A Multitask Teacher-Student Framework for Perceptual Audio Quality Assessment

Chih-Wei Wu, Phillip A. Williams, William Wolcott
Netflix, Inc.
Los Gatos, United States
{chihweiw, phill, wwolcott}@netflix.com

***Abstract*—Perceptual audio quality assessment is a task that involves the characterization and estimation of perceived quality of an audio signal. Many existing systems, depending on psychoacoustic principles and statistical models, achieve reasonable performance under specific conditions (e.g., type of artifacts, impairment levels, etc.) but do not generalize well when these conditions vary. This lack of generality often limits their utility in real-world scenarios. In this paper, we address this challenge by leveraging the domain knowledge from several state-of-the-art expert systems. Particularly, we explore the idea of training a multitask student model using unlabeled data and the pseudo labels from multiple expert (i.e. teacher) systems. Evaluation is conducted using a variety of test datasets, and the results show that our proposed system compares favorably with the state-of-the-art systems and achieves the highest overall performance.**

***Index Terms*—audio quality, multitask learning, teacher-student learning, unlabeled data**

I. Introduction

Assessing perceptual quality is a crucial step in the development of many audio algorithms (e.g. source separation [1], audio coding [2], etc.). Generally speaking, the goal is to quantify the perceived quality of the processed signals and understand the perceptual impact induced by the algorithms. A standard procedure of such assessments relies on the collection of human ratings through carefully designed subjective listening tests, in which the estimated quality of the selected audio signals are represented by their Mean Opinion Scores (MOSs). Despite being an effective approach, conducting a well-controlled and bias-free listening test is non-trivial [3] and time-consuming. A computational approach, on the contrary, has the potential of providing a consistent and less labor-intensive solution; a robust model not only allows fast experimentation of research at a smaller scale, but also enables the examination of audio quality on a large scale.

To computationally assess the perceptual audio quality, different techniques have been proposed [4], [5]. Existing systems, such as Perceptual Evaluation of Audio Quality (PEAQ) [6] and Perceptual Objective Listening Quality Assessment (POLQA) [7], typically consist of two stages: in the first stage, features that incorporate domain knowledge such as psychoacoustics and human auditory system are extracted; in the following stage, these features are mapped to a perceptually meaningful scale through a pre-trained regression model. The second stage, which relies heavily on the integrity of the training materials, tends to have a profound impact on the generality of the resulting model [4]. For example, PEAQ has been reported to be sub-optimal when tested on low quality anchors that are different from its training data [8]. Similarly, POLQA, a quality metric trained on speech signals, was found to be less suitable for non-speech content [8]. To some extent, these limitations could be overcome by re-training the regression model. For instance, POLQA Music [9] was an adaptation of POLQA to non-speech content using additional training materials. However, this option is not always feasible due to the scarcity of publicly available large and diverse datasets.

In this paper, we address the challenges of model generality and data availability with the following ideas: first, we apply the teacher-student learning paradigm [10] to train a student model using unlabeled data and the pseudo labels, generated by the teacher models. Next, motivated by the recent success of multitask learning (MTL) in audio related tasks [11], [12], we propose a multitask student model that simultaneously predicts the output of multiple teachers in order to increase the model generality. To narrow the scope of this study, we focus on the assessment of degradation from various audio coding algorithms. The main contributions of this work include: (i) insights of leveraging unlabeled data and multiple expert systems as teachers for audio quality assessment, (ii) the exploration of MTL for improving the generality of the student model, and (iii) a multitask teacher-student framework for training a full-reference perceptual audio quality assessment system that achieves the best overall performance.

II. Related Work

Existing methods for signal-based perceptual audio quality assessments can be roughly divided into two categories, namely *non-intrusive* and *intrusive* [4]. *Non-intrusive* methods, also known as no-reference or single-ended, are designed to assess the quality of a target signal (i.e., signal with potential degradation) without any reference signal (i.e., signal with near-perfect quality). These types of systems offer great flexibility with less required input, but accuracy is often compromised due to the absence of a reference. Prior work in this category mainly focuses on speech quality evaluation [13], [14], and evaluation of general audio is relatively unexplored. *Intrusive* methods, also known as full-reference, assess the quality of a target signal through a direct comparison to a

reference signal. These types of systems have a more restricted use case in exchange for robustness and accuracy. The majority of the existing methods belong to this category [6]–[8], [15]–[22], including our proposed system.

Intrusive methods, depending on their signal of interest, can be further divided into two sub-categories: *speech* and *audio*. *Speech* focused systems, such as Perceptual Speech Quality Measure (PSQM) [15], Perceptual Evaluation of Speech Quality (PESQ) [16], POLQA [7], Hearing-Aid Speech Quality Index (HASQI) [17], and Virtual Speech Quality Objective Listener (ViSQOL) [18], typically operate on signals with narrow bandwidth and often achieve high correlation with human ratings [4]. Systems designed for *audio*, such as PEAQ [6], PEMO-Q [19], ViSQOLAudio [8], [20], Hearing-Aid Audio Quality Index (HAAQI) [21], and Generalized Power Spectrum Model (GPSM) [22], operate on full-bandwidth signals, and their performance tends to be content dependent. In a comparative study [23], Torcoli and Dick found that different expert systems are sensitive to different types of artifacts, which could potentially be attributed to their inherent designs and training materials. The study also implies the benefit of having a pool of expert systems with a diverse domain expertise.

Recently, MTL has been successfully applied to several audio related tasks [11], [12], [24]. By training a model to perform multiple related tasks in parallel, MTL is able to improve the generalization through the shared representation among these tasks [25]. For example, Hung et al. [11] proposed a MTL model trained on synthetic data that can jointly predict instrument, pitch, and piano roll representation at frame-level, and the results showed that the MTL-based method generalized well on real data and achieved strong performances compared to other baselines. Chen and Su [24] combined the challenging task of chord function recognition with chord symbol recognition through the MTL framework, and the results showed promising improvements for chord function recognition compared to its single task learning (STL) counterpart. Similarly, Böck et al. [12] proposed a MTL model for beat tracking and tempo estimations simultaneously. The resulting model not only performed well on multiple test datasets, but also showed capability of learning beat tracking through tempo labels only.

Inspired by the above mentioned studies, we explore the possibility of building a model which extracts and harmonizes information from different expert systems through the use of unlabeled data and the MTL paradigm.

III. Method

A. System Overview

The overview of our proposed system is shown in Fig. 1. The processing steps can be grouped into the training and the testing phase, respectively. In the training phase, a collection of unlabeled audio data is first gathered. Four existing expert systems (see Sect. III-C) are used as teacher models to generate the pseudo labels for the unlabeled data; these pseudo labels are predictions from the teacher models and will be used as pseudo ground truth for training purposes. To proceed with training, four selected audio features (see Sect. III-B) are extracted and scaled. Subsequently, the features and the pseudo labels are used to train a MTL student model. In particular, the student model is trained to predict all different pseudo ground truth labels simultaneously. The testing phase has a similar pipeline as in training. The same audio features are extracted from the test data, followed by a feature scaling process using the parameters estimated from the training data. A trained MTL student model is then used to predict all teachers' opinions at once. Finally, these scores are scaled to a five-grade MOS range that resembles the result from a subjective listening test.

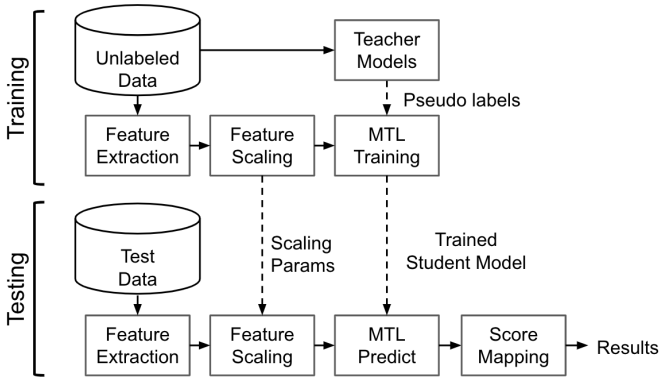

Fig. 1. Flowchart of the proposed system

B. Feature Extraction

To characterize the perceptual relevance of a signal from different perspectives, four audio features from previous studies are combined as the input representation for our system. These features are selected from the expert systems which encapsulate strong domain knowledge [23], and they are specifically chosen for their strong correlation with each expert system's predictions. These features include Noise-to-Mask Ratio (NMR) [26], weighted Perceptual Similarity Measure (PSMt) [19], Cepstral Correlation (CepCorr) [21], and Neurogram Similarity Index Measure (NSIM) [8], which are referred to as f_{nmr}, f_{psmt}, $f_{cepcorr}$, and f_{nsim} for the remainder of the paper.

The final feature vector can be summarized as $v_{all} = [f_{cepcorr}, f_{nmr}, f_{nsim}, f_{psmt}]$. All four features are scaled to a numerical range between 0 and 1 using a standard min-max scaling with the parameters estimated from the training data.

C. MTL Student Model

The MTL student model is a fully-connected deep neural network (DNN) consisting of four hidden layers. The first two hidden layers each contain 64 neurons, and the third and fourth layers contain 32 and 16 neurons, respectively. All layers use Rectified Linear Unit (ReLU) activation functions. Each layer is followed by a $dropout = 0.3$. The output layer consists of four neurons with Sigmoid activation functions. This model

architecture is chosen for its simplicity, and the exploration of more sophisticated architectures is left as a future work.

The model is trained by optimizing the following loss function:

$$\mathcal{L}_{MTL} = MSE(\hat{y}, y_{pseudo}), \quad (1)$$

where $MSE()$ is the mean squared error, $\hat{y} = [t1, t2, t3, t4]$ are the outputs (i.e. learned tasks) of the model, and $y_{pseudo} = [y_{haaqi}, y_{peaq}, y_{pemoq}, y_{visqol}]$ are the pseudo labels generated from all teacher models; the teacher models include HAAQI[1], PEAQ[2], PEMO-Q[3], and ViSQOLAudio[4]. In this paper, the objective of a task is to approximate a specific teacher model. All pseudo labels are linearly scaled to a range of 0 to 1 using their theoretical min/max values (e.g., {0, 1} for HAAQI, {-4, 0} for PEAQ and PEMO-Q, and {1, 5} for ViSQOLAudio). The resulting MTL student model is able to perform four tasks, and the output of each task can be mapped to a five-grade scale and used as an individual metric.

The model is implemented in Python using Tensorflow with Keras module.[5] All weights of the DNN are randomly initialized and optimized using Adam [27], and the model is trained using a learning rate $lr = 0.001$ for 400 epochs with the batch size = 32.

IV. Experiment

A. Experiment Setup

The following systems are included for benchmarking:

- *SNR*: a simple baseline that computes the signal-to-noise ratio between the reference and noise signal.
- *HAAQI*: a teacher model based on [21].
- *PEAQ*: a teacher model based on [6].
- *PEMO-Q*: a teacher model based on [19].
- *ViSQOLAudio*: a teacher model based on [20].
- *STL-(HAAQI, PEAQ, PEMO-Q, ViSQOLAudio)*: student models that are trained to approximate each of the four teachers independently.
- *MTL-(T1, T2, T3, T4)*: the proposed MTL student model. Each output task is evaluated individually.
- *MTL-(mean, gmean, median)*: a variant of the MTL student model by aggregating the four output tasks into one value. Three different aggregation methods are evaluated, namely the arithmetic mean, geometric mean, and median.

All of the above single task learning (STL) models have the same architecture as the MTL model except for the output layer, which contains only one Sigmoid neuron.

To account for the random initialization of DNN-based models (i.e., systems denoted by STL or MTL), the evaluation is repeated 5 times; in each run, the unlabeled data is randomly split into 90%/10% for training and validation, and the resulting model is tested on all the test datasets. Finally, the mean and standard deviation of the main metric (see Sect. IV-B) across 5 runs are computed. Note that none of the test datasets are used for training, which represents the most generalizable evaluation scenario.

[1]Matlab implementation provided by the author

[2]http://www-mmsp.ece.mcgill.ca/Documents/Software, 2021.02

[3]http://bass-db.gforge.inria.fr/peass/PEASS-Software.html, 2021.02

[4]https://qxlab.ucd.ie/index.php/audio-and-music, 2021.02

[5]https://www.tensorflow.org/api_docs/python/tf/keras, 2021.02

TABLE I
List of training and test datasets

Usage	Name	#Files	Total Dur. (hr)	Publicly Available?
Training	UnlabeledTV2020	3250	17.97	N
Test	Bs1387Conform	32	0.13	Y
	CoreSV14	280	1.47	Y
	UnbAvq2013	24	0.05	Y
	NLow	63	0.20	N
	NHigh	36	0.12	N

B. Metrics

The standard calculation of Pearson's correlation coefficient r is used as the main metric. For each test dataset, this metric is computed by correlating the predictions from each system with the human ground truth labels. In addition to the individual r, an overall metric $\bar{r}*$ is calculated using an approximately unbiased minimum-variance estimator as described in [28], which accounts for the unequal sizes of different datasets; this estimator summarizes the results across all test datasets and computes the weighted average of correlation coefficients. Additional metrics, such as root mean square error (RMSE), are also computed and made available in our online repository.[6]

C. Datasets

The list of datasets used in this paper is shown in Table I. The details of each dataset is explained as follows:

- *UnlabeledTV2020* is a proprietary unlabeled dataset created from the audio tracks of various TV shows. The audio content includes dialogue, sound effects, and music. There are 250 episodes from 5 different languages (i.e., English, Spanish, French, Japanese, and Brazilian Portuguese). A 20 sec audio clip is extracted from each episode and subsequently processed by 12 different treatments (i.e., processing methods such as audio coding and filtering), including: (i) 3.5 kHz lowpass (ii) 7 kHz lowpass (iii) 64kbps HE-AAC-v1 (iv) 96kbps HE-AAC-v1 (v) 96kbps EAC3 (vi) 32kbps MP3 (vii) 48kbps MP3 (viii) 96kbps MP3 (ix) 32kbps Opus (x) 48kbps Opus (xi) 96kbps Opus, and (xii) 96kbps Vorbis. The resulting training dataset consists of 3250 audio clips (≈ 18 hrs), which is significantly larger than other labeled test datasets listed in Table I.

For evaluation, we use five different labeled datasets:

- *Bs1387Conform*: these are 32 audio clips for validating the implementation of PEAQ[7]. The audio content includes short clips of speech or instrumental sounds (e.g., harpsichord, snare drum, triangle, etc.).
- *CoreSV14*: this is a crowd-sourced dataset[8] which consists of 35 music samples and 5 speech samples. Each sample is

[6]https://github.com/cwu307/mtl_audio_qual, 2021.02

[7]https://www.itu.int/rec/R-REC-BS.1387-1-200111-I/en, 2021.02

[8]http://listening-test.coresv.net/results.htm, 2021.02

TABLE II

EVALUATION RESULTS OF ALL SYSTEMS (SEE SECT. IV-B FOR THE DETAILS ON METRICS). ALL THE RESULTS HAVE STANDARD DEVIATION $\sigma \leq 0.05$, AND $(*)$ INDICATES THE RESULT WITH $\sigma \geq 0.04$. THE BEST PERFORMING SYSTEM OF EACH COLUMN IS IN BOLD.

Role	Systems	Test Datasets					Overall
		Bs1387Conform	CoreSV14	UnbAvq2013	NLow	NHigh	
Baseline	SNR	0.332	0.657	0.479	-0.298	0.312	0.461
Teacher	HAAQI	0.466	0.854	0.656	-0.346	0.207	0.594
	PEAQ	**0.922**	**0.945**	0.314	-0.330	0.558	0.686
	PEMO-Q	0.804	0.914	**0.680**	0.173	0.648	0.764
	ViSQOLAudio	0.706	0.782	0.620	**0.883**	0.383	0.759
STL Student	STL-HAAQI	0.436	0.876	0.631	-0.319	0.181	0.608
	STL-PEAQ	0.494	0.823	0.551	0.160*	0.679	0.684
	STL-PEMO-Q	0.802*	0.926	0.642	0.441	**0.682**	0.815
	STL-ViSQOLAudio	0.781	0.783	0.521*	0.863	0.358	0.753
MTL Student	MTL-T1	0.677	0.900	0.620	-0.269	0.525*	0.668
	MTL-T2	0.675	0.859	0.583	0.090*	0.597	0.700
	MTL-T3	0.868	0.920	0.677	0.411	0.643	0.808
	MTL-T4	0.866	0.886	0.575	0.875	0.614	**0.850**
MTL Student Aggregated	MTL-mean	0.856	0.909	0.637	0.560	0.613*	0.819
	MTL-gmean	0.888	0.902	0.644	0.782	0.613	**0.851**
	MTL-median	0.855	0.920	0.640	0.600	0.630	0.834

processed with 6 different treatments, including: (i) 48kbps FAAC (ii) 96kbps FAAC (iii) 96kbps QAAC (iv) 96kbps Opus (v) 96kbps Vorbis, and (vi) 128kbps MP3. In total, there are 280 audio clips with crowd-sourced MOSs.

- *UnbAvq2013 [29]*: this dataset consists of 6 audio samples from different scenes (e.g., sports event, music performance, news report, etc.). Each sample is processed with 3 treatments, including: (i) 32kbps MP3 (ii) 96kbps MP3, and (iii) 128kbps MP3. The resulting dataset contains 24 audio clips with subjective ratings.
- *NLow*: this is a proprietary dataset of low-bitrate coded items. This dataset consists of 7 audio samples from various acoustic scenes (e.g., dialog, environmental sounds, music, etc.); each sample is processed with 8 treatments, including: (i) 3.5 kHz lowpass (ii) 7 kHz lowpass (iii) 32kbps HE-AAC-v1 (iv) 64kbps HE-AAC-v1 (v) 24kbps HE-AAC-v2 (vi) 32kbps HE-AAC-v2 (vii) 16kbps xHE-AAC, and (viii) 24kbps xHE-AAC. A total number of 63 audio clips are included in this dataset. The subjective scores of these items were collected from an internal listening test conducted over loudspeakers.
- *NHigh*: this is another proprietary dataset of high-bitrate coded items. There are 9 audio samples from various acoustic scenes (e.g., crackling fire, gun shots, music, etc.); each sample is processed with 3 treatments, including: (i) 96kbps HE-AAC-v1 (ii) 128kbps HE-AAC-v1, and (iii) 128kbps AAC-LC. A total number of 36 audio clips are included in this dataset. The subjective scores of these items were collected from an internal headphone listening test.

All of the above mentioned datasets were decoded to the PCM format with sampling rate = 48 kHz and bit depth = 16. All the ground truth labels are linearly converted to the standard five-grade scale for consistency.

V. RESULTS AND DISCUSSION

The evaluation results are shown in Table II. The following observations can be made by comparing the baseline and teacher models: first, SNR does not achieve a comparable performance with other teacher models. This result is expected since this metric does not take into account any human perception. Second, all of the teacher models seem to perform reasonably well on at least one dataset while falling short on another. This result suggests that each teacher model has its own domain expertise, and none of them is versatile enough for all test datasets. Overall, the teacher models seem to perform well on *CoreSV14* and struggle on *NLow*. A possible explanation is that *NLow* was based on a listening test conducted on loudspeakers, which is an outlier among the test datasets. Nevertheless, ViSQOLAudio achieved a good performance on *NLow*, which shows the possibility of correctly predicting the perceptual quality of this heterogeneous challenging dataset.

A few interesting questions could be also answered by further observing the results. First, does MTL help the student model generalize better? By comparing the results between the STL and teacher models, each STL model seems to perform similarly with its corresponding teacher. This result is expected since the STL model is approximating the underlying function of a teacher model including its weaknesses. The MTL student model, on the other hand, is able to learn from multiple teachers and derive a more generalizable system. As a result, each learned task (i.e., MTL-T1 to T4) is able to outperform its corresponding teacher model and achieves a better overall performance. Second, is it beneficial to aggregate four learned tasks? Based on the overall performance, MTL-T4 and MTL-gmean seem to perform similarly. However, a further comparison of their coefficient of determination R^2 (0.52 versus 0.67) suggests a superior performance from the MTL-gmean. In short, our results show that geometric mean could be a simple yet effective way of aggregating the outputs from a MTL student model, but the optimal strategy still requires more investigations. It is worth noting that MTL-gmean does not exceed the teacher models by outperforming on any specific test dataset; instead, the MTL-gmean is able to perform comparably with the best teacher model on each test

dataset and eventually achieve the highest overall performance. This result implies the advantage of MTL-based student models in terms of fusing and harnessing the knowledge from teacher models without explicit human guidance.

VI. Conclusion

In this paper, we have presented a multitask teacher-student framework for perceptual audio quality assessment. The proposed system integrates audio features derived from different psycho-acoustic principles and simultaneously performs multiple assessments. Specifically, the MTL student model is trained to approximate several teacher models at once using an unlabeled dataset and pseudo labels. The evaluation results show that our proposed system performs consistently well on different test datasets and achieves the highest overall performance. One potential limiting factor in our proposed method is the pool of teacher models, and the impact of different compositions of the pool still requires further investigations. Nevertheless, the proposed framework provides an effective strategy of fusing the expert knowledge without any predetermined weighting mechanism. Future directions of this work include: (i) adding more features. With the ongoing research in human auditory perception, more features could be added to help the system capture more perceptually relevant information, (ii) investigating different methods for aggregating the learned tasks. In addition to geometric mean, more sophisticated methods such as linear regression could be used as a late-fusion strategy to better combine the learned tasks, (iii) exploring different architectures. Particularly, models such as Recurrent Neural Networks (RNNs) or Temporal Convolutional Networks (TCNs) could potentially leverage the temporal information and lead to improvements, and (iv) verifying the generality of the proposed framework for assessing the quality of other audio processing methods such as noise reduction and source separation.

References

[1] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011.

[2] M. Erne, "Perceptual Audio Coders "What to listen for"," in *Proceedings of the Audio Engeering Society (AES) convention*, 2001.

[3] S. Zieliński and F. Rumsey, "On Some Biases Encountered in Modern Audio Quality Listening Tests—A Review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[4] A. W. Rix, J. G. Beerends, D.-S. Kim, P. Kroon, and O. Ghitza, "Objective Assessment of Speech and Audio Quality—Technology and Applications," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1890–1901, Nov. 2006.

[5] D. Campbell, E. Jones, and M. Glavin, "Audio quality assessment techniques—A review, and recent developments," *Signal Processing*, vol. 89, no. 8, pp. 1489–1500, Aug. 2009.

[6] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PEAQ–The ITU Standard for Objective Measuremen of Perceived Audio Quality," *Journal of the Audio Engineering Society*, vol. 48, no. 1, pp. 3–29, 2000.

[7] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual Objective Listening Quality Assessment (POLQA), The Third Generation ITU-T Standard for End-to-End Speech Quality Measurement Part I–Temporal Alignment," *Journal of the Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[8] C. Sloan, N. Harte, D. Kelly, A. C. Kokaram, and A. Hines, "Objective Assessment of Perceptual Audio Quality Using ViSQOLAudio," *IEEE Transactions on Broadcasting*, vol. 63, no. 4, pp. 693–705, Dec. 2017.

[9] P. Počta and J. G. Beerends, "Subjective and Objective Assessment of Perceived Audio Quality of Current Digital Audio Broadcasting Systems and Web-Casting Applications," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 407–415, Sep. 2015.

[10] S. Kum, J.-H. Lin, L. Su, and J. Nam, "Semi-supervised learning using teacher-student models for vocal melody extraction," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) conference*, 2020.

[11] Y.-N. Hung, Y.-A. Chen, and Y.-H. Yang, "Multitask learning for frame-level instrument recognition," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[12] S. Böck, M. E. Davies, and P. Knees, "Multi-task learning of tempo and beat: learning one to improve the other," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) conference*, 2019.

[13] I. T. Union, "Recommendation ITU-T P.563: Single-ended method for objective speech quality assessment in narrow-band telphony applications," 2004.

[14] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2019.

[15] I. T. Union, "Recommendation ITU-T P.861 Objective quality measurement of telephone-band (300 - 3400 Hz) speech codecs," 1996.

[16] 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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[17] J. M. Kates, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[18] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 13, pp. 1–18, 2015.

[19] R. Huber and B. Kollmeier, "PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 1902–1911, Nov. 2006.

[20] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, Jun. 2015.

[21] J. M. Kates and K. H. Arehart, "The Hearing-Aid Audio Quality Index (HAAQI)," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 354–365, 2016.

[22] T. Biberger, J.-H. Fleßner, R. Huber, and S. D. Ewert, "An Objective Audio Quality Measure Based on Power and Envelope Power Cues," *Journal of the Audio Engineering Society*, vol. 66, no. 7/8, pp. 578–593, Aug. 2018.

[23] M. Torcoli and S. Dick, "Comparing the Effect of Audio Coding Artifacts on Objective Quality Measures and on Subjective Ratings," in *Proceedings of the Audio Engeering Society (AES) convention*, 2018.

[24] T.-P. Chen and L. Su, "Functional harmony recognition of symbolic music data with multi-task recurrent neural networks," in *Proceedings of the International Society for Music Information Retrieval (ISMIR) conference*, 2018.

[25] R. Caruana, "Multitask Learning," *Machine learning*, vol. 28, p. 35, 1997.

[26] K. Brandenburg and T. Sporer, ""NMR" and "Masking Flag": Evaluation of Quality Using Perceptual Criteria," in *Proceedings of the international conference of Audio Engeering Society (AES)*, 1992.

[27] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] R. A. Alexander, "A note on averaging correlations," *Bulletin of the Psychonomic Society*, vol. 28, no. 4, pp. 335–336, Oct. 1990.

[29] H. B. Martinez and M. C. Q. Farias, "A No-reference Audio-visual Video Quality Metric," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2014.