# An objective measure of quality for time-scale modification of audio

Timothy Roberts, and Kuldip K. Paliwal

Citation: The Journal of the Acoustical Society of America **149**, 1843 (2021); doi: 10.1121/10.0003753
View online: https://doi.org/10.1121/10.0003753
View Table of Contents: https://asa.scitation.org/toc/jas/149/3
Published by the Acoustical Society of America

# An objective measure of quality for time-scale modification of audio

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

**ABSTRACT:**
Objective evaluation of audio processed with time-scale modification (TSM) remains an open problem. Recently, a dataset of time-scaled audio with subjective quality labels was published and used to create an initial objective measure of quality (OMOQ). In this paper, an improved OMOQ for time-scaled audio is proposed. The measure uses handcrafted features and a fully connected network to predict subjective mean opinion scores (SMOS). Basic and advanced perceptual evaluation of audio quality features are used in addition to nine features specific to TSM artefacts. Six methods of alignment are explored with interpolation of the reference magnitude spectrum to the length of the test magnitude spectrum giving the best performance. The proposed measure achieves a mean root mean square error of 0.490 and a mean Pearson correlation of 0.864 to SMOS, equivalent to the 97th and 82nd percentiles of the subjective sessions, respectively. The proposed measure is used to evaluate TSM algorithms, finding that Elastique gives the highest objective quality for solo instrument and voice signals, whereas the identity phase-locking phase vocoder gives the highest objective quality for music signals and the best overall quality. The objective measure is available online at https://www.github.com/zygurt/TSM. *© 2021 Acoustical Society of America.*
https://doi.org/10.1121/10.0003753

(Received 23 June 2020; revised 17 February 2021; accepted 22 February 2021; published online 17 March 2021)

[Editor: Paavo Alku] Pages: 1843–1854

## I. INTRODUCTION

Time-scale modification (TSM) is the process of modifying the duration of a signal without modifying the pitch of the signal. To justify the quality of the processing, subjective testing must be attempted. But, it is expensive and time consuming. Objective methods are available for evaluation of audio quality; however, these methods require reference and test signals of identical duration. Consequently, most published objective measures cannot be applied to this context. Two objective measures, the signal to error ratio (SER) by Verhelst and Roelands (1993) and $D_M$ by Laroche and Dolson (1999), have been proposed. Nonetheless, they are shown to be only high level indicators of "phasiness" or quality (Laroche and Dolson, 1999). In this work, we propose the first objective measure of quality (OMOQ) for time-scale modified audio. It uses handcrafted features with deep-learning methods and is trained using a recently published dataset (Roberts, 2020). The contributions of this paper are an OMOQ for time-scaled audio, novel quality features specific to TSM and comparison of TSM methods using the objective measure.

Objective measures of quality seek to predict the worth of a test signal and can be broadly classified into two classes, traditional and machine learning. Traditional measures, such as the perceptual evaluation of speech quality of ITU-T (2001b), STOI of Gomez *et al.* (2011), and the TSM specific measures of SER and $D_M$, are purely analytical in nature. Machine learning methods use neural networks to develop a relationship between subjective evaluations of the test signal and handcrafted or data-driven features extracted from reference and test signals as in the perceptual evaluation of audio quality (PEAQ; ITU-T, 2001a). Deep learning allows for objective measures that do not require a reference file as in Avila *et al.* (2019) for speech quality; however, these methods have not yet been applied to TSM or general sound sources.

Training of deep-learning methods requires a large amount of labelled signals. Recently, a dataset of time-scaled audio with subjective labels was published for this purpose (Roberts and Paliwal, 2020). Reference files were drawn from a large variety of sources, including speech, singing, solo harmonic, and percussive instruments, as well as a variety of musical genres. The training subset, containing 5280 processed files, was generated using 6 methods to time scale 88 reference files at 10 ratios. The methods used were the phase vocoder (PV) of Portnoff (1976), the identity phase-lockingp vocoder (IPL) of Laroche and Dolson (1999), waveform similarity overlap-add (WSOLA) of Verhelst and Roelands (1993), fuzzy epoch synchronous overlap-add (FESOLA) of Roberts and Paliwal (2019), harmonic percussive separation time-scale modification (HPTSM) of Driedger *et al.* (2014), and mel-scale sub-band modelling (uTVS) of Sharma *et al.* (2017). Playback speeds of 0.3838, 0.4427, 0.5383, 0.6524, 0.7821, 0.8258, 0.9961, 1.381, 1.667, and 1.924 were used as time-scale ratios ($\beta$) for the training subset. The testing subset, containing 240 files, was created using 3 additional methods to

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

time scale 20 reference files at a random $\beta$ in each band of $0.25 < \beta < 0.5$, $0.5 < \beta < 0.8$, $0.8 < \beta < 1$, and $1 < \beta < 2$. Elastique by Zplane Development (2019), the phase vocoder using fuzzy classification of bins (FuzzyPV) of Damskägg and Välimäki (2017), and non-negative matrix factorisation time-scale modification (NMFTSM) of Roma *et al.* (2019) were used to generate the testing subset. Different TSM methods were used to ensure that the training and testing sets were independent. Finally, an evaluation subset was generated by processing the testing subset reference files with all previously mentioned methods in addition to the scaled phase-locking phase vocoder (SPL) of Laroche and Dolson (1999), IPL and SPL variants of PhaVoRIT ($\overline{\text{IPL}}$ and $\overline{\text{SPL}}$) of Karrer *et al.* (2006), and epoch synchronous overlap-add (ESOLA) of Rudresh *et al.* (2018). In the interval of $0.22 < \beta < 2.2$, 20 time-scale ratios were used, resulting in 5200 files with 400 files per method. During subjective testing, 42 529 ratings were collected from 263 participants in 633 sessions, resulting in a minimum of 7 ratings per file. Subjective median opinion scores (MedianOS) and subjective mean opinion scores (SMOS) before and after normalization were provided as labels. The dataset was published under the Creative Commons Attribution 4.0 International (CC BY 4.0, Mountain View, CA) license and is available online[1] (Roberts, 2020).

The International Telecommunications Union (ITU) Recommendation BS.1387, more commonly known as the PEAQ (Thiede *et al.*, 2000), is an OMOQ developed primarily for evaluation of audio codecs. It combines research from multiple groups and was released as an ITU standard in 2001. The PEAQ has two modes of operation, basic and advanced. The basic version (PEAQB) consists of a fast Fourier transform (FFT)-based peripheral ear model, preprocessing, calculation of 11 model output variables (MOVs), and a small neural network. The advanced version (PEAQA) follows the same framework but with a filter-bank-based ear model and five MOVs.

The FFT-based ear model aims to process the input signals in a way that is similar to the ear. The model contains a FFT, rectification, scaling of the input signal, outer and middle ear weighting, auditory filter bands, internal noise, frequency-domain spreading, and time-domain spreading. The filter-bank model is identical in aim and contains scaling of the input signals, direct current (DC) rejection, auditory filter band decomposition, outer and middle ear weighting, frequency-domain spreading, rectification, time-domain spreading, adding of internal noise, and additional time-domain spreading. Preprocessing of the resulting excitation patterns for both ear models creates patterns used in the calculation of the MOVs, the details of which can be found in ITU-T (2001a), Thiede *et al.* (2000), and Kabal *et al.* (2002).

The basic MOVs can be categorised into six groups. Modulation difference MOVs, WinModDiff1B, AvgModDiff1B, and AvgModDiff2B, are the windowed and linear averages of the modulation differences. Noise loudness MOVs, of which RmsNoiseLoudB is the only one used in the basic method, are the squared averages of the noise loudness and takes masking into account. Bandwidth MOVs, BandwidthRefB and BandwidthTestB, estimate the mean bandwidths of the reference and test signals, considering only frames with a bandwidth greater than 8 kHz. The psuedocode for the calculation is given in ITU-T (2001a). When considering auditory masking, Total NMRB, is the linear mean of the noise-to-mask ratio, whereas relative disturbed frames basic, RelDistFramesB, is the number of frames with a noise-to-mask ratio above 1.5 dB as a ratio of the number of frames for the signal. For detection probability, the maximum filtered probability of detection (MFPDB) models the smaller impact of distortions at the beginning of the file on quality assessment. The average distorted block (ADBB) uses the number of frames with a distortion detection probably above 0.5 and is calculated according to Sec. 4.7.2 in ITU-T (2001a). Finally, the harmonic structure of error (EHSB) MOV measures the harmonic structure of the error signal as strong harmonic structure may be transferred to the error signal. The advanced model uses EHSB and four additional MOVs. RmsModDiffA, RmsNoiseLoudAsymA, and AvgLinDistA are all calculated from the filterbank ear model excitation patterns while SegmentalNMRB is calculated from the FFT model. For full details, see ITU-T (2001a) and Kabal *et al.* (2002).

The PEAQ makes use of a neural network to map the MOVs to a single distortion index (DI) value. The network used with the basic model is a fully connected network with a single hidden layer of three nodes and sigmoid activation. Each feature is independently normalized to between zero and one before input to the network using

$$\hat{\text{MOV}} = \frac{\text{MOV} - \min(\text{MOV})}{\max(\text{MOV}) - \min(\text{MOV})}. \tag{1}$$

Finally, the DI is mapped to the final objective difference grade (ODG), minimizing the root mean square error (RMSE). The initial PEAQ standard (ITU-T, 2001a) was found to contain errors and to omit vital information required for a proper implementation of the standard. Kabal *et al.* (2002) clarified errors and omissions and provided a MATLAB implementation of the PEAQ-B portion of the standard.

Two quality measures for TSM have been proposed. Roucos and Wilgus (1985) used the SER, which is calculated by

$$\text{SER} = 10\,\log_{10} \frac{\displaystyle\sum_{u=0}^{U-1} \sum_{k=0}^{N/2} |X_T|^2}{\displaystyle\sum_{u=0}^{U-1} \sum_{k=0}^{N/2} (|X_R| - |X_T|)^2}, \tag{2}$$

where $X$ is shorthand for $X(u,k)$, $u$ is the frame number, $k$ is the frequency bin, $U$ is the total number of frames, $N$ is the FFT size, $X_R$ is the short time Fourier transform (STFT) of the reference signal, and $X_T$ is the STFT of the test signal. It is a measure of the difference between the magnitude spectra of the reference and test signals. Laroche and Dolson

(1999) proposed an objective phasiness measure ($D_M$) by determining the *a posteriori* consistency of the STFT synthesis reconstruction, and it is a measure of the horizontal and vertical phase coherences of the scaled signal. It is calculated by

$$D_M = \frac{\sum_{u=0}^{U-1}\sum_{k=0}^{N/2}\left(|X_T| - |X_R|\right)^2}{\sum_{u=0}^{U-1}\sum_{k=0}^{N/2}|X_R|^2}. \qquad (3)$$

Neither of these measures have seen continued use and each measure was noted to be only a high level indicator of the signal phasiness (Laroche and Dolson, 1999); however, they are beneficial to the performance of the proposed objective measure.

The paper is organized as follows. Section II presents the proposed OMOQ method. Section III presents feature and network results as well as a comparison of TSM algorithms. Availability, future research, and conclusions are presented in Secs. IV, V, and VI, respectively.

## II. METHOD

In this section, the proposed TSM objective measure is described. It uses a neural network to infer the SMOS score from handcrafted features computed from audio processed by TSM. A system block diagram can be seen in Fig. 1. Modifications to the PEAQ features are described in Sec. II A, additional features specific to TSM artefacts are described in Sec. II B, and the neural network is described in Sec. II C.

### A. Changes to the PEAQ

The PEAQ was chosen as the starting point for feature generation due to the high level of detail and specificity in the documentation for the measure. Changes were, however, made to allow for the use of signals of differing lengths, assuming a constant time-scale ratio was applied while processing the signal. Implementation of the PEAQ-B and PEAQ-A MOVs followed ITU-T (2001a) and referred to Kabal *et al.* (2002) in cases of ambiguity.

Signal preparation begins by summing all input channels before DC removal and normalization to the maximum absolute value. The proposed method uses a full scale as ±1 rather than the 16-bit integers of the PEAQ. A single channel is used in the proposed method as multichannel TSM is rarely considered (Roberts and Paliwal, 2018). Consequently, a single channel used for detection probability calculations in ITU-T (2001a) Sec. 4.7. Test and reference files are truncated to between the first and last time that the sum of the absolute of four consecutive samples exceeds 0.0061 as per ITU-T (2001a). This removes frames with low energy at the beginning and end of the signals during averaging calculations and synchronises the time-scaling starting point.

The PEAQ assumes an input sample rate of 48 kHz, however, the dataset used in this research has a sample rate of 44.1 kHz. Instead of resampling every file, the proposed method uses the bin frequency values of ITU-T (2001a). In

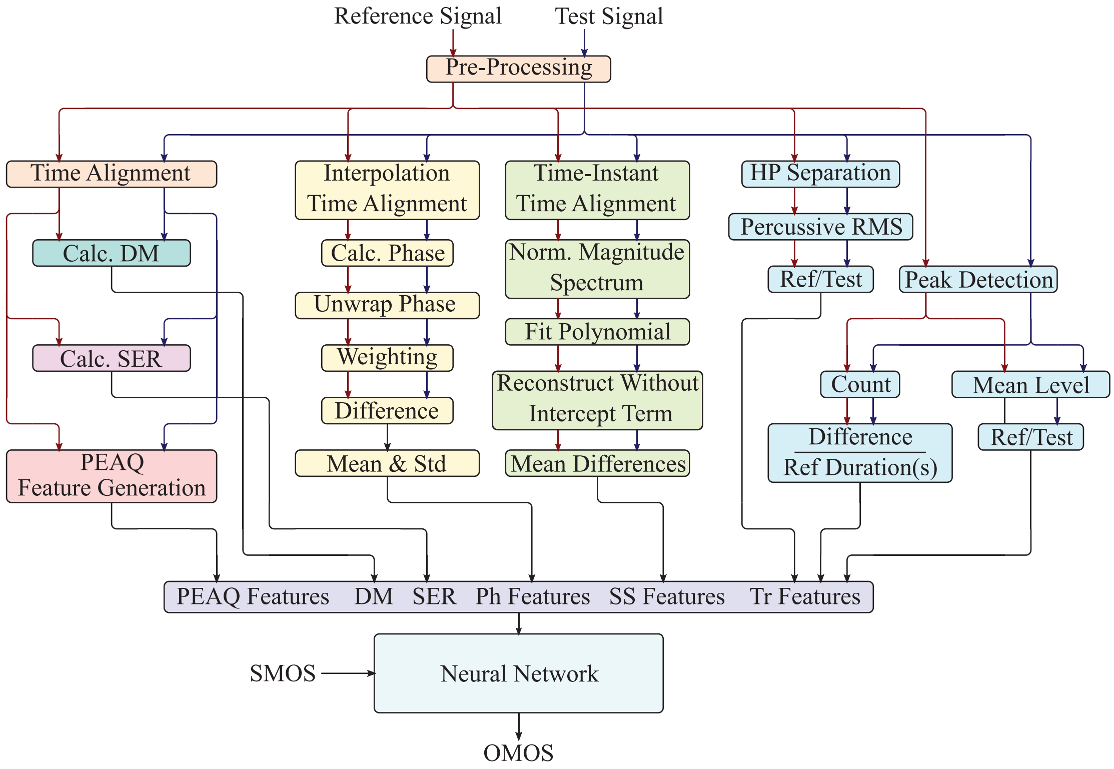

FIG. 1. (Color online) The OMOQ system block diagram. Features are colored by group with detail shown for novel features.

the calculation of BandwidthRefB and BandwidthTestB, the noise floor is calculated above 21 kHz with 8 kHz used as the bandwidth cutoff for bin inclusion during averaging. This increases the generality of the bandwidth feature generation across sample rates. The PEAQ and proposed method both assume that the frequency bandwidth will be reduced as a result of processing.

The reference signals before and after spectral adaptation are used as input for the *AvgLinDistA* calculation. However, the ITU specification is unclear as to which filter envelope modulation $\{\mathrm{Mod}[k,n]$ in Eq. (57)$\}$ to use in Eq. (67). The proposed implementation uses the reference modulation in the calculations of $s_{\mathrm{ref}}$ and $s_{\mathrm{test}}$ for Eq. (66) of ITU-T (2001a).

The final change to the ITU standard in the proposed method is the interpretation of "related to" in the calculation of RelDistFramesB. The proposed method uses the interpretation of Kabal *et al.* (2002) as meaning that the fraction of frames exceeds 1.5 dB.

Six methods of alignment were investigated during development, time-instant framing anchored to the reference or test signal and four methods of interpolating the magnitude spectrum frequency bins along the time-axis. Time-instant framing extracts frames from the reference and test signals at identical time-instants by scaling the frame locations by $\beta$ such that $S_R = u\beta S_T$ where $u$ is the frame number, $S_R$ is the reference signal shift in samples, and $S_T$ is the test signal shift in samples. In cases where $\beta$ is not known, the ratio between the lengths of the truncated input signals is used.

Although the alignment through time-domain resampling is not suitable because of the resulting changes in pitch, it is possible to resample the magnitude spectrum. This causes a change in the time-axis evolution of the signal without changing the positioning of the signal on the frequency axis and is similar to the filterbank TSM method proposed by Sharma *et al.* (2017). In the proposed method, interpolation for basic PEAQ features is applied prior to the ear model using one of four targets: the longest signal, the shortest signal, the reference signal, or the test signal. For advanced PEAQ features, interpolation to the test or reference duration is applied after application of the ear model. If the time scale is unknown, it is estimated by assuming a constant time-scale ratio. Through a simple thought experiment, we can observe that as we scale signals through interpolation, the transient components of the signal will also be scaled, whereas the same transients will not be scaled through time-instant framing. As such, it is necessary to consider all and combinations of the alignment methods.

### B. Additional features

When calculating the PEAQ bandwidth features, asymmetric thresholds are used with +10 dB used for *BandwidthRefB* and +5 dB used for *BandwidthTestB*. The test bandwidth, calculated with a +10 dB threshold (*BandwidthTestNew*), has been included as an additional feature.

The two traditional TSM OMOQs were included as features in the proposed method. SER was bounded to a maximum of 80 to avoid possible infinite results when processing identical files. This empirical value was the maximum finite feature value for identical files.

One cause of phasiness is phase unwrapping errors that occur when the time-scaling parameter ($\alpha = 1/\beta$) is not an integer (Laroche and Dolson, 1999). In this work, we propose a method for estimating the level of phasiness by considering the phase progression of the reference and test signals. The proposed phasiness features track phase progression through time for the reference and test tracks, accounts for the change of the time scale, and calculates the difference between the resulting unwrapped phase progression. Weighting is applied to the phase difference with unity and magnitude spectrum weighting applied in separate features within the proposed method. Weighting restricts phasiness to audible portions of the signal. These features are calculated in the following manner where $\angle$ denotes the arctan2 calculation. The phase spectra of the reference and test signals are calculated from the STFT and adjusted to be between 0 and $2\pi$ using

$$\angle\hat{X} = \begin{cases} \angle X, & \angle X > 0, \\ \angle X + 2\pi, & \text{otherwise}, \end{cases} \tag{4}$$

forming $\angle\hat{X}$. $2\pi$ is then successively added to each bin until it is greater than the same frequency bin in the previous frame using

$$\acute{X} = \min(\angle\hat{X} + 2\pi P) > \angle\hat{X}(u-1,k), \tag{5}$$

where $P \in \mathbb{Z}$. The longer $\acute{X}$ signal is then resampled to match the length of the shorter signal, forming $\tilde{X}$. The weighted angle difference ($\Delta\varphi$) can then be calculated using

$$\Delta\varphi = \begin{cases} W(k)(\acute{X}_R - \beta\tilde{X}_T), & U_T \geq U_R, \\ W(k)(\beta\tilde{X}_R - \acute{X}_T), & \text{otherwise}, \end{cases} \tag{6}$$

where weighting is calculated with

$$W(k) = \begin{cases} \dfrac{|X_R|}{\max|X_R|}, & U_T \geq U_R, \\ \dfrac{|X_T|}{\max|X_T|}, & \text{otherwise}, \end{cases} \tag{7}$$

or $W(k) = 1$ for no weighting, where $U_R$ and $U_T$ are the total number of frames in the reference and test signals, respectively. The time and frequency means of the angle differences form *MPhNW* for no weighting and form *MPhMW* for magnitude weighting. Similarly, the standard deviation of the frequency mean of the absolute weighted difference forms *SPhNW* and *SPhMW*. A number of additional measures were explored, including the power spectrum weighting, Fletcher-Munson curve weighting, and the mean first difference along the time dimension, however, they were found to be poor measures or contribute little toward the network training.

Figure 2 shows the phasiness features compared to both the SMOS and TSM ratio. Phasiness can be seen to increase as the TSM ratio moves away from 100% and as the SMOS decreases as expected. Animated three-dimensional plots rotating between features as functions of the SMOS and $\beta$, color coded to each TSM method can be found online.[2].

Phasiness causes spectral coloration of the signal (Laroche and Dolson, 1999), allowing for spectral similarity to be used as an indicator of the phasiness. Two features [spectral similarity mean absolute difference (SSMAD) and spectral similarity mean difference (SSMD)] were developed using the differences in the smoothed spectrum between the reference and test signals. Frames, aligned using reference frame anchors, are converted to normalized magnitude spectra using the STFT and Hann windowing. Third-order polynomials are then fit to the spectra. The resulting polynomials without the intercept term are applied to a linearly spaced vector $N/2$ in length. Removal of the intercept term removes any overall level difference between the frames. The mean absolute difference and mean difference between the reference and test signals are calculated for each frame with the means of these values forming the

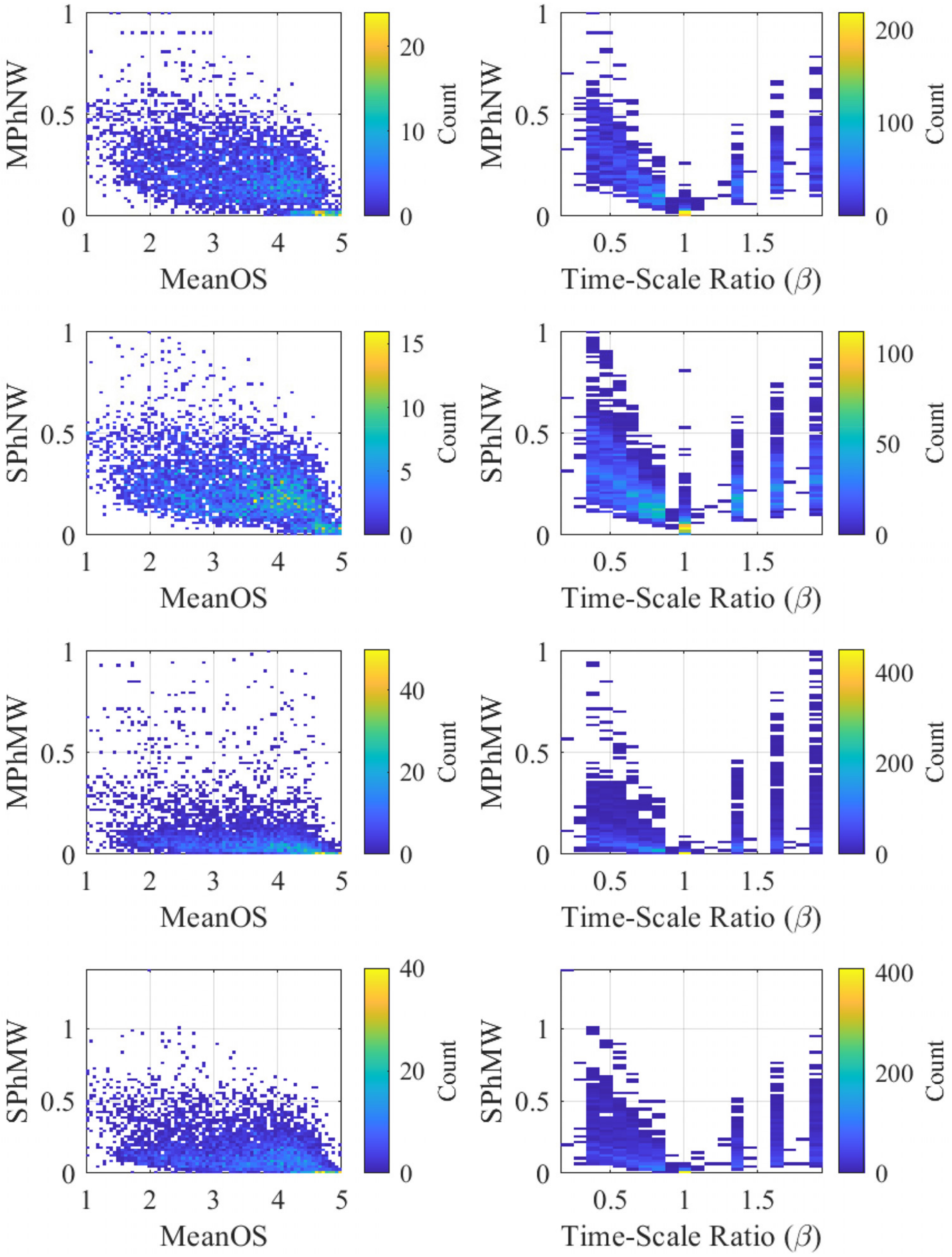

FIG. 2. (Color online) Phasiness features as functions of SMOS and the TSM ratio. The means and standard deviations for no weighting and magnitude weighting.

two spectral similarity features. These features also give a measure of signal coloration that is introduced by the TSM algorithm. Figure 3 shows the spectral similarity features in relation to the SMOS and TSM ratio. Further analysis found groupings for individual and classes of TSM methods within the features. Time-domain methods inherently introduce less or no phasiness by avoiding the phase unwrapping and vertical phase coherence problems of the frequency-domain methods, and FESOLA and WSOLA tend to have better spectral similarity than do the frequency-domain methods.

Changes in the transient content of the signal are common TSM artefacts. Three features have been developed for the proposed method, peak delta (ΔP), transient ratio, and harmonic percussive separation transient ratio, with no requirement for alignment between signals. ΔP is the difference in the number of onsets between the reference and test signals per second. Onset detection is applied to both signals using the spectral features method described by Bello *et al.* (2005). A weighting function, $W[k] = |k|$, is applied to the power spectrum using

$$\tilde{E}[u] = \sum_{k=0}^{N/2-1} W[k]|X|^2 \tag{8}$$

to suppress low frequency content and produce sharp peaks at transients before the first backward difference of the logarithmic transform is calculated using

$$\Delta\tilde{E}[u] = \log_{10}\tilde{E}[u] - \log_{10}\tilde{E}[u-1]. \tag{9}$$

Peak picking is applied to the onset results in which we define a peak as greater than its four surrounding values with

$$\mathrm{P}[u] = \begin{cases} 1, & \Delta\tilde{E}[u] > \Delta\tilde{E}[u-2:u+2], \\ 0, & \text{otherwise}. \end{cases} \tag{10}$$

No threshold for peak detection is used as it is expected that spurious peaks should exist in both the reference and test signals. Finally, the difference in the number of peaks per second, calculated using

$$\Delta P = \frac{f_s}{\dim(x_R)}\left(\sum P_T[u] - \sum P_R[u]\right), \tag{11}$$

is used as the feature, where $f_s$ is the sampling frequency and $\dim(x_R)$ is the length of the reference signal in samples.

The transient ratio (TrRat) is a measure of the change in the transient level caused by processing and makes use of the peak locations calculated previously in Eq. (10). It is calculated by selecting peaks where the onset peak level is greater than one standard deviation above the mean onset level $(\overline{\Delta\tilde{E}} + \sigma_{\Delta\tilde{E}})$, resulting in a vector of onset peak locations $(\hat{P})$. Peak values are then used to calculate the ratio of the mean transient levels between the reference and test signals using

$$\mathrm{TrRat} = \frac{\mathrm{mean}(\Delta\tilde{E}_R[\hat{P}])}{\mathrm{mean}(\Delta\tilde{E}_T[\hat{P}])}. \tag{12}$$

The harmonic percussive separation transient ratio (HPSTrRat) compares the root mean square (RMS) levels of the reference and test transients. The transients are extracted from the reference and test signals using the median filtering method of Driedger *et al.* (2014). The RMS levels of the extracted signals are calculated before the final feature is

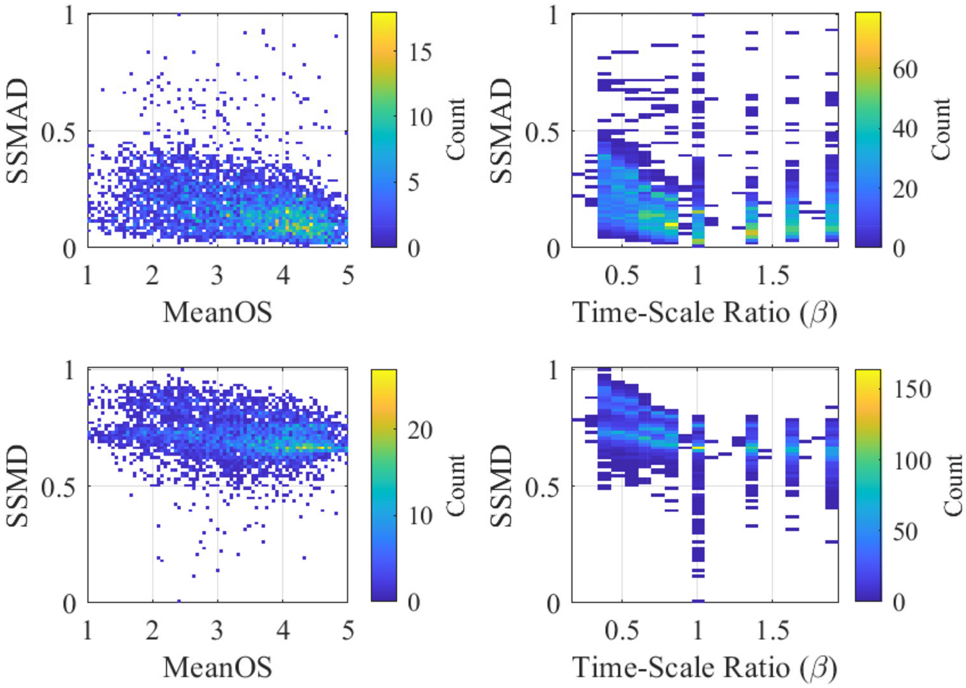

FIG. 3. (Color online) The SSMAD and SSMD features as functions of SMOS and the TSM ratio.

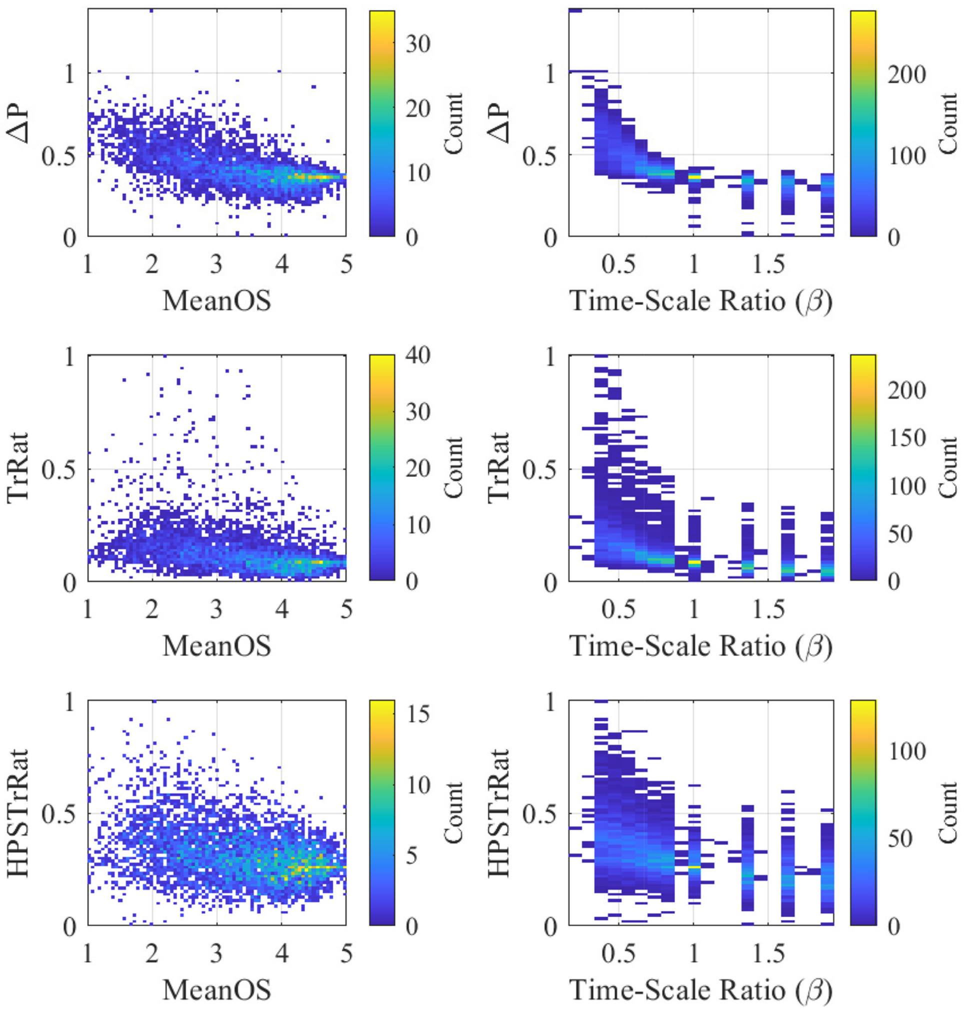

FIG. 4. (Color online) Transient features as functions of SMOS and the TSM ratio.

computed by the ratio of the reference to test. Figure 4 compares each of the transient features to the SMOS and TSM ratio.

Musical noise is a known artefact introduced by the frequency-domain TSM, likely caused be periodicity introduced to noise components of the signal due to the sum-of-sines model of the STFT. This results in holes and/or peaks in the power spectrum that are heard as musical noise (Torcoli, 2019) and was explored as a possible feature. Spectral kurtosis, as proposed by Torcoli (2019), was explored using all previously discussed methods of alignment. Lower, middle, and upper frequency bands were used in addition to the maximum across all bands. As all time-alignment methods produced highly correlated results, interpolation to test was chosen as the alignment method. However, inclusion of these features reduced neural network performance and as a result, they were removed from the features used in the final proposed network. This is likely due to the subtlety of the musical noise in comparison to other TSM artefacts such as phasiness and transient smearing.

Prior to network training, features and target SMOS scores were scaled to the interval [0,1].

### C. Network structure

The estimation of the opinion scores was formulated as a regression problem using a fully connected neural network with 3 hidden layers of 128 output nodes as shown in Fig. 5. Layer normalization and Rectified Linear Unit (ReLU) activation were used with residual connections around the second and third layers, facilitated by adding the input of a layer to its output. Sigmoid activation is applied to the final output. The network has 36 737 trainable parameters.

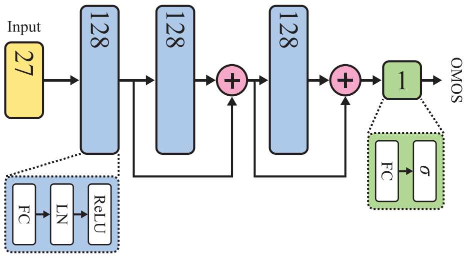

FIG. 5. (Color online) The neural network of the proposed measure. The numbers denote the number of layer output nodes, FC is a fully connected layer, LN is layer normalization, ReLU denotes the activation function, ⊕ is the element-wise summation of the layer input and output values, and $\sigma$ denotes a sigmoid activation layer.

Ten percent of the training dataset was reserved for validation. The network was trained for 800 epochs using a single batch, RMSE loss ($\mathcal{L}$), AdamW optimization (Loshchilov and Hutter, 2017), and a learning rate of $1e^{-4}$. Networks that were still improving after 800 epochs were trained for an additional 800 epochs. Internal loss values were calculated using estimates in the interval of [0,1], whereas reported loss values were calculated using estimates scaled back to the original interval of [1,5]. The Pearson Correlation Coefficient (PCC; $\rho$) and $\mathcal{L}$ were used as network performance measures. The composite measure of Roberts and Paliwal (2020) was used when selecting the ideal epoch after training. The optimal epoch was chosen as the epoch with the minimum overall distance ($\mathcal{D}$), calculated by

$$\mathcal{D} = \sqrt{\hat{\rho}^2 + \hat{\mathcal{L}}^2}, \tag{13}$$

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

$$\hat{\rho} = \sqrt{(1 - \bar{\boldsymbol{\rho}})^2 + \Delta\boldsymbol{\rho}^2}, \tag{14}$$

$$\hat{\mathcal{L}} = \sqrt{\bar{\mathcal{L}}^2 + \Delta\mathcal{L}^2}, \tag{15}$$

where $\boldsymbol{\rho} = [\rho_{\text{tr}}, \rho_{\text{val}}, \rho_{\text{te}}]$, $\mathcal{L} = [\mathcal{L}_{\text{tr}}, \mathcal{L}_{\text{val}}, \mathcal{L}_{\text{te}}]$, $[.,.]$ denotes concatenation, tr, val, and te denote training, validation, and testing, respectively, $\bar{\mathcal{L}}$ is the mean of $\mathcal{L}$, $\bar{\boldsymbol{\rho}}$ is the mean of $\boldsymbol{\rho}$, $\Delta\boldsymbol{\rho} = \max(\boldsymbol{\rho}) - \min(\boldsymbol{\rho})$, and $\Delta\mathcal{L} = \max(\mathcal{L}) - \min(\mathcal{L})$.

## III. RESULTS

### A. Feature results

An initial larger set of features was heuristically optimized based on changes in network performance and similarity to other features to reduce redundant features. If the $\rho$ between features of the same type was above approximately 0.95, one of the features was removed with Fig. 6 showing the correlation between each of the features in the proposed measure. This process increased the performance of the trained network. The features that were removed include the spectral kurtosis features, alternative weightings of phasiness features, and most standard deviations of time-domain features before averaging. Due to the nonlinear nature of the relationship between $\beta$ and the SMOS, absolute $\rho$ was calculated separately for $\beta < 1$ and $\beta > 1$ and then averaged. The novel TSM features were found to have a greater correlation to the SMOS than most of the PEAQ features. Of interest is the lack of individual features highly correlated with the SMOS or $\beta$ while still resulting in excellent network performance. Features were generated at approximately 400 files per hour using 16 threads on a Xeon E5–2630 (Intel, Santa Clara, California).

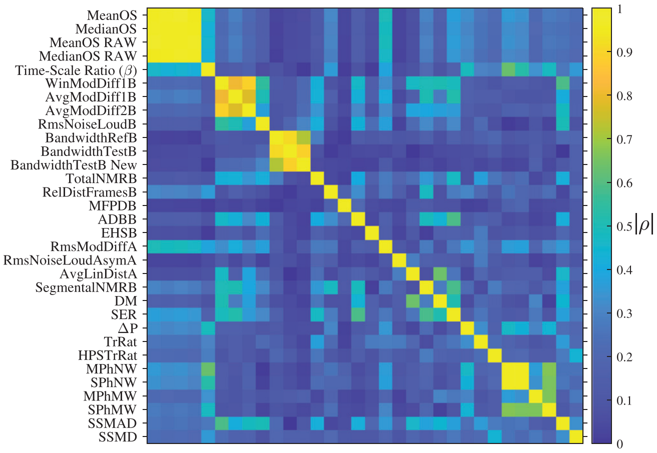

FIG. 6. (Color online) The feature correlation matrix for the final features. The absolute correlation averaged across $\beta < 1$ and $\beta > 1$ due to the non-monotonic nature of SMOS as a function of $\beta$ is shown.

### B. Network performance

A wide range of testing and network configurations were considered during the development of the proposed method. Network hyper-parameters were optimized through a systematic non-exhaustive search. Each method of alignment was trained to the SMOS, MedianOS, raw SMOS, and raw MedianOS targets, and raw values were calculated prior to subjective session normalization in Roberts and Paliwal (2020). Additionally, baseline conditions, the inclusion of reference files within the training set, concatenation of logarithmic transforms of features, and combinations of multiple alignment methods were considered. Deterministic training of the network was conducted using seeds from 0 to 99. Figure 7 shows the box plot distribution of the best $\mathcal{D}$ for each of the seed values used while training to the SMOS. Lower values are better with a smaller range meaning less reliance on the initial seed.

Across all cases, networks trained to mean, rather than median, targets processed better $\mathcal{L}$ and $\rho$ results. Consequently, the results discussed below will be solely focused on networks trained to mean targets. To increase

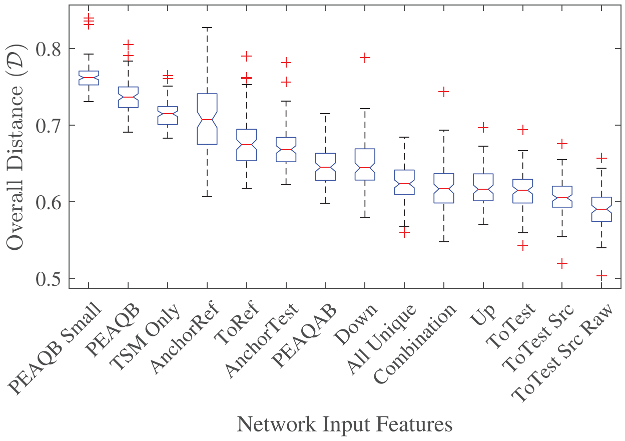

FIG. 7. (Color online) The box plot of the best distance measure for each seed and training configuration ordered by median $\mathcal{D}$ is shown. The PEAQB neural network (NN) uses the original PEAQ network, and all others use the network described in Sec. II C. Lower is better, and less spread means less reliance on the initial seed.

TABLE I. RMSE loss mean ($\bar{\mathcal{L}}$) and range ($\Delta\mathcal{L}$), PCC mean ($\bar{\rho}$) and range ($\Delta\rho$), median overall distance ($\tilde{\mathcal{D}}$), and minimum overall distance [min($\mathcal{D}$)]. Trained to SMOS unless specified. The best results appear in bold.

| Features (alignment) | $\bar{\mathcal{L}}$ | $\Delta\mathcal{L}$ | $\bar{\rho}$ | $\Delta\rho$ | $\tilde{\mathcal{D}}$ | min($\mathcal{D}$) |
|---|---|---|---|---|---|---|
| Original PEAQB (to test) | 0.668 | 0.054 | 0.719 | 0.075 | 0.762 | 0.731 |
| PEAQB (to test) | 0.636 | 0.104 | 0.753 | 0.028 | 0.737 | 0.691 |
| TSM only (to test) | 0.630 | 0.115 | 0.764 | 0.026 | 0.715 | 0.683 |
| All (anchor reference) | 0.540 | 0.205 | 0.834 | 0.086 | 0.707 | 0.607 |
| All (to reference) | 0.549 | 0.203 | 0.827 | 0.093 | 0.675 | 0.617 |
| All (anchor test) | 0.524 | 0.268 | 0.842 | 0.124 | 0.668 | 0.622 |
| PEAQAB (to test) | 0.558 | 0.109 | 0.820 | 0.043 | 0.645 | 0.598 |
| All (to shorter) | 0.543 | 0.120 | 0.836 | **0.024** | 0.644 | 0.580 |
| All unique (all alignments) | 0.524 | 0.117 | 0.844 | 0.036 | 0.623 | 0.560 |
| Combination (to test and anchor test) | 0.477 | 0.221 | **0.873** | 0.085 | 0.617 | 0.548 |
| All (to longer) | 0.534 | 0.109 | 0.834 | 0.030 | 0.616 | 0.571 |
| All (to test) | 0.500 | 0.150 | 0.860 | 0.050 | 0.615 | 0.543 |
| All (to test including reference) | 0.490 | 0.101 | 0.864 | 0.030 | 0.605 | 0.519 |
| All (to test including reference; SMOS raw) | **0.474** | **0.089** | 0.859 | 0.028 | **0.590** | **0.503** |

readability, median overall distance ($\tilde{\mathcal{D}}$) and the best case $\mathcal{D}$ with associated $\bar{\mathcal{L}}$, $\Delta\mathcal{L}$, $\bar{\boldsymbol{\rho}}$, and $\Delta\boldsymbol{\rho}$ values can be found in Table I. Values were calculated as per Sec. II C.

The baseline performance for the traditional methods was determined by correlation with the target. SER and $D_M$ gave overall $\rho$ with subjective scores of 0.3708 and 0.1574, respectively. The machine learning baseline performance was obtained by applying time-aligned PEAQB features to the original PEAQB network described by ITU-T (2001a), shown as "original PEAQB (to test)." By increasing the complexity of the network to that in Sec. II C, $\bar{\mathcal{L}}$ and $\bar{\boldsymbol{\rho}}$ improved, shown as "PEAQB (to test)." Performance was further improved through the inclusion of the PEAQA features, shown as "PEAQAB (to test)." Interpolating to the length of the test signal was found to give the best performance followed by, in order, interpolating up to the longer signal, down to the shorter signal, anchoring frame locations to the test signal, interpolating to the reference signal length, and anchoring frame locations to the reference signal. Using only the new TSM features gave improved performance over the PEAQB features. Combinations of features generated using interpolation to test and time-instant anchoring to test (combination) alignment were also applied to the network. This improved performance over individual alignments; however, network performance was highly reliant on the initial seed selection. All of the unique features were also combined and applied to a larger network with 512 nodes per layer but did not improve over previously tested feature sets. Combinations of concatenating logarithmic transforms of the features, including reference signals, and combining different alignment features were applied to the network but all resulted in reduced performance. The best overall network aligned signals using interpolation to the length of the test signals and included reference signals when training with the SMOS targets set to five. The loss and correlation for each epoch of the proposed network can be seen in Fig. 8.

Given that the network performance in predicting the raw SMOS outperforms the prediction of the normalized SMOS, investigation of the objective mean opinion score (OMOS) differences was undertaken. The mean difference between the normalized and raw SMOS was found to be −0.0023, whereas the mean difference was found to be 0.016 for the OMOS. Normalizing was found to slightly extend the range of the SMOS values with higher ratings for high quality files and lower ratings for low quality files. Given the ITU-T (2019) recommendation of normalization, the final proposed OMOQ was trained using the normalized SMOS.

The proposed network achieved a mean PCC of 0.864 and a RMSE of 0.490 to the SMOS was trained to normalized SMOS using interpolation to test for alignment and included reference files within the training set. These results place the proposed network at the 82nd and 97th percentiles of the subjective sessions for the PCC and RMSE, respectively, resulting in a system that effectively predicts the mean opinion scores (MOS) for the signals with little or no consensus.

### C. TSM algorithm evaluation

TSM algorithms were compared using the evaluation subset, which is described in Sec. I. The uTVS implementation used in subjective testing ($\overline{\text{uTVS}}$), and an IPL by

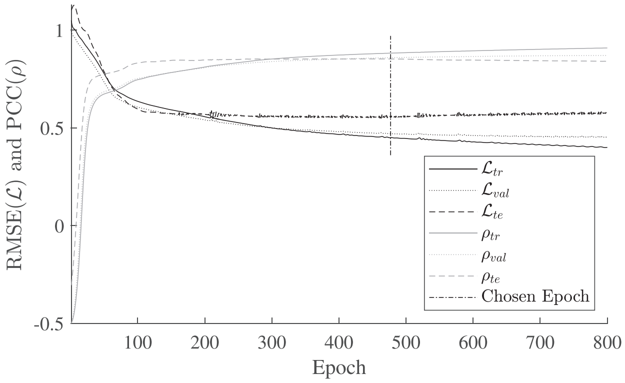

FIG. 8. The loss and correlation for training, validation, and test sets for each epoch. the best epoch is shown as a vertical line.

TABLE II. Mean objective mean opinion score (OMOS) for each class of file and overall result. Means are calculated without $\beta$ of 0.2257 and 1. Methods in order fromleft to right are NMFTSM, ESOLA, FESOLA, PV, FuzzyPV, Phavorit SPL, uTVS, subjective testing uTVS, Phavorit IPL, HPTSM, SPL, WSOLA, Elastique, Driedger's IPL, and IPL.

| | NMF | ES | FES | PV | FPV | $\overline{\text{SPL}}$ | uTVS | $\overline{\text{uTVS}}$ | $\overline{\text{IPL}}$ | HP | SPL | WS | EL | DIPL | IPL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Music | 2.931 | 2.782 | 2.987 | 3.572 | 3.663 | 3.678 | 3.591 | 3.615 | 3.715 | 3.597 | 3.672 | 3.538 | 3.721 | 3.771 | **3.835** |
| Solo | 2.932 | 3.635 | 3.652 | 3.463 | 3.399 | 3.586 | 3.628 | 3.630 | 3.663 | 3.673 | 3.646 | 3.792 | 3.796 | **3.850** | 3.773 |
| Voice | 2.987 | 3.302 | 3.412 | 3.052 | 3.160 | 3.201 | 3.320 | 3.317 | 3.212 | 3.415 | 3.457 | 3.507 | **3.660** | 3.596 | 3.621 |
| Overall | 2.948 | 3.194 | 3.314 | 3.383 | 3.433 | 3.507 | 3.521 | 3.530 | 3.548 | 3.565 | 3.600 | 3.605 | 3.725 | 3.742 | **3.752** |

Driedger and Muller (2014; DIPL) have also been included. Although $\beta = 1$ was used in the evaluation, in practice, time-scaling is only applied at ratios other than one. Additionally, $\beta = 0.25$ was the minimum available for Elastique using SonicApi (GmbH & Co., Berlin, Germany). Consequently, all results for $\beta = 1$ and $\beta < 0.25$ were excluded from averaging calculations. Table II shows the mean OMOS for each of the TSM methods tested in addition to means for each file class ordered by the ascending overall mean.

The analysis is split into each class of the reference files followed by the overall average results. The poor performance of the uTVS subjective testing implementation for $\beta$ close to one is also visible with the updated implementation showing monotonic improvement toward $\beta = 1$. The noisy nature of the results for Elastique, FuzzyPV, and NMFTSM in Roberts and Paliwal (2020) has been smoothed with all of the results following those of Roberts and Paliwal (2020).

For musical files, the OMOQ effectively differentiates between the frequency and time-domain methods where the quality worsens faster for the time-domain methods. WSOLA fares the best out of the time-domain methods, diverging from the frequency-domain methods for $\beta < 0.8$ as shown in Fig. 9(a). When averaged, the OMOQ rates the IPL highest, followed by Elastique. All other frequency-domain methods gave similar results.

For solo files, all methods except the NMFTSM perform similarly with a maximum difference between methods of 0.576 for $\beta = 0.87$. The method means at each time scale can be seen in Fig. 9(b). DIPL has the highest mean OMOS, followed by Elastique, WSOLA, and IPL as shown in Table II. The strong performance of WSOLA is expected due to individual harmonic and percussive signals.

The voice file OMOS shows the greatest variance between methods. Of interest is the exponential shape of the curve for $\beta < 1$ compared to the logarithmic shape for

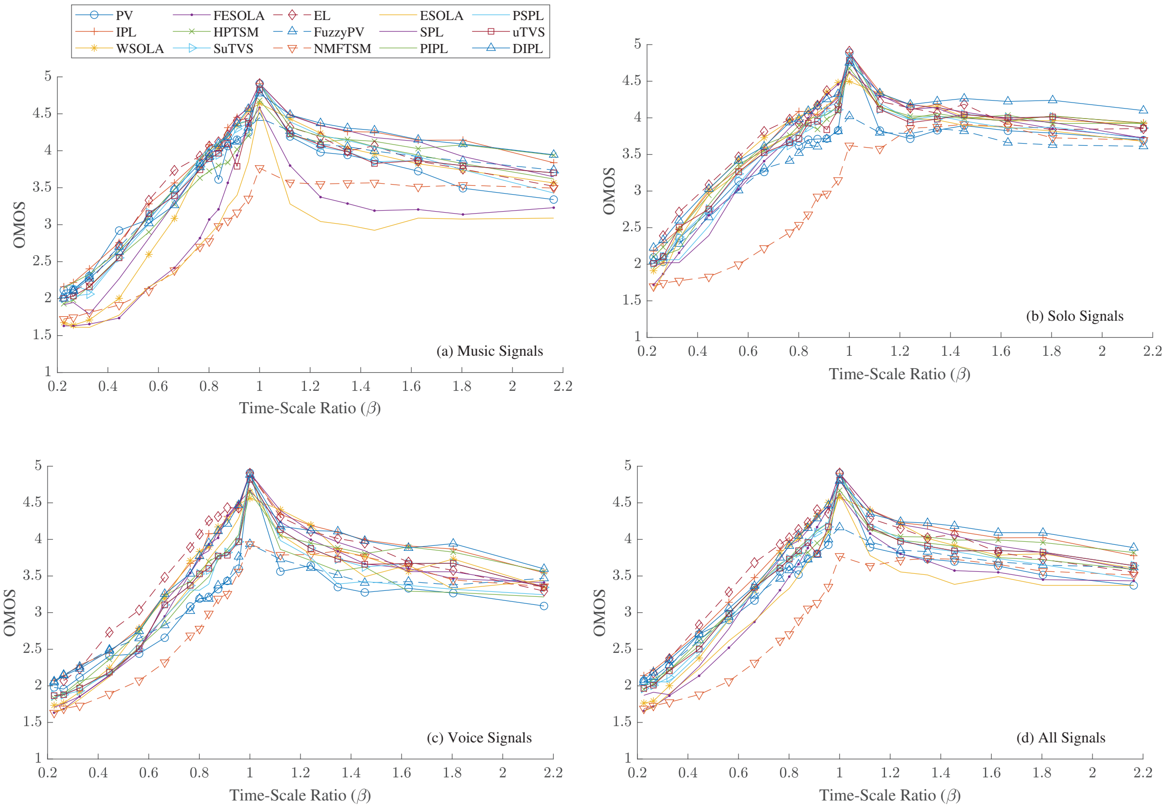

FIG. 9. (Color online) The mean OMOS for each TSM method as a function of $\beta$ for (a) musical signals, (b) solo signals, (c) voice signals, and (d) all signals combined. Each point is the average OMOS for a subset of files processed by one TSM method at one time-scale ratio. Higher is better.

musical and solo classes, indicating that harsher subjective evaluation of the voice files was learned by the network. The method means at each time scale can be seen in Fig. 9(c). Elastique has the highest mean OMOS, followed by IPL, DIPL, and WSOLA. ESOLA and FESOLA give improved performances for this class relative to other methods.

By averaging all OMOS, the IPL has the highest average rating, followed by DIPL and Elastique, separated by only 0.03 OMOS. Only 0.098 separates WSOLA through SPL. The overall low performance of FuzzyPV is unexpected given that it builds on the IPL. However, other methods that perform decomposition of the signal, such as NMFTSM and HPTSM, also perform below the methods they build upon, suggesting that simpler artefacts are preferred over those introduced by multiple processing methods. The overall means can be seen in Fig. 9(d). Two-sample $t$-test analysis ($\alpha = 0.05$) of all of the OMOS shows the null hypothesis of equal means to be rejected in almost all of the cases when the absolute difference of mean OMOS is greater than 0.098. ESOLA and FESOLA are the only exceptions with an absolute difference of 0.1201 and $P$-values of 0.069.

## IV. AVAILABILITY

The proposed tool is available online.[3] This includes the MATLAB scripts for the feature generation, PyTorch code feature evaluation, and features for all dataset files in "csv" and ".mat" formats. A bash script is also included, which creates a virtual environment and installs required modules. The features are also available with the subjective dataset online.[4]

## V. FUTURE RESEARCH

Future research is multifaceted. First, evaluation of a wide range of commercial and lesser known published TSM methods should be considered in addition to comparisons of different implementations of the same TSM method. Second, expansion into alternative and deeper neural networks should also be considered. Initial testing resulted in a $\rho_{te}$ of 0.71 for a random forest network using the handcrafted features, whereas using blind data-driven features created by a convolutional neural network (CNN) used as input to a fully connected network resulted in a $\rho_{te}$ of 0.65.

## VI. CONCLUSION

An objective measure for time-scaled audio was proposed with the performance superior to most subjective listeners. The measure used handcrafted features and a fully connected network to predict the SMOS. The PEAQB and PEAQAd features were used in addition to nine novel features specific to TSM artefacts. Six methods of alignment were explored with interpolation of the magnitude spectrum to the duration of the test signal giving the best performance, achieving a mean RMSE of 0.490 and a mean PCC of 0.864. Using the proposed method to evaluate algorithms, it was found that Elastique gave the highest objective quality for voice signals while the IPL variants gave the highest objective quality for music and solo instrument signals, as well as the best overall performance. Future work includes optimization of feature generation, exploration of other network structures, and evaluation of additional TSM algorithms.

[1] See http://ieee-dataport.org/1987 the subjective dataset (Last viewed 7 March 2021).

[2] See https://zygurt.github.io/TSM/objective for the animated three-dimensional plots rotating between features (Last viewed 7 March 2021).

[3] See https://github.com/zygurt/TSM for the scripts for the feature generation.

[4] See http://ieee-dataport.org/1987 the subjective dataset (Last viewed 7 March 2021).

Avila, A. R., Gamper, H., Reddy, C., Cutler, R., Tashev, I., and Gehrke, J. (**2019**). "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, New York), pp. 631–635.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. B. (**2005**). "A tutorial on onset detection in music signals," IEEE Trans. Speech Audio Process. **13**(5), 1035–1047.

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 *Proc. of the 17th Int. 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.

Gomez, A. M., Schwerin, B., and Paliwal, K. (**2011**). "Objective intelligibility prediction of speech by combining correlation and distortion based techniques," in *Twelfth Annual Conference of the International Speech Communication Association*.

ITU-T. (**2001a**). "ITU-R BS. 1387-1: Method for objective measurements of perceived audio quality," Technical Report (International Telecommunications Union, Geneva, Switzerland).

ITU-T. (**2001b**). "ITU-R p.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Technical Report (International Telecommunications Union, Geneva, Switzerland).

ITU-T. (**2019**). "ITU-R BS. 1284-1: General methods for the subjective assessment of sound quality," Technical Report (International Telecommunications Union, Geneva, Switzerland).

Kabal, P. (**2002**). "An examination and interpretation of ITU-R BS. 1387: Perceptual evaluation of audio quality," TSP Lab Technical Report (Department of Electrical and Computer Engineering, McGill University, Montreal, Canada), pp. 1–89, available at https://github.com/NikolajAndersson/PEAQ (Last viewed 7 March 2021).

Karrer, T., Lee, E., and Borchers, J. (**2006**). "PhaVoRIT: A phase vocoder for real-time interactive time-stretching," Technical Report (Aachen, Germany).

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

Loshchilov, I., and Hutter, F. (**2017**). "Decoupled weight decay regularization," arXiv:1711.05101.

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

Roberts, T. (**2020**). "A time-scale modification dataset with subjective quality labels," available at http://dx.doi.org/10.21227/ny9p-rv41 (Last viewed 7 March 2021).

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

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

Roberts, T., and Paliwal, K. K. (**2020**). "A time-scale modification dataset with subjective quality labels," J. Acoust. Soc. Am. **148**(1), 201–210.

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

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

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

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 *2017 Twenty-Third National Conference on Communications (NCC)* (IEEE, New York), pp. 1–5.

Thiede, T., Treurniet, W., Bitto, R., Schmidmer, C., Sporer, T., Beerends, J., Colomes, C., Kehyl, M., Stoll, G., Brandenburg, K., and Feiten, B. (**2000**). "Peaq-the ITU standard for objective measurement of perceived audio quality," J. Audio Eng. Soc. **48**(1/2), 3–29.

Torcoli, M. (**2019**). "An improved measure of musical noise based on spectral kurtosis," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, IEEE, New York), 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 *Proceedings of ICASSP '93* (IEEE, New York), Vol. 2, pp. 554–557.

Zplane Development. (**2019**). "Èlastique time stretching and pitch shifting sdks (version 3.2.5) [computer program]," available at http://licensing.z-plane.de/technology\#elastique (Last viewed October 31, 2019).