T. Roberts, A. Nicolson and K.K. Paliwal, "Deep Learning-Based Single-Ended Quality Prediction for Time-Scale Modified Audio"
J. Audio Eng. Soc., vol. 69, no. 9, pp. 644–655, (2021 September).
DOI: https://doi.org/10.17743/jaes.2021.0031

Deep Learning-Based Single-Ended Quality Prediction for Time-Scale Modified Audio

TIMOTHY ROBERTS,[1*] AARON NICOLSON,[2] AND KULDIP K. PALIWAL[1]
(timothy.roberts@griffithuni.edu.au) (aaron.nicolson@csiro.au) (k.paliwal@griffith.edu.au)

[1]*Griffith University, Nathan, Australia*
[2]*Australian eHealth Research Centre, CSIRO, Herston, Australia*

Objective evaluation of audio processed with Time-Scale Modification (TSM) has recently seen improvement with a labeled time-scaled audio dataset used to train an objective measure. This double-ended measure was an extension of Perceptual Evaluation of Audio Quality and required reference and test signals. In this paper two single-ended objective quality measures for time-scaled audio are proposed that do not require a reference signal. Internal representations of spectrogram and speech features are learned by either a Convolutional Neural Network (CNN) or a Bidirectional Gated Recurrent Unit (BGRU) network and fed to a fully connected network to predict Subjective Mean Opinion Scores. The proposed CNN and BGRU measures respectively achieve average Root Mean Square Errors of 0.61 and 0.58 and mean Pearson Correlation Coefficients of 0.77 and 0.79 to the time-scaled audio dataset. The proposed measures are used to evaluate TSM algorithms and comparisons are provided for 15 TSM implementations. A link to implementations of the objective measures is provided.

0 INTRODUCTION

Time-Scale Modification (TSM) aims to manipulate the temporal domain of a signal independent of pitch and timbre. The time-scale ratio (β) denotes time-expansion (slower playback) for $\beta < 1$ and time compression (faster playback) for $\beta > 1$. Subjective testing is undertaken in order to justify the quality of the processing. However the testing is expensive and time consuming. Recently [1] published a dataset of time-scaled signals with subjective evaluation labels and initial work toward an objective measure of quality. However this method requires reference and test signals and additional interpolation to align low-bandwidth representations of the signals. In this work we propose multiple single-ended objective measures of quality for audio processed with TSM and extend the use of deep learning-based MOS estimation for single-ended measures. A convolutional or recurrent neural network front-end generates data-driven features, while a Fully Connected Neural Network (FCNN) back-end predicts the overall quality. The measures are trained using the dataset of [1], with the dataset referred to as TSMDB from this point.

*Address correspondence to E-mail: timothy.roberts@griffithuni.edu.au

0.1 Quality Evaluation

Subjective evaluation, such as BS.1284 [2], is the gold standard for evaluating quality of speech and audio processing. Participants are asked to rate the processing quality of audio files, often using ratings of Bad, Poor, Fair, Good, and Excellent that map linearly to the interval [1,5]. Opinion scores are then averaged, giving a Mean Opinion Score (MOS) per file. However, this process is lengthy and expensive. Consequently many objective measures of quality have been proposed to predict MOS.

Objective measures can be classified into double-ended (DE; invasive) and single-ended (SE; non-invasive) methods. The former calculates differences between reference and processed signal pairs, whereas the latter operates solely on the processed signal. This allows non-invasive measures to be used in a variety of use cases such as testing of in-service real-time systems using multiple tests through a signal path, as in [3] and [4]. SE measures have seen considerable use for speech quality [4–10] with little use for general audio quality. These general audio measures, including [11–15], are DE. Additionally some SE measures have been trained to DE measures, including [16] and [17].

SE measures are often compared to baseline DE measures such as Perceptual Evaluation of Speech Quality [18] and Perceptual Evaluation of Audio Quality (PEAQ) [12].

However there is no standard for objective quality of TSM, with minimal published literature on the topic. The total published research is found in the following papers. [19] published preliminary work toward an objective measure by using linear regression of the Mean Square Error (MSE) of transient, tonal, and noise energy deviations to predict the Subjective Mean Opinion Score (SMOS). [1] and [20] proposed measures that use hand-crafted features, derived using novel and modified PEAQ [2] features, as input to an FCNN to predict SMOS.

The DE method of [1], referred to as OMOQDE from this point, aligned signals through time-axis interpolation of the reference magnitude spectrum to the processed magnitude spectrum. The resulting PEAQ features were used to retrain the PEAQ basic neural network. Formulated as a regression problem, an FCNN was used to predict the MOS targets of the TSMDB. OMOQDE achieved an average Pearson Correlation Coefficient (PCC; $\overline{\rho}$) of 0.719 and average Root Mean Square Error (RMSE) loss ($\overline{L}$) of 0.668 using the MOS range of 1–5 for the training, validation, and test sets. A distance measure that penalized over-fitting was used to select the ideal network and is discussed further in Sec. 1. These results were improved in [20] to $\overline{\rho}$ of 0.864 and an average RMSE $\overline{L}$ of 0.490 and were able to resolve statistically significant differences in mean quality between TSM methods of 0.1 MOS.

0.2 Time-Scale Modification Dataset

OMOQDE was trained using the TSMDB [1], which contains a training subset of 5,280 files and testing subset of 240 files. Six TSM methods at 10 β values were used to process 88 reference signals to create the training subset. Three additional methods at randomized β within four bands were used to process 20 additional reference signals to create the testing subset. This resulted in no overlap between training and testing subsets when considering TSM methods, β, or reference signals. The test set reference signals were also processed by 15 TSM methods at 20 β values between 0.2 and 2 to create an evaluation set used to compare the TSM methods. These methods are:

- Phase Vocoder (PV) [21],
- Identity Phase-Locking Phase Vocoder (IPL) and Scaled Phase-Locking Phase Vocoder (SPL) [22],
- Waveform Similarity Overlap Add (WS) [23],
- Fuzzy Epoch Synchronous Overlap-Add (FES) [24],
- Harmonic Percussive Separation Time-Scale Modification (HP) [25],
- Mel-Scale Sub-Band Modeling (uTVS) [26] and the version used in subjective testing (uTVS),
- Elastique (EL) [27],
- Phase Vocoder using fuzzy classification of bins (FPV) [28],
- Non-Negative Matrix Factorization Time-Scale Modification (NMF) [29],
- PhaVoRIT (IPL and SPL) [30],
- Epoch Synchronous Overlap-Add (ES) [31], and
- IPL implementation of [25] (DIPL).

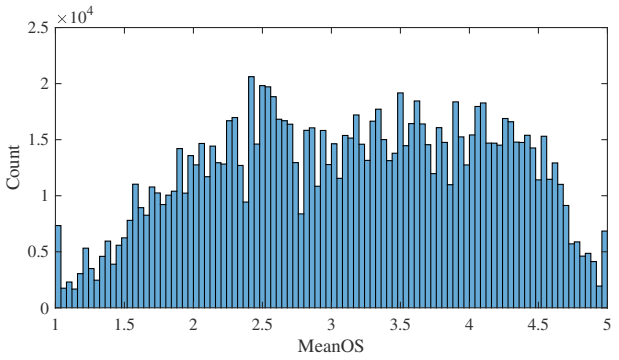

Fig. 1. Distribution of frames per Mean Opinion Score (MeanOS) in training set.

These were chosen to cover most TSM approaches. Time-domain, frequency-domain, and source separation methods are represented as well as filterbank and non-negative matrix factorization methods. TSMDB quality labels were provided as MOS and median opinion scores calculated before and after session normalization in the interval [1,5]. The scores were collated from 42,529 ratings by 263 participants in 633 sessions, with a minimum of seven ratings per file. All files in the dataset have a single channel, sampling rate of 44.1 kHz, and bit depth of 16 bits. Some reference files are stereo and were converted to a single channel by summation and normalization to the interval [–1,1].

The files of the TSMDB were split into training, validation, and testing subsets as per [1] and [20]. The first six TSM methods in the list above were used for training and validation, while the next three methods were used for testing, resulting in no overlap of TSM method, source file, or β. 10% of the training subset was randomly selected to be used as the validation set. The evaluation set uses the training subset source files processed by all of the TSM methods listed above with a unique set of β values [20]. As TSM quality changes based on β and signal source [1], the aim of the TSMDB therefore is to cover a wide variety of signal sources and β for a variety of TSM methods.

Objective measures have historically been difficult to develop because of the lack of a large enough set of files with subjective ratings. Within the recently published dataset, the number of frames for $\beta < 1$ in the training subset is significantly greater than for $\beta > 1$; however the number frames is relatively uniform for $2 \leq SMOS \leq 4.5$, see Fig. 1. The reduced number of frames for $1 \leq SMOS \leq 1.5$ and $4.5 \leq SMOS \leq 5$ may impact the estimation at the extremes of the Objective Mean Opinion Score (OMOS) range as there is less data for network training. Additionally the difference in time-base between original and processed signals has meant traditional objective measures cannot be applied. Deep learning is allowing for non-intrusive measures of quality, bypassing the need for identical time-bases between original and processed signals.

0.3 Deep Learning and Features

Deep learning is often used in objective measures of quality. Convolutional Neural Networks (CNNs) [32] are

commonly used on spatial domain tasks, such as image classification. They have also found use in speech and audio because of the spatio-temporal representation of short-time frequency analysis, as in [11]. CNNs learn weights of convolutional kernels that are applied successively, creating higher-order representations of the signal. Recurrent Neural Networks (RNNs) differ from standard fully connected networks through the inclusion of a memory cell and suitability to time-series data. In this paper Long Short-Term Memory (LSTM) [33] and Gated Recurrent Units (GRU) [34] are the used cell types. LSTM cells are controlled by three gates (input, output, and forget) that determine what information is added to or removed from the cell. GRU is a variant of LSTM that removes the output gate and has fewer parameters. Bidirectional Recurrent Neural Networks [35], such as the Bidirectional Long Short-Term Memory (BLSTM) and Bidirectional Gated Recurrent Units (BGRU), extend RNNs with forward and backward passes over the time-series.

Introduced by [36], Mel Frequency Cepstral Coefficients (MFCCs) have found extensive use as a lower-bandwidth transformed signal representation in speech processing, as in [37]. MFCCs are computed by first estimating the periodogram of the short-time power spectrum. A bank of triangular-shaped filters spaced uniformly on the Mel-Scale is then applied, resulting in the energy of each filter. The logarithm of the filterbank energies is then taken, followed by a Discrete Cosine Transform to decorrelate the filterbank energies. Differential and acceleration coefficients are often used to give an indication of the dynamics of the MFCCs and are generally known as Deltas (D) and Delta-Deltas (D′).

The contributions of this paper include the application of CNN and BGRU networks in the context of non-intrusive quality assessment of general audio. It also presents the first SE quality measures for time-scaled audio as well as the effectiveness of different input features such as magnitude, phase, MFCCs, and Deltas for this task. Finally the paper makes general inference of the relative quality of TSM methods in three different classes of signals: music, solo instruments, and voice. The paper is organized as follows: Sec. 1 presents the proposed Single Ended Objective Measure of Quality (OMOQSE) methods; Sec. 2 presents network results as well as a comparison of TSM algorithms. Availability, future research, and conclusions are presented in Secs. 3, 4, and 5, respectively.

1 METHOD

First we describe the audio processing. Signals were prepared by normalizing to the interval [–1,1] and trimming silence at the beginning and end of the signal. Silence was determined, according to [38], as the first and last time the sum of four consecutive samples is greater than 0.0061. The magnitude spectrum ($|X|$), magnitude and phase spectra ($[|X|;\angle X]$), power spectrum ($|X|^2$), MFCCs, MFCCs and D ([MFCCs;D]), and MFCCs, D and D′ ([MFCCs;D;D′]), where $[\cdot\ ;\ \cdot]$ is concatenation, were tested during development. The magnitude, phase, and power spectra used a frame length of $N = 2{,}048$ samples, overlap of N/2, and Hann window. MFCCs were of length 128, with D and D′ width 9 from $t - 4$ to $t + 4$ with respect to the current time-step. Overall or per frequency-bin standardization of the input features was explored.

Because of the variable length of the input signal, truncating and duplicating the signal were explored. For the CNN, sequences were truncated to the overall minimum length (L), starting from a different random location in each epoch. During testing the OMOS was averaged over 16 segments to capture more information of the processed signal for a wider sampling of the signal. An alternative method of pooling could be considered in future work. Repeating the input signal to the duration of the longest signal was also considered for RNNs; however as LSTM and GRU operate sequentially on each frame, input signals were used in their entirety. An attention mechanism could be used in the future to improve the loss function.

Prior to network training, target scores were scaled to the interval [0,1] using

$$SMOS \leftarrow \frac{SMOS - 1}{4}. \tag{1}$$

1.1 Network Structure

The proposed CNN data structure, shown in Fig. 2, was based on that of [11]. It contains four convolution layers, of filter sizes 16, 32, 64, and 32, with batch normalization and a 5 x 5 kernel for the first layer and 3 x 3 for the remaining layers. The first two convolutional layers are followed by max pooling layers, with 2 x 2 kernels and 2 x 2 stride. After concatenation and 10% dropout, three fully connected layers of output size 128 are used. The final layer has a single output. This results in 821,857 trainable parameters. Rectified linear unit (ReLU) activation is used throughout except for the output layer where the Sigmoid activation is used. Residual connections around the second and third fully connected layers are used. RMSE is used as the loss function. Features were concatenated in time-aligned input panes.

The proposed final-frame (FF) model for LSTM, BLSTM, GRU, and BGRU networks can be seen in Fig. 3. FF RNN models use backpropagation through time to learn from the error between the final output and SMOS. The total feature dimension (D_F) is set by the concatenation of input features. For the proposed network using [MFCCs;D] features, D_F is 256. Two RNN layers were used with the memory layer size (D_H) set to the number of directions (n) multiplied by D_F. L is the sequence length and ranged from 53 to 2,179 frames. An RNN architecture of many-to-one was used, with the final frame used as input to an FCNN after 10% dropout. The FCNN contained 3 layers of output size 256, 128, and 1, respectively. Layer normalization and ReLU activation were used for layers 1 and 2, while Sigmoid activation was used for the output layer. This results in 16,370,945 trainable parameters. Again RMSE is used as the loss function. Magnitude, phase, and power spectra ($D_F = 1025$, $D_H = 512$) were also explored as input to this network.

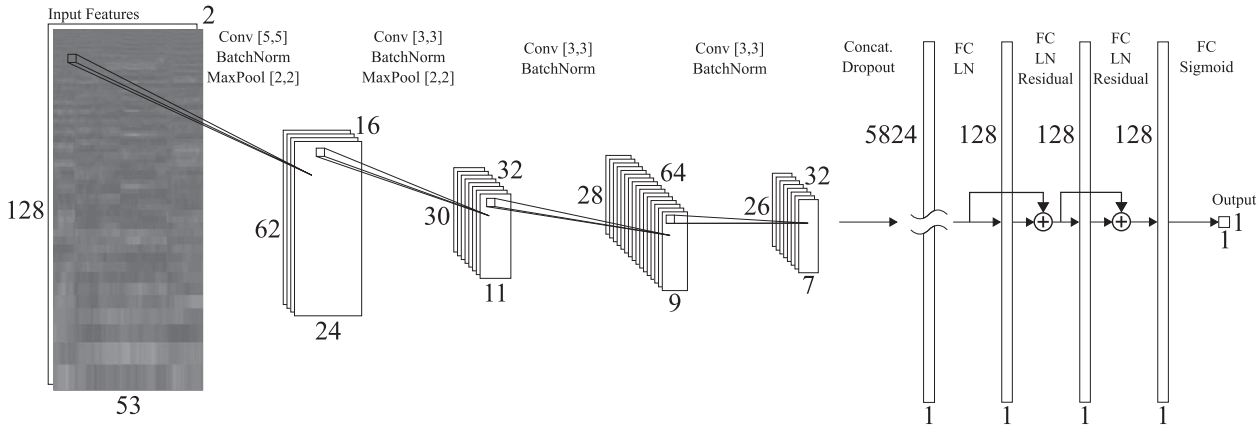

Fig. 2. Proposed CNN dataflow. Kernel sizes in brackets, numbers denote layer size and number of channels, FC is a fully connected layer, LN is layer normalization, and ReLU activation used unless specified. Conv[x,y] is Convolution[Filtersize], BatchNorm is Batch Normalization, MaxPool is Max Pooling, and Concat. is Concatenation.

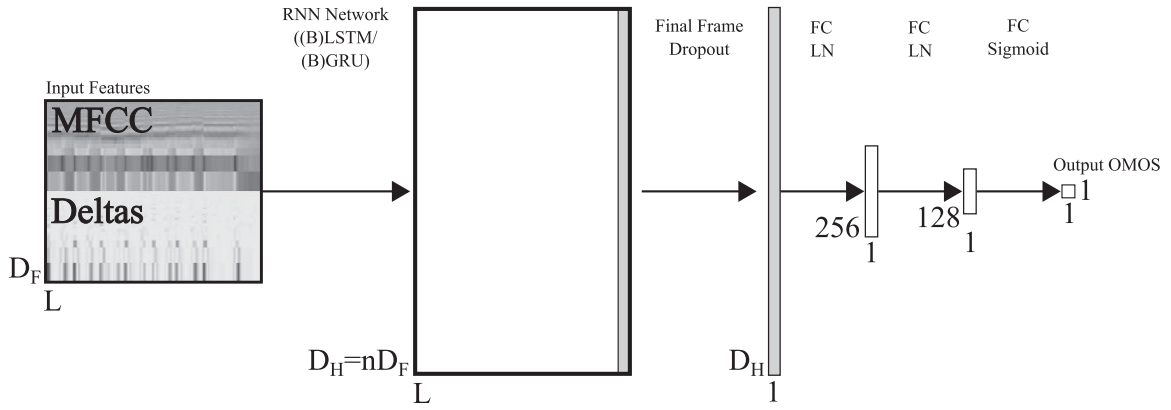

Fig. 3. Proposed FF RNN dataflow. D_F is feature depth, D_H is hidden dimensions, n is the number of directions, numbers denote layer sizes, FC is a fully connected layer, LN is layer normalization, and ReLU activation used unless specified. Mel-Frequency Cepstral Coefficients (MFCC) and first differences (Deltas) are used as input.

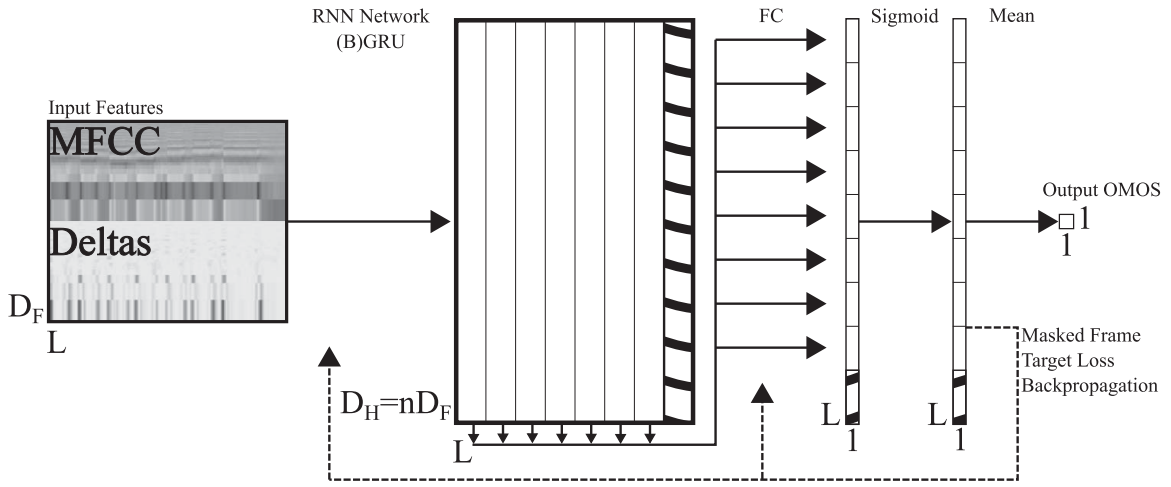

Fig. 4. Proposed GRU-FT network dataflow. D_F is feature depth, D_H is hidden dimensions, n is the number of directions, L is sequence length, numbers denote layer sizes, FC is a fully connected layer, and hashed sections are zero-padding to longest file in mini-batch. Mel-Frequency Cepstral Coefficients (MFCC) and first differences (Deltas) are used as input.

The proposed frame-target (FT) model for GRU and BGRU networks (GRU-FT and BGRU-FT) can be seen in Fig. 4. Two GRU layers of $D_H = 256$ with 10% dropout were used in a similar structure to the previous RNN. However a single fully connected layer with Sigmoid activation reduces feature dimensionality to $L \times 1$. The network has 1,972,737 trainable parameters. The MSE between the target SMOS and each frame estimate is used as loss. Frame targets are averaged for the length of the sequence to calculate the OMOS. As this calculation is independent of training, median, minimum, and maximum values of frame targets were also considered. Minimum frame targets were considered as quality evaluation of time-scaled signals in a degradation style analysis, where subjective quality is heavily influenced by the quality of the worst part of the signal. As the number of frames as a function of β or MOS is not a uniform distribution, the impact was explored by training on signals truncated to the minimum signal length and signals repeated to the maximum signal length.

1.2 Training

10% of the training dataset was reserved for validation. The CNN was trained for 100 epochs using a mini-batch size of 132, while RNNs were trained for 30 to 60 epochs with a mini-batch size of 48. A learning rate of $1e^{-4}$ was used in most cases, with $1e^{-5}$ if network performance stopped improving within the first 10 epochs. AdamW [39] was used as the optimizer for all networks. Loss for back-propagation was calculated using estimates in the interval of [0,1]. Reported loss values (L) were calculated using RMSE and estimates scaled back to the original interval of [1,5], for comparison with OMOQDE. As the prediction of opinion scores for novel TSM methods is the use case, early stopping based on validation loss was not used. The optimal epoch minimized the distance measure of [1], where the minimum overall distance ($\mathcal{D}$) is calculated by

$$\mathcal{D} = \|[\hat{\rho}, \hat{L}]\|_2, \tag{2}$$

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

$$\hat{\rho} = \|[1 - \overline{\boldsymbol{\rho}}, \Delta\boldsymbol{\rho}]\|_2, \tag{3}$$

$$\hat{L} = \|[\overline{\boldsymbol{L}}, \Delta\boldsymbol{L}]\|_2, \tag{4}$$

where $\boldsymbol{\rho} = [\rho_{tr}, \rho_{val}, \rho_{te}]$; $\boldsymbol{L} = [L_{tr}, L_{val}, L_{te}]$; *tr*, *val*, and *te* denote training, validation, and testing; $\overline{\boldsymbol{L}}$ is the mean of $\boldsymbol{L}$; $\overline{\boldsymbol{\rho}}$ is the mean of $\boldsymbol{\rho}$; $\Delta\boldsymbol{\rho} = max(\boldsymbol{\rho}) - min(\boldsymbol{\rho})$; and $\Delta\boldsymbol{L} = max(\boldsymbol{L}) - min(\boldsymbol{L})$. This scheme preferences networks with similar training, validation, and testing performance by penalizing over-training and allows for the novel artifacts of the test subset to inform the chosen optimal network, without their direct use in training.

An evaluation set of 6,000 files, published as part of [40], was generated from the test set reference files. Twenty new time-scales in the range of $0.22 < \beta < 2.2$, with all TSM methods listed in Sec. 0, were used to process the reference files. During evaluation, averages do not include $\beta = 0.2257$, as the minimum for EL is $\beta = 0.25$. $\beta = 1$ has also been excluded from averaging as it should be a unity system. Checking this is useful during development, but average OMOS should not improve for letting the output equal the input if $\beta = 1$.

2 RESULTS

2.1 Network Performance

A wide range of testing and network configurations were considered during the development of the proposed measures. Network hyper-parameters were optimized through a systematic non-exhaustive search. Deterministic training of all networks was conducted using seeds from 0 to 29. The use of RMSE, MSE, and Mean Absolute Error were explored, with the best performing loss function used in each proposed method. Fig. 5 shows the box plot distribution of the best $\mathcal{D}$ for each seed, where lower is better.

Median overall distance ($\widetilde{\mathcal{D}}$) and best case $\mathcal{D}$ with associated training, validation, and test loss and correlation values can be found in Table 1. While the improvement in performance appears linear in Fig. 5, many network configurations have not been included. Most networks trained with [MFCCs;D] achieved $0.55 < L_{te} < 0.67$, with only BGRU-FT achieving $L_{te} > 0.68$ or $\mathcal{D} < 0.72$. This appears to be the L_{te} and $\mathcal{D}$ limit for these network configurations and input features, even with ρ_{tr} approaching 1 when allowed to over-train.

The results in Table 1 can be summarized as follows: The proposed CNN achieved an L_{te} of 0.801 and ρ_{te} of 0.637, while the proposed BGRU-FT network achieved an L_{te} of 0.762 and ρ_{te} of 0.682. In comparison to subjective session mean loss ($\overline{\boldsymbol{L}}$) and mean PCC ($\overline{\boldsymbol{\rho}}$), the CNN was placed at the 74th and 32nd percentiles, and the proposed BGRU-FT network was placed at the 84th and 39th percentiles.

To give an indication of what the networks may be learning, correlation between OMOQSE, OMOS, and OMOQDE features was calculated for CNN and BGRU-FT networks. No significant correlation was found with maximum correlations of 0.210 and 0.206 for CNN and BGRU-FT, respectively.

Several trends were seen across testing. Networks trained using [MFCCs;D] features out-performed those trained using [MFCCs;D;D′] as well as solely MFCCs, magnitude spectra, magnitude and phase spectra, and the power spectrum. In all cases, magnitude only features out-performed combined magnitude and phase features. Decreased performance due to the inclusion of phase is likely because of its noise-like quality. Results for networks trained on the power spectrum are not shown in plots to increase comprehension. Improved performance was found using MFCCs generated with Librosa over TorchAudio with identical settings. For RNNs, FT measures outperformed FF measures, GRU outperformed LSTM, and bidirectional networks generally outperformed single direction networks for the same input features and network size. As such, RNN analysis will focus on BGRU-FT, alongside analysis of CNN performance.

The BGRU-FT network consistently rates TSM methods that use signal decomposition highly, whereas the CNN does not favor this method of TSM, with results

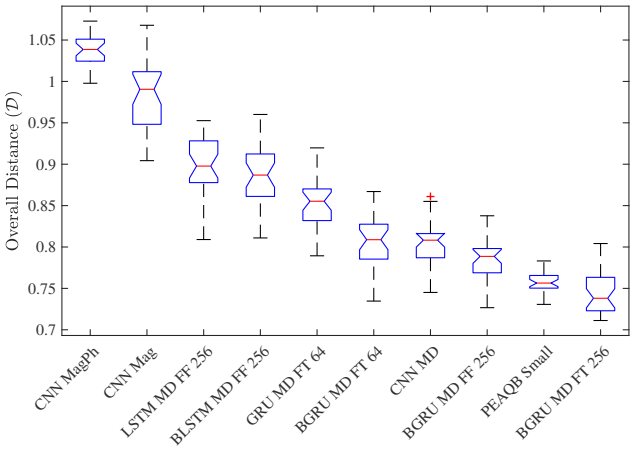

Fig. 5. Distance measure for each network configuration, ordered by median. Network input of $|X|$ is denoted by Mag, $\angle X$ by Ph, [MFCCs;D] by MD, and hidden size by 64 or 256.

more closely matching SMOS method ordering. While the BGRU-FT network achieves better loss and correlation results, it does not rate files greater than four and shows very little variation in OMOS, between TSM methods, and across the time-scale range, showing the CNN model to be the better predictor of SMOS.

2.1.1 CNN Performance

The CNN improved significantly through the use of [MFCCs;D] over $|X|$, $\angle X$, and $|X|^2$, with similar performance to FF RNNs. Normalization of input spectra reduced network performance. Small gains were found through optimizing the kernel size; however the maximum kernel size was limited by the length of the shortest file. Repeating signals to the length of the longest files decreased network performance as did using a combination of repeating or truncating to 500 or 1,000 frames. The CNN predicts across most of the OMOS range, shown in Fig. 6, but fails to correctly predict values below 2 for the test set. The network also does not predict scores above 4 when using the evaluation set, Fig. 8, even though it contains files with $\beta = 1$. Overall loss and correlation can be found in Table 1, while signal class loss and correlation can be found in Table 2. The network gives similar performance between signal classes for each set.

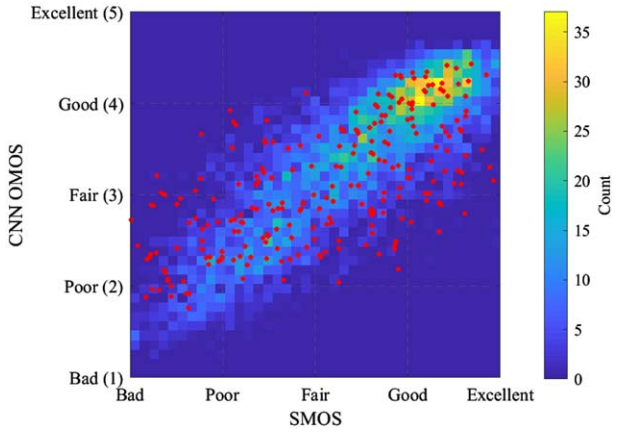

Fig. 6. Training subset confusion matrix for CNN OMOQSE and SMOS. Testing subset overlaid as dots.

2.1.2 BGRU-FT Performance

The BGRU-FT was found to give the best performance of the tested SE networks according to the distance measure

Table 1. Training, Validation, and Test Loss (L) and PCC (ρ), median overall distance ($\widetilde{\mathcal{D}}$), and minimum overall distance ($\min(\mathcal{D})$). Best SE results in bold.

Ended	Network	Features	Hidden	L_{tr}	ρ_{tr}	L_{val}	ρ_{val}	L_{te}	ρ_{te}	$\widetilde{\mathcal{D}}$	$\min(\mathcal{D})$
SE	BLSTM-FF	$\|X\|^2$	512	0.929	0.002	0.976	0.221	1.123	0.244	1.364	1.344
SE	LSTM-FF	$\|X\|^2$	512	0.929	0.011	0.974	0.239	1.064	0.262	1.350	1.322
SE	CNN	$[\|X\|;\angle X]$	...	0.754	0.584	0.855	0.502	0.942	0.484	1.039	0.998
SE	CNN	$\|X\|$	...	0.606	0.757	0.685	0.713	0.944	0.553	0.991	0.904
SE	LSTM-FF	[MFCCs;D]	256	0.854	0.581	0.663	0.295	**0.720**	0.221	0.898	0.809
SE	BLSTM-FF	[MFCCs;D]	256	0.567	0.795	0.593	0.757	0.849	0.581	0.887	0.811
SE	GRU-FT	[MFCCs;D]	64	0.632	0.746	0.646	0.707	0.820	0.649	0.855	0.789
SE	BGRU-FT	[MFCCs;D]	64	0.491	0.854	0.563	0.782	0.778	0.675	0.809	0.735
SE	CNN	[MFCCs;D]	...	0.500	0.843	0.522	**0.834**	0.801	0.637	0.808	0.745
SE	BGRU-FF	[MFCCs;D]	256	0.536	0.820	0.546	0.801	0.784	0.667	0.789	0.728
DE	FCNN	PEAQB	3	0.677	0.674	0.637	0.734	0.691	0.749	0.756	0.731
SE	BGRU-FT	[MFCCs;D]	256	**0.454**	**0.874**	**0.512**	0.824	0.762	**0.682**	**0.738**	**0.711**

Table 2. Loss and correlation per class for CNN validation and test sets. Best results in bold.

Class	L_{val}	ρ_{val}	L_{te}	ρ_{te}
Music	0.551	0.820	0.840	**0.682**
Solo	0.516	0.825	**0.752**	0.600
Voice	**0.502**	**0.841**	0.791	0.570

Table 3. Loss and correlation per class for BGRU-FT validation and test sets. Best results in bold.

Class	L_{val}	ρ_{val}	L_{te}	ρ_{te}
Music	**0.496**	0.832	0.873	0.627
Solo	0.510	**0.829**	0.696	0.678
Voice	0.531	0.799	**0.671**	**0.713**

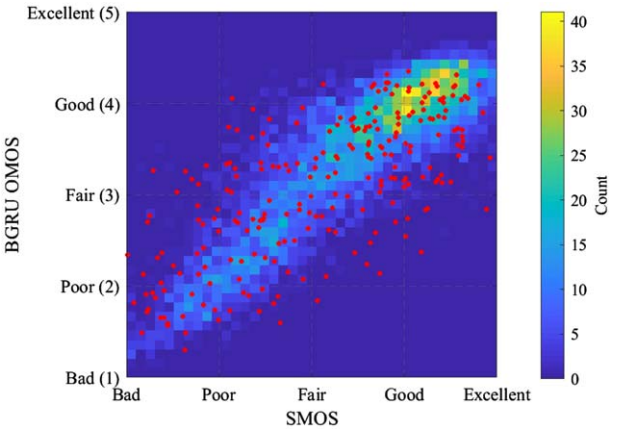

Fig. 7. Training subset confusion matrix for BRGU-FT OMOQSE and SMOS. Testing subset overlaid as dots.

and gives similar performance to OMOQDE. The proposed method gives the best L_{tr}, ρ_{tr}, L_{val}, and ρ_{te} performance, resulting in the best $\widetilde{\mathcal{D}}$ and min($\mathcal{D}$) scores. Exact results are shown in Table 1. When collapsing estimated frame targets, no significant difference was found between mean or median of predictions, while selecting the minimum or maximum prediction reduced performance.

The BGRU-FT predicts across most of the range of SMOS, seen in Fig. 7; however scores above 3.5 were not predicted when using the evaluation set, Fig. 9. The network gives similar performance between signal classes for each set, seen in Table 2. A hidden size of 256 outperformed 64, 128, and 512 sizes, with 10% dropout outperforming 0%, 25%, and 50%. Including D′ was found to reduce performance, as did increasing the number of MFCCs to 256. Multiple fully connected layers were also explored but did not improve performance. FT RNNs slightly improved performance over FF RNNs, with FF improvements following BGRU-FT results, with the best FF network shown in Fig. 5 and Table 1.

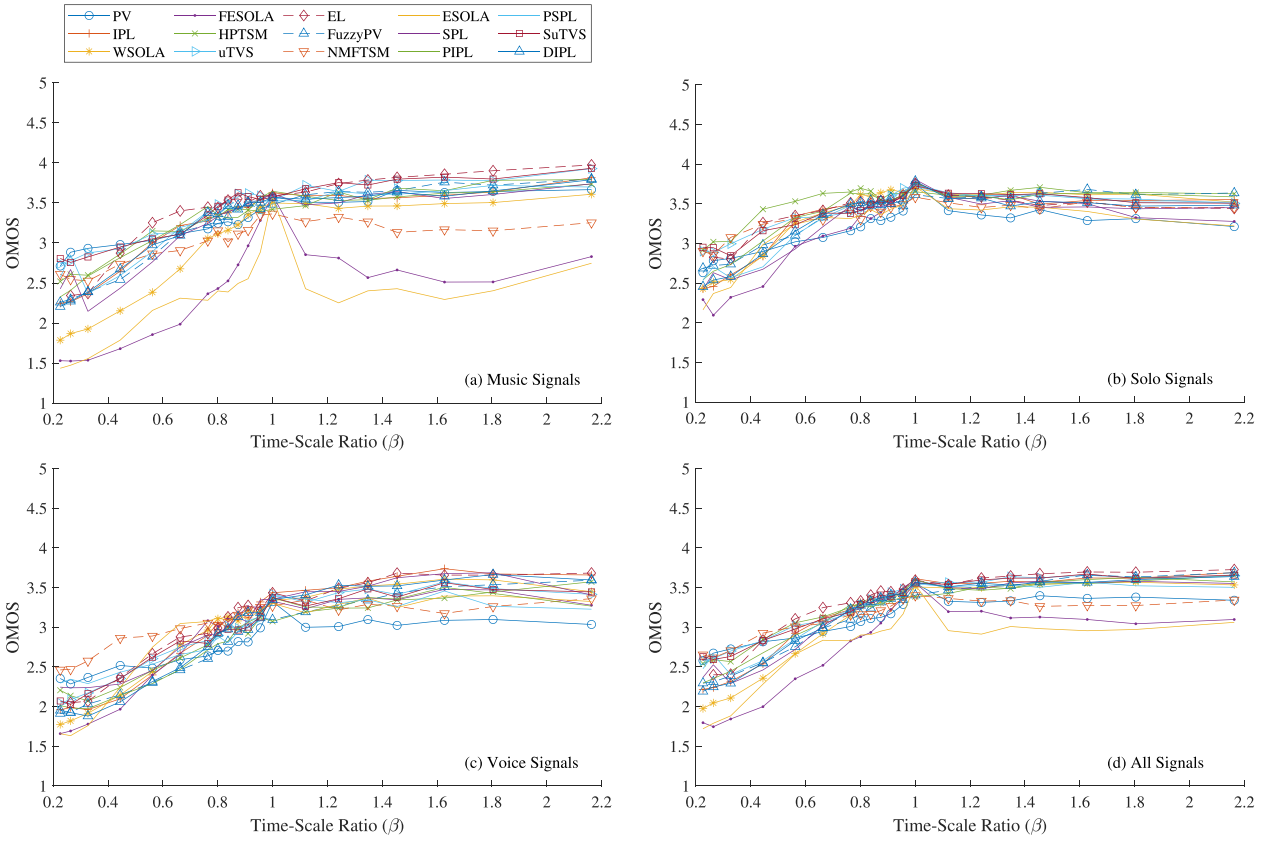

Fig. 8. CNN estimated Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals, and (d) All signals combined.

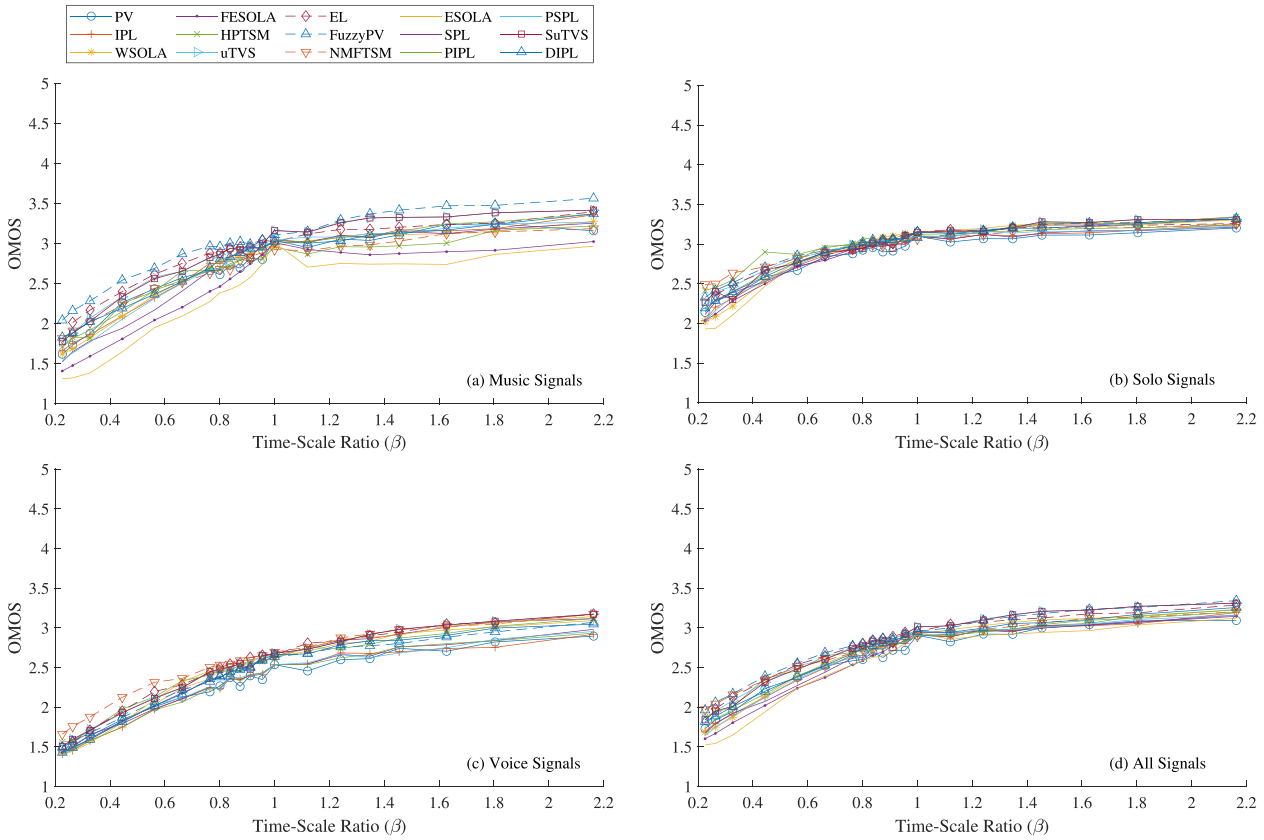

Fig. 9. BGRU-FT estimated Mean OMOS for each TSM method as a function of β for: (a) Musical signals, (b) Solo signals, (c) Voice signals, and (d) All signals combined.

Experiments showed using truncated random segments with BGRU-FT reduced performance, as did extending signals through repetition. Repeating input for the CNN also reduced performance.

2.2 TSM Algorithm Evaluation

In this section TSM methods are evaluated using the aforementioned evaluation set. Tables 4 and 5 show average OMOS for each signal class per TSM method ordered by overall mean OMOS, Figs. 8 and 9 show average OMOS per TSM method and β, and Figs. 10 and 11 show TSM methods for which differences in mean are statistically significant. As in [20] the design choice of excluding results for $\beta = 1$ and $\beta < 0.25$ from averaging calculations forming Tables 4 and 5 was used, as time-scaling is applied for $\beta \neq 1$, while $\beta = 0.25$ was the minimum available for EL. Common trends are presented, followed by CNN and then BGRU-FT analysis.

Estimation of signals time-scaled using NMF was particularly challenging for all networks. This is likely due to novel artifacts described by [29] and SMOS distribution skewed toward low scores. This provides a challenge for network design as novel TSM methods may not have similar artifacts or SMOS distributions to those in the training set. However the relative rating of EL and FPV to other TSM methods follows that of subjective testing. As suggested by network L_{te} and ρ_{te} only a general sense of TSM quality is obtained. Small details, such as the reduced quality of uTVS used in subjective testing at $\beta \approx 1$, are not visible. The networks have also not learned the non-linearity of SMOS as a function of β, continuing to increase for $\beta > 1$, seen in Figs. 8 and 9. The uniform quality of methods at $\beta = 1$ is however visible, as is the reduction in TSM quality for $\beta < 1$.

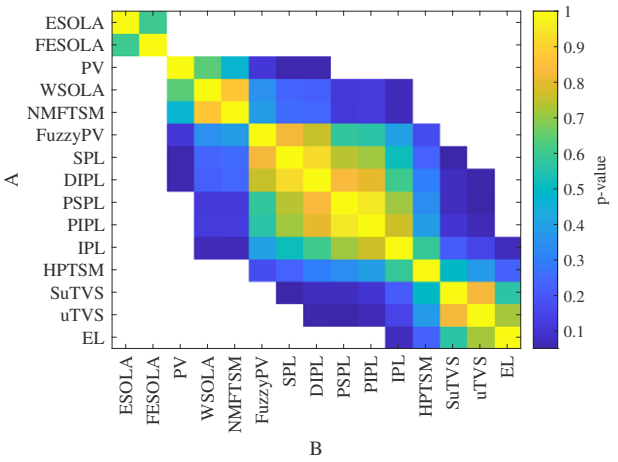

Fig. 10. Masked two-sample t-test (A vs. B) for all CNN OMOS estimates for each TSM method. Showing $p > 0.05$ for TSM method comparisons where the difference in mean is not statistically significant. Unequal means indicated by white.

Table 4. Mean OMOS for each class of file and overall result using the proposed CNN OMOQ. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$.

	ES	FES	PV	WS	NMF	FPV	SPL	DIPL	$\overline{\text{SPL}}$	$\overline{\text{IPL}}$	IPL	HP	$\overline{\text{uTVS}}$	uTVS	EL
Music	2.291	2.424	3.318	3.045	3.053	3.290	3.259	3.299	3.327	3.335	3.294	3.325	3.460	3.469	3.445
Solo	3.248	3.209	3.202	3.416	3.370	3.396	3.297	3.343	3.375	3.441	3.372	3.553	3.424	3.435	3.419
Voice	2.966	2.938	2.793	3.020	3.082	2.886	3.064	2.982	2.948	2.879	3.053	2.925	2.988	2.999	3.104
Overall	2.781	2.814	3.126	3.149	3.157	3.200	3.212	3.217	3.227	3.230	3.245	3.274	3.307	3.318	3.335

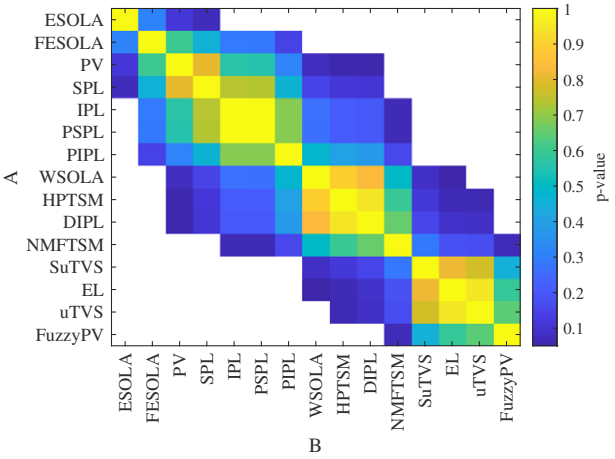

Fig. 11. Masked two-sample t-test (A vs. B) for all BGRU-FT OMOS estimates for each TSM method. Showing $p > 0.05$ for TSM method comparisons where the difference in mean is not statistically significant. Unequal means indicated by white.

2.2.1 CNN Evaluation of TSM Methods

For musical files, Fig. 8(a), the CNN differentiates between frequency and time-domain methods, where quality rapidly falls for time-domain methods when $\beta < 1$. WS fares the best of the time-domain methods, diverging from frequency domain methods for $\beta < 0.8$. The relative improvement in PV quality is also visible for $\beta < 0.5$ and slower falloff of EL. When averaged, the CNN rates uTVS and subjective uTVS highest followed by EL. For solo files, Fig. 8(b), HP exceeds other methods for $\beta < 1$. This class is the only evaluation where the highest OMOS is at $\beta = 1$. HP has the highest mean OMOS, followed by PhaVoRIT IPL, both uTVS methods, EL, and WS, as shown in Table 4. Voice file OMOS, Fig. 8(c), is comparatively low for time-domain methods, which is unexpected, as speech is often scaled well by time-domain methods. The high quality of NMF is also unexpected based on subjective results in [1]. EL has the highest mean OMOS, followed by NMF, SPL, and IPL. All other methods gave similar averaged quality.

For all CNN OMOS, EL has the highest average rating, followed by both uTVS methods and HP. The overall means can be seen in Fig. 8(d). The five best methods are separated by < 0.1 OMOS with a maximum difference of 0.554 for all methods. The overall low quality of FPV is unexpected, given that it builds on IPL; however further analysis is required to determine if the difference is statistically significant. A two-sample t-test ($\alpha = 0.05$) of all OMOS shows the null hypothesis, TSM methods having equal means, to be rejected in almost all cases when the absolute difference of mean OMOS is greater than 0.098. Fig. 10 shows p-values for the t-tests that were unable to reject equal means.

2.2.2 BGRU-FT Evaluation of TSM Methods

BGRU-FT OMOS shows the most variance for music files, Fig. 9(a). Again, time-domain methods rate lower. FPV is rated highest, followed by uTVS and EL. For multiple TSM methods, OMOS continues to increase for $\beta > 1$. By combining this information with the improvement when D is included as an input feature, we theorize that BGRU-FT is learning a relationship between SMOS and the velocity and duration between D events. As β increases, the time between sound events decreases and the attack portion of the energy envelope becomes sharper.

For solo files, Fig. 9(b), there is very little variance between methods, with a maximum difference $\approx$0.5 OMOS for $\beta = 0.2257$. Solo files have the highest overall OMOS of the 3 classes, which is consistent with subjective findings. HP has the highest mean OMOS, followed by uTVS and FPV. Voice file OMOS, Fig. 9(c), shows the lowest TSM quality of the 3 classes with a continued increase in OMOS for $\beta > 1$ across all TSM methods. NMF has the highest mean OMOS, which is unexpected based on [1]. EL is the next highest, followed by uTVS methods and ES. The high quality of ES is expected as the method was designed for TSM of speech.

Table 5. Mean OMOS for each class of file and overall result using the proposed BGRU-FT OMOQ. Means calculated for $\beta \neq 0.2257$ and $\beta \neq 1$.

	ES	FES	PV	SPL	IPL	$\overline{\text{SPL}}$	$\overline{\text{IPL}}$	WS	HP	DIPL	NMF	$\overline{\text{uTVS}}$	EL	uTVS	FPV
Music	2.378	2.511	2.723	2.711	2.764	2.733	2.764	2.741	2.699	2.787	2.720	2.890	2.899	2.901	3.016
Solo	2.925	2.947	2.891	2.917	2.917	2.943	2.976	2.978	3.035	2.978	2.988	2.983	2.988	3.005	3.011
Voice	2.503	2.468	2.323	2.350	2.329	2.345	2.334	2.472	2.492	2.443	2.591	2.526	2.544	2.532	2.444
Overall	2.580	2.629	2.653	2.664	2.680	2.680	2.699	2.732	2.738	2.741	2.762	2.809	2.819	2.822	2.843

For overall OMOS, Fig. 8(d), FPV has the highest average rating followed by uTVS methods and EL. The ordered ranking of methods is close to expected, with only NMF ranking unexpectedly. This is possibly due to the method retaining the shape of percussive elements during time-scaling. The six best methods are separated by < 0.102 OMOS, with a maximum difference of 0.263 for all methods. A two-sample t-test analysis ($\alpha = 0.05$) of all OMOS shows the null hypothesis of equal means to be rejected in almost all cases when the absolute difference of mean OMOS is greater than 0.098. Fig. 11 shows p-values for the t-tests that were unable to reject equal means.

The OMOQDE took approximately 15 h to evaluate the 6,000 files of the evaluation set (approximately 7 h of audio), whereas the proposed networks took approximately 15 min on system with a Xeon E5-2630 and NVIDIA GTX1080. The majority of this improvement is due to the removal of time-frequency spreading when calculating PEAQ features. The elimination of alignment between reference and test signals is also beneficial as it removes an additional temporal manipulation before feature calculation. While OMOQDE is a more accurate estimate of time-scaling quality, the proposed OMOQSE measures give a fast relative quality assessment and provide a platform for future SE objective measures.

3 AVAILABILITY

The proposed CNN and BGRU-FT tools are available from github.com/zygurt/TSM. This includes Python scripts for feature generation, proposed methods implemented in PyTorch, and evaluation methods. A bash script is also included to simplify use.

4 FUTURE RESEARCH

This study shows promise in non-invasive evaluation of the quality of TSM methods. However improvements can be made through input feature selection and exploring the use of phase derivatives or instantaneous frequency. Generalization to unseen TSM methods and sound sources is also an area for future research. More research needs to be conducted regarding duration invariant transformations that limit the network's ability to learn simple relationships such as the duration of musical events within the signal to SMOS. Additional attention could also be given to network architectures, such as Transformer Networks [41]. Pre-training using a large task-related dataset could also be explored.

5 CONCLUSION

Two single-ended objective measures for time-scaled audio are proposed with performance matching that of simple OMOQDE measures with reduced processing time. The SMOS score is estimated from [MFCCs;D] inputs using CNN and BGRU-FT network architectures, by training the networks to the SMOS of the TSMDB. The CNN uses four convolutional layers with batch normalization followed by concatenation before an FCNN with residual connections estimates the OMOS. The BGRU-FT network gives a single output for each frame that is fed into an FCNN for final OMOS prediction. The CNN achieves an $\overline{L}$ of 0.608 and $\overline{\rho}$ of 0.771, whereas BGRU-FT achieves an $\overline{L}$ of 0.576 and $\overline{\rho}$ of 0.794. The proposed measures are used to evaluate TSM methods, with estimates consistent with relative quality found in subjective testing. Future work includes exploration of alternative features and network architectures.

6 REFERENCES

[1] T. Roberts and K. K. Paliwal, "A Time-Scale Modification Dataset With Subjective Quality Labels," *J. Acoust. Soc. Am.*, vol. 148, no. 1, pp. 201–210 (2020 Jul.). https://doi.org/10.1121/10.0001567.

[2] ITU-R, "General Methods for the Subjective Assessment of Sound Quality," *Recommendation ITU-R BS. 1284-2* (2019 Jan.).

[3] D.-S. Kim and A. Tarraf, "ANIQUE+: A New American National Standard for Non-Intrusive Estimation of Narrowband Speech Quality," *Bell Labs Tech. J.*, vol. 12, no. 1, pp. 221–236 (2007 Mar.). https://doi.org/10.1002/bltj.20228.

[4] T. H. Falk and W.-Y. Chan, "Single-Ended Speech Quality Measurement Using Machine Learning Methods," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1935–1947 (2006 Nov.). https://doi.org/10.1109/TASL.2006.883253.

[5] L. Malfait, J. Berger, and M. Kastner, "P. 563—The ITU-T Standard for Single-Ended Speech Quality Assessment," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1924–1934 (2006 Nov.). https://doi.org/10.1109/TASL.2006.883177.

[6] D. Sharma, Y. Wang, P. A. Naylor, and M. Brookes, "A Data-Driven Non-Intrusive Measure of Speech Quality and Intelligibility," *Speech Commun.*, vol. 80, pp. 84–94 (2016 Jun.). https://doi.org/10.1016/j.specom.2016.03.005.

[7] B. Patton, Y. Agiomyrgiannakis, M. Terry, et al., "AutoMOS: Learning a Non-Intrusive Assessor of Naturalness-of-Speech," presented at the *NIPS 2016 End-to-End Learning for Speech and Audio Processing Workshop* (2016 Nov.). https://arxiv.org/abs/1611.09207.

[8] C.-C. Lo, S.-W. Fu, W.-C. Huang, et al., "Mosnet: Deep Learning-Based Objective Assessment for Voice Conversion," *arXiv preprint arXiv:1904.08352* (2019 Apr.).

[9] Y. Choi, Y. Jung, and H. Kim, "Deep MOS Predictor for Synthetic Speech Using Cluster-Based Modeling," in *Proceedings of the 21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1743–1747 (Shanghai, China) (2020 Oct.). https://doi.org/10.21437/Interspeech.2020-2111.

[10] Y. Choi, Y. Jung, and H. Kim, "Neural MOS Prediction for Synthesized Speech Using Multi-Task Learning With Spoofing Detection and Spoofing Type Classification," in *Proceedings of the 8th IEEE Spoken Language Technology Workshop (SLT)*, pp. 462–469 (Shenzhen, China) (2021 Jan.). https://doi.org/10.1109/SLT48900.2021.9383533.

[11] H. Gamper, C. K. A. Reddy, R. Cutler, I. J. Tashev, and J. Gehrke, "Intrusive and Non-Intrusive Perceptual Speech Quality Assessment Using a Convolutional Neural Network," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 85–89 (New Paltz, NY) (2019 Oct.). https://doi.org/10.1109/WASPAA.2019.8937202.

[12] T. Thiede, W. C. Treurniet, R. Bitto, et al., "PEAQ-The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29 (2000 Feb.).

[13] J. G. Beerends, C. Schmidmer, J. Berger, et al., "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I—Temporal Alignment," *J. Audio Eng. Soc.*, vol. 61, no. 6, pp. 366–384 (2013 Jul.).

[14] R. Huber and B. Kollmeier, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1902–1911 (2006 Nov.). https://doi.org/10.1109/TASL.2006.883259.

[15] M. Chinen, F. S. C. Lim, J. Skoglund, et al., "ViSQOL v3: An Open Source Production Ready Objective Speech and Audio Metric," *arXiv preprint arXiv:2004.09584* (2020 Apr.).

[16] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-Net: An End-to-End Non-Intrusive Speech Quality Assessment Model Based on BLSTM," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 1873–1877 (Hyderabad, India) (2018 Sep.). https://doi.org/10.21437/Interspeech.2018-1802.

[17] A. A. Catellier and S. D. Voran, "Wawenets: A No-Reference Convolutional Waveform-Based Approach to Estimating Narrowband and Wideband Speech Quality," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 331–335 (2020 May). https://doi.org/10.1109/ICASSP40776.2020.9054204.

[18] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)-A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in *Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001) (Cat. No. 01CH37221)*, vol. 2, pp. 749–752 (Salt Lake City, UT) (2001 May). https://doi.org/10.1109/ICASSP.2001.941023.

[19] L. Fierro and V. Välimäki, "Towards Objective Evaluation of Audio Time-Scale Modification Methods," in *Proceedings of the 17th Sound and Music Computing Conference (SMC)*, pp. 457–462 (Torino, Italy) (2020 Jun.).

[20] T. Roberts and K. K. Paliwal, "An Objective Measure of Quality for Time-Scale Modification of Audio," *J. Acoust. Soc. Am.*, vol. 149, no. 3, pp. 1843–1854 (2021 Mar.). https://doi.org/10.1121/10.0003753.

[21] M. Portnoff, "Implementation of the Digital Phase Vocoder Using the Fast Fourier Transform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 24, no. 3, pp. 243–248 (1976 Jun.). https://doi.org/10.1109/TASSP.1976.1162810.

[22] J. Laroche and M. Dolson, "Improved Phase Vocoder Time-Scale Modification of Audio," *IEEE Trans. Speech Audio Process.*, vol. 7, no. 3, pp. 323–332 (1999 May). https://doi.org/10.1109/89.759041.

[23] W. Verhelst and M. Roelands, "An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 554–557 (1993). https://doi.org/10.1109/ICASSP.1993.319366.

[24] T. Roberts and K. K. Paliwal, "Time-Scale Modification Using Fuzzy Epoch-Synchronous Overlap-Add (FESOLA)," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 31–34 (New Paltz, NY) (2019 Oct.). https://doi.org/10.1109/WASPAA.2019.8937258.

[25] J. Driedger, M. Müller, and S. Ewert, "Improving Time-Scale Modification of Music Signals Using Harmonic-Percussive Separation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 105–109 (2014 Jan.). https://doi.org/10.1109/LSP.2013.2294023.

[26] N. Sharma, S. Potadar, S. R. Chetupalli, and T. V. Sreenivas, "Mel-Scale Sub-Band Modelling for Perceptually Improved Time-Scale Modification of Speech and Audio Signals," in *Proceedings of the 23rd National Conference on Communications (NCC)*, pp. 1–5 (Chennai, India) (2017 Mar.). https://doi.org/10.1109/NCC.2017.8077073.

[27] Zplane Development, "Èlastique Time Stretching & Pitch Shifting SDKs (Version 3.2.5)" (accessed October 31, 2019).

[28] E.-P. Damskägg and V. Välimäki, "Audio Time Stretching Using Fuzzy Classification of Spectral Bins," *Appl. Sci.*, vol. 7, no. 12, paper 1293 (2017 Dec.). https://doi.org/10.3390/app7121293.

[29] G. Roma, O. Green, and P. A. Tremblay, "Time Scale Modification of Audio Using Non-Negative Matrix Factorization," in *Proceedings of the 22nd International Conference on Digital Audio Effects (DAFx)*, pp. 213–218 (Birmingham, UK) (2019 Sep.).

[30] T. Karrer, E. Lee, and J. Borchers, "PhaVoRIT: A Phase Vocoder for Real-Time Interactive Time-Stretching," in *Proceedings of the International Computer Music Conference*, pp. 708–715 (2006 Nov.).

[31] S. Rudresh, A. Vasisht, K. Vijayan, and C. S. Seelamantula, "Epoch-Synchronous Overlap-Add (ESOLA) for Time- and Pitch-Scale Modification of Speech Signals," *arXiv preprint arXiv:1801.06492* (2018 Jan.).

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444 (2015 May). https://doi.org/10.1038/nature14539.

[33] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780 (1997 Nov.). https://doi.org/10.1162/neco.1997.9.8.1735.

[34] K. Cho, B. Van Merriënboer, C. Gulcehre, et al., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv preprint arXiv:1406.1078* (2014 Sep.).

[35] M. Schuster and K. K. Paliwal, "Bidirectional Recurrent Neural Networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681 (1997 Nov.). https://doi.org/10.1109/78.650093.

[36] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366 (1980 Aug.). https://doi.org/10.1109/TASSP.1980.1163420.

[37] A. Nicolson, J. Hanson, J. Lyons, and K. Paliwal, "Spectral Subband Centroids for Robust Speaker Identification Using Marginalization-Based Missing Feature Theory," *Int. J. Signal Process. Syst.*, vol. 6, no. 1, pp. 12–16 (2018 Mar.). https://doi.org/10.18178/ijsps.6.1.12-16.

[38] ITU-R, "Method for Objective Measurements of Perceived Audio Quality," *Recommendation ITU-R BS.1387-1* (2001 Nov.).

[39] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," *arXiv preprint arXiv:1711.05101* (2017 Nov.).

[40] T. Roberts and K. K. Paliwal, "An Objective Measure of Quality for Time-Scale Modification of Audio," *J. Acoust. Soc. Am.*, vol. 149, no. 3, pp. 1843–1854 (2021 Mar.). https://doi.org/10.1121/10.0003753.

[41] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention is All You Need," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, pp. 5998–6008 (Long Beach, CA) (2017 Dec.).

THE AUTHORS


Timothy Roberts


Aaron Nicolson


Kuldip K. Paliwal

Timothy Roberts was born in Stanthorpe, Australia, in 1989. He received BMus (Tech), BEng (Class 1A Hons.), and Ph.D. degrees from Griffith University, Brisbane, Australia, in 2009, 2016, and 2021, respectively. He is currently a research engineer and lecturer at Griffith University. His research interests include music technology, time-scale modification, expressive musical controllers, and digital signal processing.

•

Aaron Nicolson was born in Brisbane, Australia, in 1994. He received a BEng (Class 1 A Hons.) and Ph.D. degree from Griffith University, Brisbane, Australia, in 2016 and 2020, respectively. He is currently a postdoctoral research fellow at the Australian eHealth Research Centre, CSIRO. His research interests include speech, natural language, image, and multimodal processing using deep learning.

•

Kuldip K. Paliwal was born in Aligarh, India, in 1952. He received a B.S. degree from Agra University, Agra, India, in 1969; M.S. degree from Aligarh Muslim University, Aligarh, India, in 1971; and Ph.D. degree from Bombay University, Bombay, India, in 1978. He has worked at a number of organizations including Tata Institute of Fundamental Research, Bombay, India, and AT&T Bell Laboratories, Murray Hill, New Jersey, USA. Since July 1993, he has been a professor at Griffith University, Brisbane, Australia, in the School of Microelectronic Engineering. His current research interests include speech recognition, speech coding, speaker recognition, speech enhancement, face recognition, image coding, pattern recognition, and artificial neural networks.