of individual clicks simply requires temporal realignment of each click; however, a recording of sustained notes played on a violin would require the extension of the sustain section of the note's envelope. Further, in the case of a piano, one must consider whether the transient or harmonic nature of the source should be maintained. If a staccato melody played in the upper register without damping is to be slowed, should note decay be lengthened or should the decay be maintained with each note shifted to the new time scale? We argue that as the piano is a percussive instrument and unable to modify its amplitude envelope, the note decay should be maintained. This is counter to the processing applied by almost all published TSM algorithms. We propose that an ideal TSM algorithm would be sensitive to the signal source and be capable of modifying only the sustained portion of the amplitude envelope. This raises many questions in the processing of reverberation, vibrato, specific phonemes, and more. We consider that content aware or source sensitive TSM is an area with considerable potential for improving the quality of TSM.

The remainder of the paper is laid out as follows. Section II describes the TSM algorithms used to create the dataset and previous methodologies for quality evaluation. Section III describes the source files used in the creation of the dataset and the processing of the source material to create the processed dataset. Section IV describes the subjective testing methodology, opinion score normalization, results and analysis of the subjective testing, and dataset availability. Section V compares subjective results with published objective measures and provides preliminary results for a novel objective measure of quality. Finally, Sec. VI summarizes and draws conclusions from this research.

II. ALGORITHMS AND QUALITY EVALUATION

The Phase Vocoder (PV) is a frequency-domain method that uses the known phase progression between frames at the original time scale to calculate the phase progression between frames at the adjusted time scale. The digital implementation by Portnoff (1976) uses the STFT to calculate phase spectra and forms the basis for all PV methods published since. The PV is effective at scaling signals with a complex harmonic structure, however it introduces "phasiness" for non-integer values of α and is prone to transient smearing. See Laroche and Dolson (1999) for a detailed explanation.

The Identity Phase Locking Phase Vocoder (IPL) (Laroche and Dolson, 1999) reduces phasiness introduced by the PV algorithm. The PV maintains horizontal phase coherence within each STFT bin; however, the vertical phase coherence between bins is not maintained. In IPL, the phase of magnitude spectrum peaks are modified, with nearby bins locked to the phase progression of the closest peak. This method was extended through multi-resolution peak-picking and accounting for added or removed peaks by Karrer *et al.* (2006). These methods reduce phasiness, however they can introduce a spectral roughness known as metallic or musical noise.

The Waveform Similarity Overlap Add (WSOLA) algorithm (Verhelst and Roelands, 1993) is a time-domain method that uses the similarity between a frame and its natural progression in the input signal to minimize discontinuities in the time scaled signal. This is in contrast to previous methods that compare with the output signal (Moulines and Charpentier, 1990; Roucos and Wilgus, 1985). WSOLA effectively processes speech and monophonic musical signals, however due to the reliance on the fundamental frequency for alignment, produces low quality results for polyphonic signals.

Fuzzy Epoch Synchronous Overlap-Add (FESOLA) (Roberts and Paliwal, 2019) uses cross-correlation of glottal closure instants, known as epochs, for aligning frames of speech. Epochs are calculated using a Zero Frequency Resonator before smearing in the time-domain. The smearing improves the cross-correlation of epochs and accounts for changes in fundamental frequency. This method works well for speech and monophonic signals, however it is not effective at processing polyphonic signals.

Harmonic-Percussive Separation Time Scale Modification (HPTSM) of Driedger *et al.* (2014) is a hybrid method that uses median filtering of spectrograms for signal separation. WSOLA and IPL are used for percussive and harmonic components, respectively. Improved quality was shown over both individual methods. The method was also shown to compete with contemporary commercial state-of-the-art algorithms.

Multi-component Time-Varying Sinusoidal (uTVS) decomposition (Sharma *et al.*, 2017) uses a Mel-scale filterbank and the Hilbert transform to calculate instantaneous phase and frequency, bypassing phase unwrapping and the quasi-stationary assumption of traditional frequency-domain methods. As a result, temporal smearing and phasiness artefacts are reduced. This method slightly improves quality over HPTSM, with large improvements over traditional methods.

Elastique (Zplane Development, 2018) is a widely used commercial TSM method. While the algorithm is not publicly available, it is currently a state-of-the-art method and has been used in recent TSM subjective evaluations.

https://doi.org/10.1121/10.0001567

Fuzzy classification of spectral bins (FuzzyPV) (Damskägg and Välimäki, 2017), is an extension of the IPL. Spectral bins are given a degree of membership to three classes, sinusoidal, noise and transient, resulting in a fuzzy classification of each bin. Sinusoidal bins are scaled using IPL with phase locking applied to sinusoidal bins, while random phase is added to noise bins. Analysis phases of transients bins are simply relocated in time. Subjective evaluation shows improvement over HPTSM and similar performance to Elastique.

Non-Negative Matrix Factorization Time-Scale Modification (NMFTSM) by Roma *et al.* (2019) decomposes the signal into percussive events and harmonic components. Percussive events are copied directly to the output signal, while IPL is used for harmonic components. The duration of percussive events is preserved, however it is highly reliant on correct detection of the events and introduces novel artefacts.

Little formal subjective testing has been used to evaluate proposed methods, with most proposed methods providing results from informal testing. A wide variety of time scales and algorithms are used, with little consistency. Time scales are often limited with two to five time scales ($0.5 \leq \beta \leq 2$) reported in formal testing, with a bias toward $\beta < 1$. This reduces the number of files that require rating, but also limits algorithm evaluation. The difference in quality between $\beta < 1$ and $\beta > 1$ was mentioned briefly by Sylvestre and Kabal (1992). Since the release of the MATLAB TSM Toolbox (Driedger and Muller, 2014), PV, IPL, WSOLA, and HPTSM have been used in most evaluations, while comparisons to commercial algorithms are rare (Damskägg and Välimäki, 2017; Driedger *et al.*, 2014; Karrer *et al.*, 2006). The source audio used during testing also varies between papers with some papers using the files provided with the MATLAB TSM Toolbox. It was noted by Moulines and Laroche (1995) that a thorough perceptual evaluation of TSM approaches had not yet been undertaken.

Two objective measures have been proposed, Signal to Error Ratio (SER) by Roucos and Wilgus (1985) and synthesis consistency (D_M) by Laroche and Dolson (1999). SER accounts only for successive magnitude spectra, with no attention paid to phase spectra. D_M also compares the output frame's magnitude to the reconstructed signal's magnitude, however the "measure is not a clear indicator of phasiness" (Laroche and Dolson, 1999). Neither of these measures has seen continued use.

III. DATASET DESCRIPTION

The source material for the dataset was collated from the author's previous creative projects including films, concert, and field recordings as well as music written specifically for the dataset. Files were selected to give a broad spectrum of content with variation in TSM difficulty. The number of source files, methods, and time scales was determined by balancing the amount of content required to train a neural network and the number of ratings required for a "true" Mean Opinion Score (MOS). All content was converted to mono by averaging each pair of samples to remove the influence of poor handling of multi-channel files (Roberts and Paliwal, 2018) and normalized to ± 1 before TSM. All files are 16-bit with a sample rate of 44.1 kHz and range in signal pressure level from 56.62 to 86.92 dB with a mean and standard deviation of 73.37 and 6.75 dB.

The full dataset contains 34 musical, 37 solo instrument, and 37 voice files with a complete listing provided with the dataset. The total playback length of the source files is 6 min and 42 s. Duration was kept short, with a mean of 3.7 s and standard deviation of 1.6 s, to limit the duration after time-scaling. Files were recorded using a combination of close microphone placement, multi-microphone concert recording, digital synthesis and sampling techniques, and shotgun, lapel, and large diaphragm condenser microphones. These variations in source material allow for extended subjective evaluation of future TSM methods. The musical and solo files contain synthetic and organic sound sources across classical, rock, jazz, and electronic genres. Voice files contain singing and male, female, and child speech. Finally, the evaluation source files contain a mix of each file type and were used in the generation of the test and evaluation subsets. Table I shows an overview of the signal sources.

To form the training set, the source dataset was processed using the first six methods previously mentioned at 10 time scale ratios resulting in 5280 processed files. Time scale ratios of 0.3838, 0.4427, 0.5383, 0.6524, 0.7821, 0.8258, 0.9961, 1.381, 1.667, and 1.924 were generated randomly, but adjusted to ensure coverage across the range of interest. The testing set used Elastique, FuzzyPV, and NMFTSM at four random time scales in four bands across $0.25 \leq \beta \leq 2$, resulting in 240 testing files. Subjective evaluation was conducted for both the training and testing sets. An additional evaluation set was created and is discussed in Sec. V. Full dataset generation took approximately 3 days on a medium to high end workstation.

TABLE I. Signal sources in each dataset class. Sources considered are Total, Brass, Percussion, Piano, Rhythm Section, Sound Effects, Strings, Synthesizers, Woodwinds, Child, Female, Male, and Singing. All sources within a file are counted separately.

	Total	Br.	Perc.	Piano	Rhythm	SFX	String	Synth.	Wood.	Ch.	F.	M.	Sing.
Music	27	6	7	6	8	2	3	9	12	—	—	1	2
Solo	31	—	11	3	4	1	1	3	11	—	—	—	—
Voice	30	—	—	—	—	—	—	—	—	3	12	15	4
Eval.	20	1	2	2	3	1	1	2	9	1	3	3	—

The MATLAB TSM Toolbox (Driedger and Muller, 2014) was used with default settings for WSOLA, HPTSM, and Elastique time-scaling. FuzzyPV and NMFTSM used provided implementations with default settings. Author implementations of PV, IPL, uTVS, and FESOLA were used with Hann windowing throughout and parameters chosen to maximize informal subjective evaluation. All files were normalized after processing. The PV and IPL used a frame length of 2048 samples (46.4 ms) and synthesis hop of 512 samples. FESOLA used a frame length of 1024 samples (23.2 ms). WSOLA used a frame length of 1024 samples (23.2 ms), a synthesis hop of 512 samples, and a tolerance of 512 samples. HPTSM used identical IPL parameters while WSOLA had a frame size of 256 samples (5.8 ms) and a synthesis hop of 64 samples. uTVS was implemented using six times oversampling and a filterbank containing 88 filters to maintain the relationship between the signal sample rate and filterbank length of the original paper. During testing, an error in the uTVS implementation was found that introduced discontinuities within spectra during processing at $0.9 \leq \beta \leq 1.1$ for some files. However, as the purpose of the subjective testing was to rate multiple files with a variety of artefacts, they were not removed from the dataset. The error was rectified before creation of the evaluation subset.

IV. SUBJECTIVE TESTING

Subjective testing was undertaken in two phases. Initial testing was conducted internally within the laboratory. Due to the large number of responses needed per file, testing transitioned to an online browser-based test using the Web Audio Evaluation Tool (WAET) (Jillings *et al.*, 2015), shown in Fig. 1. Remote testing greatly increased the number of participants in the study. Participants were contacted in person, directly through social media and email, through mailing lists, and public posts on websites such as Reddit and Facebook.

Testing followed ITU-R BS.1248–1 (ITU-T, 2019) recommendations for general methods for the subjective assessment of sound quality as close as practicable, resulting in the following testing parameters. Files were presented in reference-processed pairs with no limits placed on the amount of playback before moving to the next file. Checks were included to ensure both files were played at least once. A continuous grading scale was used in conjunction with a quality scale, where Poor–Excellent corresponds to scores of 1–5. Sessions contained a randomised selection of processed files, presented in random order, with participants free to choose the session they would evaluate. The amount of content per session was refined during testing, for a maximum session duration of 20 min. Toward the end of testing, the sessions were restricted to files that had limited responses to reduce MOS standard deviation.

Initial testing was undertaken using a bespoke MATLAB GUI that presented individual reference-processed pairs, allowed for saving and restoring of sessions, user input of name, sound transducer, and a check that the participant had no known hearing issues. Participants received training before beginning testing, including explanations of the purpose of TSM and common artefacts with audio examples. A small initial test session of 33 files was completed before a random session was assigned. Each session contained 18 min of audio, approximately 200 files, randomly selected from the pool of processed audio files. Participants could elect to evaluate additional sessions following a break equal in length to the completed session.

To increase the number of participants, the WAET was used. A small number of sessions were evaluated containing 100 files before reduction to 60 files based on participant feedback of session duration. Training identical to laboratory testing was available from the index page, which contained links to each test session. The index page contained reminders to use headphones in a quiet space during testing and a random number generator to suggest which test session the participant should complete. Before each session, name, age, sound transducer, experience in critical evaluation of sound, and any known hearing issues were collected. Participants could also elect to provide an email address to

FIG. 1. WAET user interface used for remote testing. Shown with two file pairs.

https://doi.org/10.1121/10.0001567

be contacted for future studies. Each session was split into pages containing six reference-processed pairs.

To remove bias and variability between sessions, opinion scores were normalized according to ITU-R BS1284 (ITU-T, 2019) using

$$Z_i = \frac{x_i - \bar{x}_{si}}{\sigma_{si}}\sigma_s + \bar{x}_s, \tag{2}$$

where Z_i is the normalized result, x_i is the opinion score of subject i, $\bar{x}_{si}$ is the mean score for subject i in session s, $\bar{x}_s$ is the mean score of all subjects in session s, σ_s is the standard deviation for all subjects in session s, and σ_{si} is the standard deviation for subject i in session s. As the files in each session were unique, means and standard deviations were calculated on the subset of files matching those in the session. Normalized opinion scores were not truncated, however MOS were limited to the subjective interval of 1–5.

A. Results

A total of 42 529 file ratings were collected from 263 participants across 633 sessions, with 10 354 ratings collected during laboratory testing. Participants ranged in age from 16 to 66 with a median age of 30. Expert listeners contributed 52.36% of ratings. Twelve files were limited to a MOS of 1, while 28 files were limited to a MOS of 5.

Due to the different files and time scale ratios used for the testing subset, direct comparison between methods in training and testing subsets was not appropriate. However, a general comparison was achieved through local averaging of MOS, centered around training time scale ratios. Means of adjacent time scale ratios, bounded by 0.3 and 3, defined the local areas. While 0.3 is greater than some time scales used within the testing set, it was set empirically to include enough data points, while limiting the impact of much slower time scales. Mean MOS for testing subset methods are noisier due to the smaller number of files, and non-uniform difficulty in processing each signal.

Two measures of reliability were used for each session. The Root Mean Squared Error (RMSE) denoted by $\mathcal{L}$ is given by

$$\mathcal{L} = \sqrt{\frac{\sum_{i=1}^{N}(\bar{x}_i - x_i)^2}{N}}, \tag{3}$$

where the number of files within the session is denoted by N, x_i is the participants opinion score for the file, and $\bar{x}_i$ is the overall MOS for the file. The Pearson Correlation Coefficient (PCC), denoted by ρ, given by

$$\rho = \frac{\mathrm{cov}(\boldsymbol{x}, \bar{\boldsymbol{x}})}{\sigma_{\boldsymbol{x}}\sigma_{\bar{\boldsymbol{x}}}}, \tag{4}$$

was also used where $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ denote sets of opinion scores and MOS for the session and $\sigma_{\boldsymbol{x}}$ and $\sigma_{\bar{\boldsymbol{x}}}$ are the standard deviation of $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$. These measures were calculated for each session before and after normalization. Outliers, calculated prior to normalization and shown in Fig. 2, were determined as sessions in which $\mathcal{L}$ or ρ were further than three scaled median absolute deviations away from their respective medians. This resulted in the removal of 45 sessions containing a total of 2102 ratings (4.94%) from the final pool of sessions.

Following outlier removal and normalization, $\mathcal{L}$ and ρ means of 0.771 and 0.791 improved to 0.682 and 0.799. Distributions of $\mathcal{L}$ and ρ pre- and post-normalization can be seen in Fig. 3.

The use of Intraclass Correlation Coefficients (ICC) was explored, however as the subjective results are neither fully crossed nor fully nested, ICC cannot be used. Instead, the interrater reliability for Ill-Structured Measurement Designs of Putka *et al.* (2008) was used, calculated by

$$G(q,k) = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_T^2 + \left(q\hat{\sigma}_R^2 + \frac{\hat{\sigma}_{TR,e}^2}{\hat{k}}\right)}, \tag{5}$$

where $\hat{\sigma}_T^2$ is the estimated variance for file main effects (true score), $\hat{\sigma}_R^2$ is the estimated variance for participant main effects, $\hat{\sigma}_{TR,e}^2$ is the estimated variance components for the combination of residual effects and file-participant interaction, and $\hat{k}$ is the harmonic mean of the number of participants per file. q scales the contribution of $\hat{\sigma}_R^2$ based on the overlap between the sets of participants who rate each file, and is calculated by

$$q = \frac{1}{\hat{k}} - \frac{\sum_i \sum_{i'} \frac{c_{i,i'}}{k_i k_{i'}}}{N_t(N_t - 1)}, \tag{6}$$

where $c_{i,i'}$ is the number of participants that each pair of files (i, i') share, k_i and k_i' are the number of participants who rated files i and i', respectively, and N_t is the total number of participants in the sample. This measure gives an overall

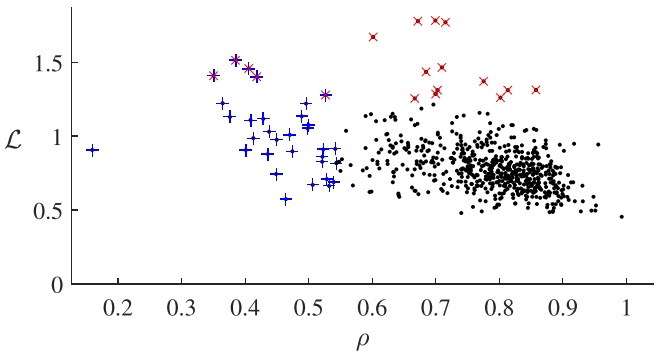

FIG. 2. (Color online) Distribution of PCC and RMSE for all sessions before normalization and outlier removal. Blue plus symbols mark PCC outliers, while red crosses mark RMSE outliers.

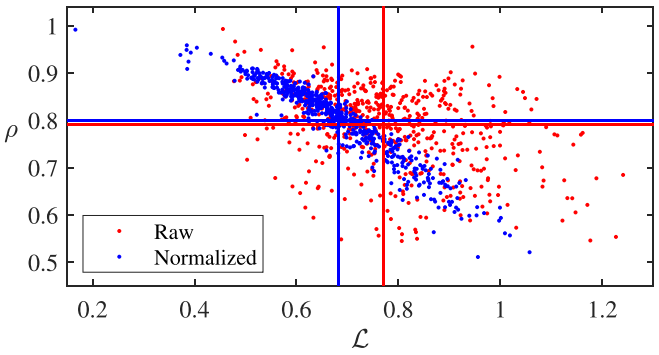

FIG. 3. (Color online) Distribution of PCC and RMSE for each session before normalization. Horizontal and vertical lines denote means.

rater reliability [$G(q, k)$] of 0.871 prior to normalization and 0.909 post normalization.

For an overview of all results, Fig. 4 shows all normalized file ratings ordered by ascending MOS. All opinion scores are shown in the histogram with the overlaid red line showing the MOS for each file. It can be seen that when the TSM quality is very high or very low there is a greater consensus amongst participants, however there is a large variance in opinion for files with mid-range quality. It can also be seen that the MOS tracks below the majority of responses in the Good to Excellent range, suggesting a difference between MOS and a majority of opinion scores. Median opinions scores were explored, based on Jamieson (2004), resulting in tighter groupings, however there was no significant change in averaged scores nor improvement in session reliability. Median opinion scores have nonetheless been included as labels with the dataset, along with mean and median opinion scores calculated before normalization.

All methods show improvement in quality as β approaches 1, as is to be expected. However, the implementation of uTVS gave poor performance when time-scaling at 0.9961, see Sec. III, but achieved state-of-the-art performance for all other time scales. Figure 5 shows the results of each method for each time scale, averaged across all files. When comparing two inverse time scale ratios, for example $\beta = 0.5$ and $\beta = 2$, the slower of the pair is lower in quality, suggesting that slowing a file down is perceptually more

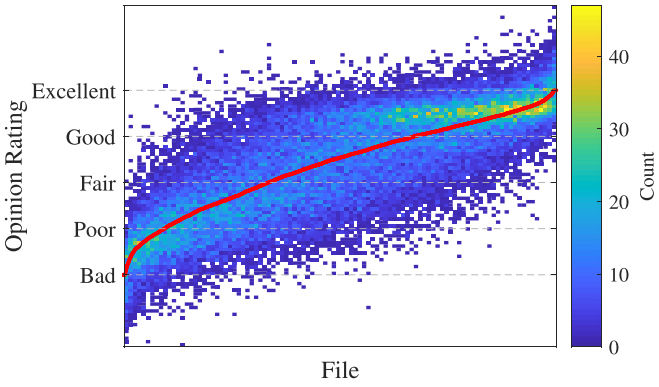

FIG. 4. (Color online) Two-dimensional histogram of normalized responses, ordered by ascending MOS (red line).

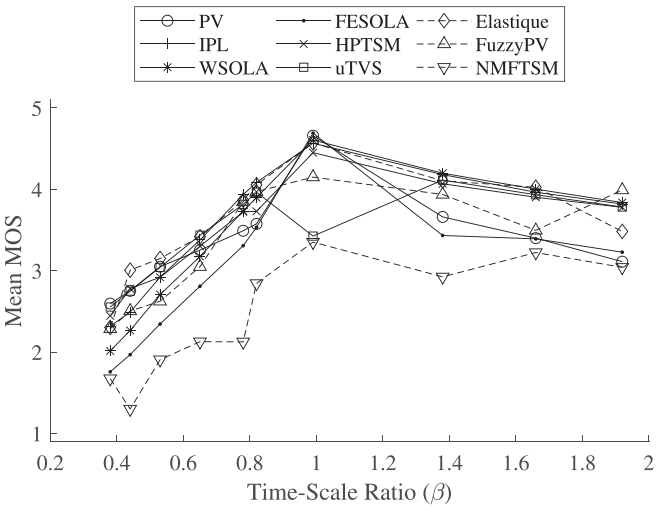

FIG. 5. Overall means for each method at each time scale for all evaluated files.

difficult than increasing its speed. This is consistent with the testing of Sharma *et al.* (2017), however the effect is more pronounced within this testing. Of interest are two specific cases, that of PV and WSOLA. For $\beta < 1$, PV is perceived to have a higher quality than WSOLA, however this is reversed for $\beta > 1$. It can then be inferred that different artefacts are perceived as having a greater impact on the quality of the TSM. We propose that for $\beta < 1$, the transient-doubling of WSOLA is perceived as worse than the phasiness and transient smearing of the PV, while for $\beta > 1$ transient skipping is less detrimental than the artefacts introduced by the PV. This is a similar finding to Moinet and Dutoit (2011), who noted that some listeners preferred PV artefacts in some cases. Similarly, comparison of PV and IPL shows a change in preference toward the smeary PV artefacts for large reductions in speed, over the metallic artefacts of IPL. The PV was rated comparably to state-of-the-art methods for the three smallest β.

A surprising result is the high performance of IPL in comparison to HPTSM and uTVS. HPTSM achieved numerically similar results to those given in Driedger *et al.* (2014). However, while HPTSM was shown to be greater in MOS by 1, our testing found IPL to be rated higher for all except the two slowest time scale ratios. Artefacts due to harmonic-percussive separation, the use of WSOLA with a very short frame length or the lower sample-rate of the files used in the MATLAB TSM Toolbox may be the cause. Similarly, the reduced sample-rate in original uTVS testing may have contributed to the variance in MOS between testing. Future research should include comparisons between different IPL implementations.

Algorithm performance per class generally follows that of the overall results. As expected however, there are differences in performance quality between methods dependent on the source material. When the mean MOS for each class are considered and $\beta = 0.9961$ results excluded, uTVS is preferred for music and solo instrument sources while WSOLA is preferred for voice sources. However, the

https://doi.org/10.1121/10.0001567

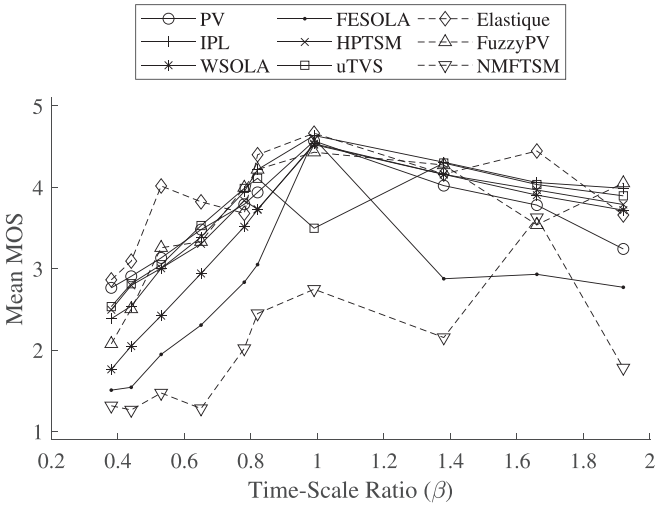

FIG. 6. Mean MOS for each method at each time scale for musical source material.

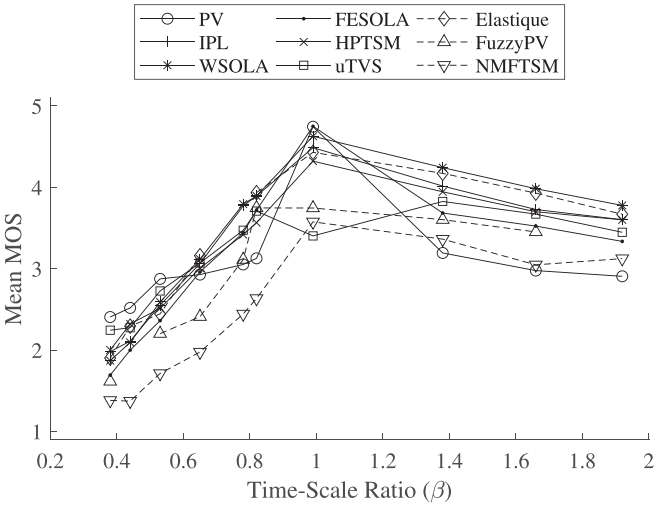

FIG. 8. Mean MOS for each method at each time scale for voice source material.

differences in averaged ratings are minor in most cases. Exact mean results have not been reported here as the primary focus is rating time scaled files, rather than definitive evaluation of different TSM methods.

Perception of processing quality for musical sources, Fig. 6, confirms the lower quality of time-domain methods, with FESOLA and WSOLA giving poor results. The most interesting result here is that the PV is consistently rated higher than other methods for $\beta < 0.7$ and is comparable for other β. If ratings are averaged for each source file, it is possible to identify "difficult" files to process. Files with uncorrelated high frequency content were rated poorly, while clean, harmonically simple musical excerpts were rated highly. Signals containing more transient material were rated lower than less transient material. Mean ratings ranged from 2.76 for Jazz_1.wav to 3.94 for Yellow_2.wav.

Mean MOS results for the solo instrument class of signals, shown in Fig. 7, improve over musical and voice classes with the exception of the PV for $\beta > 1$. Synthesizer bass sounds were the lowest rated, followed by noisy percussion, polyphonic instruments, and tuned percussion, with monophonic harmonic instruments rated highest. The combination of low frequencies with significant transients within the synthesizer bass was particularly troublesome for all TSM methods. Mean file ratings ranged from 2.54 for Synth_Bass_1.wav to 4.17 for Ocarina_01.wav.

In considering mean MOS for voice signals, shown in Fig. 8, WSOLA is preferred for $\beta > 1$, while the preference is less clear for $\beta < 1$. Most methods, except the PV and NMFTSM, were rated similarly for $0.6 < \beta < 1$, however the PV is clearly preferred for $\beta < 0.6$. After this point, smoothness is preferred over transient doubling and metallic artefacts. When considering mean file ratings, the 11 lowest rated files were all male voices, with female and child voices as the seven highest rated files. This mirrors results by Sylvestre and Kabal (1992) who suggested poor frequency resolution for lower frequencies as well as short frame sizes as causes for lower quality. Mean file ratings ranged from 2.73 for Male_18.wav to 3.59 for Child_01.wav.

The mean standard deviation across all files was 0.802 and 0.718, before and after normalization, respectively. As can be seen in Fig. 9, the range of standard deviation values

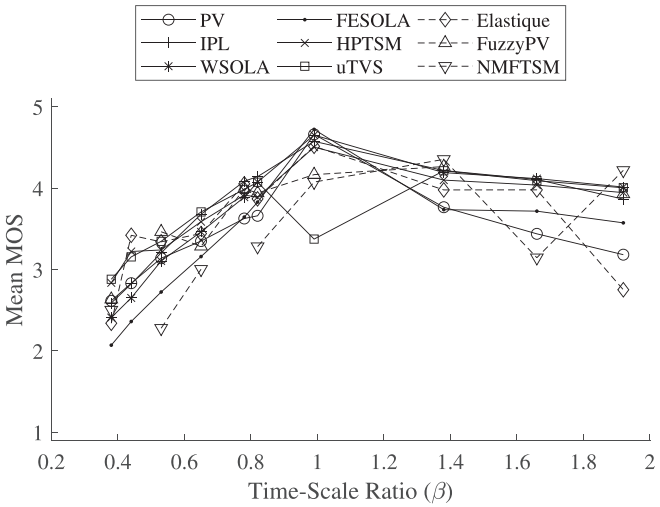

FIG. 7. Mean MOS for each method at each time scale for solo instrument source material.

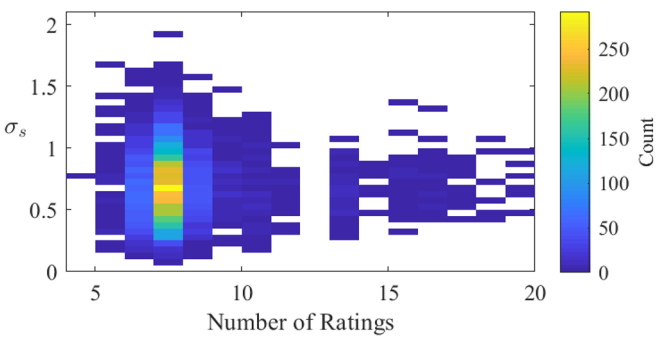

FIG. 9. (Color online) MOS standard deviation against the number of responses for that file.

converges as the number of responses for the file increases. During testing (around 19 000 ratings) this graph showed convergence at around seven ratings per file. As a result, a minimum of seven ratings per file was set as the target to give a true representation of the quality of the audio file. While there are files that have yet to converge, this is a small subset of the total dataset.

Comparisons between expert and non-expert listeners, participants with and without known hearing issues and testing modalities were undertaken using the two one sided tests (TOST) of Hauck and Anderson (1984) and Lakens (2017). TOST begins with the null hypothesis of non-equivalent means and uses TOST to show equivalence within a given interval. The interval can be given as a raw score or a standardized difference. If the confidence interval (CI) for the difference of the means falls within the equivalence interval, the null hypothesis is rejected and equivalence can be claimed. Analysis was undertaken on session RMSE and PCC values before normalization. The equivalence interval was calculated at 5% of the reference sample's mean and CIs of 95% were used throughout. Cohen's sample d is also given for indication of effect size, where d $\approx$ 0.2 is a small effect size.

ITU Recommendation BS.1284 (ITU-T, 2019) recommends investigation of the relationship between expert and non-expert listeners. Participants selected if they had experience critically evaluating the quality of audio. RMSE and PCC for non-expert listeners were found to be equivalent to those of expert listeners, with equivalence intervals shown in Fig. 10. Testing RMSE gave a maximum p value of 0.0498 and d of 0.1273. Testing PCC gave a maximum p value of 4.67×10^{-6} and d of 0.1059. We propose that equivalence is a result of the reference-test style of testing and the medium to large impairment in the processed signal, reducing the importance of highly trained critical listening skills for this type of subjective testing.

Participants also reported any known hearing issues, with an open answer text box given for responses. Results were not excluded if known issues were reported, but were instead manually sorted into a binary classification of "No known hearing issues" and "Any known hearing issues." Hearing issues included highly descriptive explanations such as "-6 dB above 14 kHz," a range of tinnitus severity, age related hearing changes, and "I like punk music." PCC for participants with any hearing issues were found to be equivalent to those without issue, while RMSE was not found to be equivalent. Equivalence intervals are shown in Fig. 11. Testing RMSE gave a maximum p value of 0.2467 and d of 0.0958. Testing PCC gave a maximum p value of 0.0245 and d of 0.1219. Our proposed explanation is two-fold. Those participants who reported known hearing issues in great detail were also expert listeners and familiar with the shortcomings of their own auditory system. Additionally, as the participants were presented with the source and processed files and asked to rate the quality of the processing, any issue within the auditory system would affect perception of both files. The small number of sessions classified as "any issue," 33 compared to 554 for "no issue," also impacts this result, greatly increasing the standard error. A t-test applied to RMSE was unable to reject that the means are equal with a p-score of 0.4985. Increasing the equivalence interval to $\pm$9.32% allows RMSE equivalence to be claimed. Due to the strong PCC equivalence and close RMSE equivalence, we find no reason to reject sessions in which hearing issues were reported.

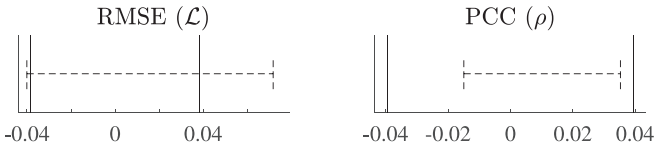

FIG. 11. TOST $(1-\alpha)$ 100% CI for equivalence of means of participants with and without hearing issues for $\alpha = 0.05$. Equivalence interval of $\pm$5% of mean for participants without hearing issues.

As testing was undertaken in different modalities, comparative analysis of results is necessary. PCC for remote participants were found to be equivalent to laboratory participants, while RMSE was not found to be equivalent. Equivalence intervals are shown in Fig. 12. Testing RMSE gave a maximum p value of 0.3474 and d of 0.2126. Testing PCC gave a maximum p value of 0.0013 and d of 0.0931. A t-test applied to RMSE was unable to reject that the means are equal with a p-score of 0.4693. Increasing the equivalence interval to $\pm$8.14% allowed RMSE equivalence to be claimed. Due to the strong PCC equivalence and close RMSE equivalence, we found no reason to reject either testing mode.

Analysis of the possible impact of age on the quality of the participant's responses was undertaken. Correlations of 0.108 and -0.001 were found between the age of the participant and the RMSE or PCC, respectively, showing no impact of age on evaluation ability.

The labeled dataset is available, under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, through IEEE-Dataport at http://ieee-dataport.org/1987. Implementation and additional source code is available at github.com/zygurt/TSM.

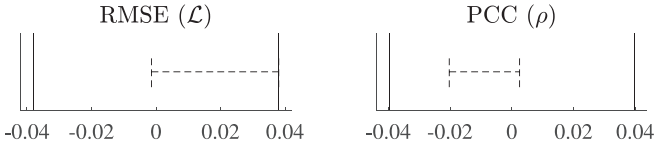

FIG. 10. TOST $(1-\alpha)$ 100% CI for equivalence of participant experience for $\alpha = 0.05$. Equivalence interval of $\pm$5% of expert participant means.

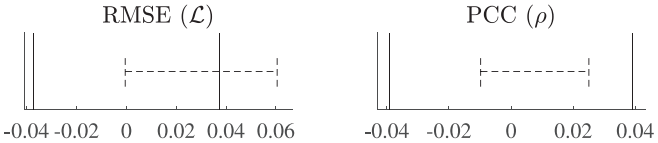

FIG. 12. TOST $(1-\alpha)$ 100% CI for equivalence of testing modality means for $\alpha = 0.05$ Equivalence interval of $\pm$5% of laboratory participant means.

https://doi.org/10.1121/10.0001567

V. TOWARD AN OBJECTIVE MEASURE OF QUALITY

Comparison between MOS and previous objective measures, SER and D_M, found correlations of 0.3707 and 0.1574, respectively, by averaging absolute correlations for $\beta < 1$ and $\beta > 1$. Signals were aligned through time axis interpolation of the reference magnitude spectrum to the duration of the test spectrum.

Perceptual Evaluation of Audio Quality (PEAQ) (ITU-T, 2001; Thiede *et al.*, 2000) is often used for objective quality evaluation. PEAQ extracts perceptually informed features, using differences between reference and test signals, which are fed into a small neural network to predict subjective scores. Direct application to time-scaled signals is not possible however, due to a loss of alignment during TSM. Initial testing, applying the dataset in the design of an objective measure of quality, was undertaken using a modified version of PEAQ. Signals were aligned as above and gave similar correlation to MOS as SER and D_M. The original PEAQ basic neural network was retrained to the subjective MOS, with 10% of the training set reserved for validation. Training used seeds of 0 to 99, with the optimal epoch given by the minimum overall distance ($\mathcal{D}$)

$$\mathcal{D} = \|[\hat{\rho}, \hat{\mathcal{L}}]\|_2, \tag{7}$$

where $\hat{\rho}$ and $\hat{\mathcal{L}}$ are calculated by

$$\hat{\rho} = \|[1 - \bar{\boldsymbol{\rho}}, (\max(\boldsymbol{\rho}) - \min(\boldsymbol{\rho}))]\|_2, \tag{8}$$

$$\hat{\mathcal{L}} = \|[\bar{\boldsymbol{\mathcal{L}}}, (\max(\boldsymbol{\mathcal{L}}) - \min(\boldsymbol{\mathcal{L}}))]\|_2, \tag{9}$$

where $\boldsymbol{\rho} = [\rho_{tr}, \rho_{val}, \rho_{te}]$, $\boldsymbol{\mathcal{L}} = [\mathcal{L}_{tr}, \mathcal{L}_{val}, \mathcal{L}_{te}]$ and *tr*, *val*, and *te* denote training, validation, and testing. The best network achieved a $\mathcal{D}$ of 0.731 and an $\bar{\boldsymbol{\mathcal{L}}}$ of 0.668 and $\bar{\boldsymbol{\rho}}$ of 0.719, placing it at the 11th and 17th percentiles of subjective sessions.

An evaluation set was created by processing the testing subset source files with all methods previously mentioned, at 20 time scale ratios in the range of $0.22 < \beta < 2.2$. The mean objective output for each method across the range of time scales is shown in Fig. 13.

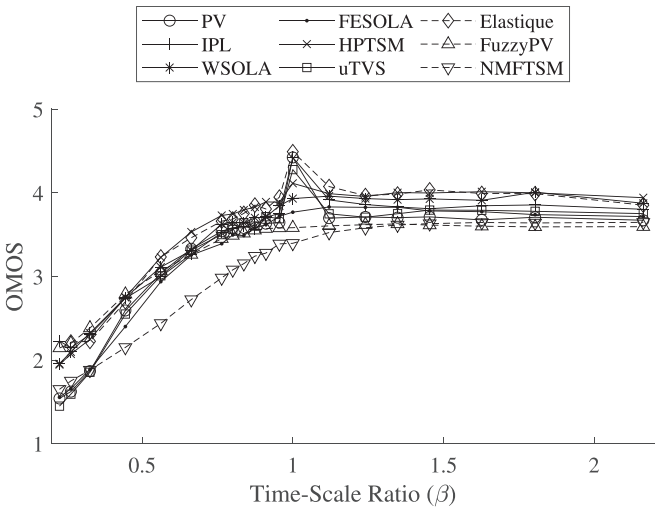

FIG. 13. Objective MOS for each method in the evaluation set, averaged at each time-scale ratio.

The output exhibits a similar shape to the subjective results, however it only moves away from the mean for $\beta < 0.75$ and $\beta = 1$. Development of an accurate objective measure of quality for TSM algorithms is now achievable and the aim of future work.

VI. CONCLUSION

This paper detailed the creation, subjective evaluation, and analysis of a dataset and its use in the development of an objective measure of quality for time-scaled audio. Six TSM methods processed 88 source files at 10 time scales resulting in 5280 processed signals for a training subset. Three additional methods at four random time scales resulted in 240 signals for a testing subset. Ratings (42 529) were collected from 633 sessions using laboratory and remote collection methods. Preliminary results for an objective measure of quality were presented, which achieved an RMSE loss of 0.668 and PCC of 0.719. The aim of future work is the design of an improved objective measure of quality for TSM using the dataset, to assist in comparative evaluation of novel methods.

ACKNOWLEDGMENTS

The authors would like to acknowledge and thank all of the participants who assisted in the subjective evaluation of the dataset as well as the reviewers for their comments that significantly improved the paper.

Damskägg, E., and Välimäki, V. (**2017**). "Audio time stretching using fuzzy classification of spectral bins," Appl. Sci. **7**(12), 1293.
Driedger, J., and Muller, M. (**2014**). "TSM toolbox: MATLAB implementations of time-scale modification algorithms," in *Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14)*, Erlangen, Germany, pp. 1–8.
Driedger, J., Muller, M., and Ewert, S. (**2014**). "Improving time-scale modification of music signals using harmonic-percussive separation," IEEE Signal Process. Lett. **21**(1), 105–109.
Hauck, W. W., and Anderson, S. (**1984**). "A new statistical procedure for testing equivalence in two-group comparative bioavailability trials," J. Pharmaco. Biopharma. **12**(1), 83–91.
ITU-T (**2001**). "Itu-r bs. 1387-1: Method for objective measurements of perceived audio quality," Technical Report.
ITU-T (**2019**). "Itu-r bs. 1284-1: General methods for the subjective assessment of sound quality," Technical Report.
Jamieson, S. (**2004**). "Likert scales: How to (ab)use them," Med. Ed. **38**(12), 1217–1218.
Jillings, N., Moffat, D., De Man, B., and Reiss, J. D. (**2015**). "Web Audio Evaluation Tool: A browser-based listening test environment," in *12th Sound and Music Computing Conference*, Maynooth, Ireland, pp. 147–152.
Karrer, T., Lee, E., and Borchers, J. (**2006**). "PhaVoRIT: A phase vocoder for real-time interactive time-stretching," Technical Report.
Lakens, D. (**2017**). "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses," Soc. Psychol. Person. Sci. **8**(4), 355–362.
Laroche, J. (**2002**). *Time and Pitch Scale Modification of Audio Signals* (Springer, Boston, MA), pp. 279–309.
Laroche, J., and Dolson, M. (**1997**). "Phase-vocoder: About this phasiness business," in *IEEE Proceedings of the 1997 Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 1–4.

Laroche, J., and Dolson, M. (**1999**). "Improved phase vocoder time-scale modification of audio," IEEE Trans. Speech Audio Process. **7**(3), 323–332.

Moinet, A., and Dutoit, T. (**2011**). "PVSOLA: A phase vocoder with synchronized overlap-add," in *Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11)*, Paris, France, pp. 269–275.

Moulines, E., and Charpentier, F. (**1990**). "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Commun. **9**(5–6), 453–467.

Moulines, E., and Laroche, J. (**1995**). "Non-parametric techniques for pitch-scale and time-scale modification of speech," Speech Commun. **16**(2), 175–205.

Portnoff, M. (**1976**). "Implementation of the digital phase vocoder using the fast Fourier transform," IEEE Trans. Acoust., Speech, Signal Process. **24**(3), 243–248.

Portnoff, M. (**1981**). "Time-scale modification of speech based on short-time Fourier analysis," IEEE Trans. Acoust., Speech, Signal Process. **29**(3), 374–390.

Putka, D. J., Le, H., McCloy, R. A., and Diaz, T. (**2008**). "Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability," J. Appl. Psychol. **93**(5), 959–981.

Roberts, T., and Paliwal, K. K. (**2018**). "Stereo time-scale modification using sum and difference transformation," in *IEEE 2018 12th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–5.

Roberts, T., and Paliwal, K. K. (**2019**). "Time-scale modification using fuzzy epoch-synchronous overlap-add (FESOLA)," in *IEEE 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 31–34.

Roma, G., Green, O., and Tremblay, P. (**2019**). "Time scale modification of audio using non-negative matrix factorization," in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx-19)*, Birmingham, United Kingdom, pp. 1–6.

Roucos, S., and Wilgus, A. (**1985**). "High quality time-scale modification for speech," in *IEEE Proceedings of ICASSP '85*, Vol. 10, pp. 493–496.

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

Sylvestre, B., and Kabal, P. (**1992**). "Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation," in *IEEE Proceedings of ICASSP '92*, Vol. 1, pp. 81–84.

Thiede, T., Treurniet, W. C., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J. G., and Colomes, C. (**2000**). "PEAQ - The ITU standard for objective measurement of perceived audio quality," J. Audio Eng. Soc. **48**(1/2), 3–29; available at http://www.aes.org/e-lib/browse.cfm?elib=12078.

Torcoli, M. (**2019**). "An improved measure of musical noise based on spectral kurtosis," in *IEEE 2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 90–94.

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

Zplane Development (**2018**). "Èlastique time stretching & pitch shifting sdks (version 3.2.5) [computer program]," http://licensing.zplane.de/technology\#elastique (Last viewed October 31, 2019).