Interspeech 2022
18-22 September 2022, Incheon, Korea

TRANSFORMER-BASED QUALITY ASSESSMENT MODEL FOR GENERALIZED USER-GENERATED MULTIMEDIA AUDIO CONTENT

Deebha Mumtaz[1], *Ajit Jena*[1], *Vinit Jakhetiya*[1], *Karan Nathwani*[1], *Sharath C. Guntuku*[2]

[1]IIT JAMMU, [2]University of Pennsylvania

(2018rcs0011, 2020pcs2021, vinit.jakhetiya, karan.nathwani)@iitjammu.ac.in,
sharathg@seas.upenn.edu

Abstract

In this paper, we propose a computational measure for the quality of audio in user-generated multimedia (UGM) in accordance with the human perceptual system. To this end, we first extend the previously proposed IIT-JMU-UGM Audio dataset by including samples with more diverse context, content, distortion types, and intensities, along with implicitly distorted audio that reflect realistic scenarios. We conduct subjective testing on the extended database containing 2075 audio clips to obtain the mean opinion scores for each sample. We then introduce transformer-based learning to the domain of audio quality assessment, which is trained on three vital audio features: Mel-frequency cepstral coefficients, chroma, and Mel-scaled spectrogram. The proposed non-intrusive transformer-based model is compared against state-of-the-art methods and found to outperform Simple RNN, LSTM, and GRU models by over 4%. The database and the source code will be made public upon acceptance.

Index Terms: Non-intrusive Audio Quality Assessment, Transformer-based Learning, User-generated Multimedia.

1. Introduction

The availability of portable devices such as digital cameras and smartphones, as well as information sharing technology has resulted in the explosive generation of user-generated multimedia (UGM) content. Post COVID-19 outbreak, digital multimedia has further influenced every aspect of most individual's lives, be it work, e-learning, entertainment, social media, communication, speech-based human-machine interaction, etc. However, the diversity in capturing devices, varying bandwidth network, background disturbances, etc., highly influence the quality of UGM. This, in turn, affects the user's quality of experience (QoE), which motivates service providers and audio researchers to provide optimal and reliable solutions to maximize consumer's QoE. Audio and speech quality assessment techniques are categorized into two main types: a) intrusive methods, which require both reference as well as test/distorted signal, and b) non-intrusive methods that determine audio quality based only on the implicit properties of a given signal. In their work, Rix *et al.* [1] developed an intrusive quality evaluation technique known as Perceptual Evaluation of Speech Quality (PESQ), for narrow-band speech signals transmitted over telecommunication channels. The Perceptual Objective Listening Quality Assessment (POLQA) was proposed in [2] to predict the quality of narrow-band along with super-wide band signals by employing masking functions and projecting them in the frequency domain. Hines *et al.* [3], explored the spectro-temporal similarity between reference and distorted signal to obtain the quality of speech signals. Another study by Taal *et al.* [4] explored time-frequency domain properties to determine audio quality intrusively. The authors in [5] determined the quality score of reverberated and de-reverberant signals by employing spectral modulation. Lately, Chen *et al.* [6] proposed recurrent neural network-based metric for the assessment of voice conversion. Further, work done in [7] proposed a model for determining the quality of separated audio source signals using timbre features. In [8], Joan *et al.* proposed a semi-supervised learning method to determine the quality of speech signals to deal with the lack of annotated dataset. The authors in [9] proposed a model based on BLSTM and 1D-CNN to determine the quality of speech. Some other recent quality assessment metrics include [10–16]. As analyzed in this study, most of the audio quality estimation metrics focus on speech signals, while, the domain of UGM audio signals is less explored. These signals consist of not only human voice but a variety of sound types (such as music, background sounds, songs, environment sounds, noise, added sound effects, etc.), each pertaining to their context suggesting audio quality assessment of UGMs is different from prior works [17].

In [18], authors studied the effect of various distortions in the UGM, such as wind noise, handling noise, etc. However, the dataset consisted of limited content, and size (128 audio samples) and was based on multiple regression model for quality assessment. Further, in previous work [17], IIT-JMU-UGM Audio Dataset was proposed, which consisted of 1150 UGM audio clips and used a gated recurrent unit (GRU) for predicting the perceptual quality of audio samples. Though the paper proposed a successful benchmark for UGM, there are certain shortcomings of this dataset. First, the dataset consists of mainly two types of distortions; low bit rate and background noise. Second, only about 10% of the samples implicitly consist of distortions with limited diversity in context. Consequently, to make a more generalized and diverse audio dataset consisting of several real-world scenarios, we extend the dataset by including more real-life distortions and contexts. The major contributions of the proposed work are as follows:

1. The proposed IIT-JMU-UGM Audio Dataset-2 consists of more diverse, real-world scenarios and is created by incorporating clips with different contexts, content, distortion types, and intensities, implicitly distorted audio along with the IIT-JMU-UGM Audio Dataset. The newly created dataset has 2,075 audio samples along with their respective subjective scores.
2. The proposed non-intrusive audio quality technique introduces the application of transformer-based learning in the domain of audio quality assessment and achieves over 4% better performance in comparison to the existing state-of-the-art algorithm.

This work is supported by the CRG-SERB Scheme (Project No. CRG/2018/003920) and Microsoft Research Travel Grant.

10.21437/Interspeech.2022-10386

2. Proposed Method

In this work, we propose a non-intrusive audio quality assessment technique based on the transformer learning model for assessing the quality of UGM audio data. Firstly, we develop the repository called IIT-JMU-UGM Audio Dataset-2. Next, we propose a model consisting of two parts. The first part is the feature extraction block, which is followed by a transformer-based deep learning module for predicting audio quality.

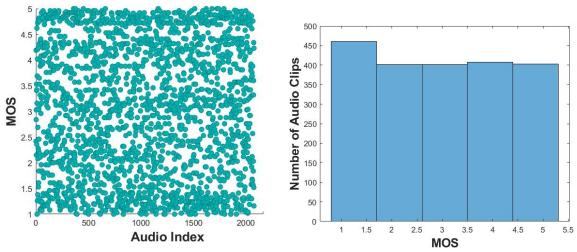
2.1. Data Preparation

The earlier proposed IIT-JMU-UGM Audio Dataset consists of 1150 audio samples. These samples contain two main types of distortions caused by low bit rate and the inclusion of background noise. In order to include more distortions that are present in real-world scenarios, we extend the database as IIT-JMU-UGM Audio Dataset-2.

In extending the database, we focused on both the implicit distortions as well as the synthetically/explicitly added distortions. For implicit distortion, we specifically searched samples with varying degrees of distortion. The degradation in such samples was caused due to various factors such as high compression due to repeated uploading of samples, captured in the noisy background (e.g., restaurant, stadium, market), captured in low bandwidth conditions (e.g., zoom-meeting, video-games, live streaming), clips from old records (e.g., interviews, movies, songs, radio-recordings), captured by non-professionals, include synthetic sound effects (e.g., comedy clips), etc. Moreover, these clips were recorded from different devices such as digital recorders, mobile phones, tabs, web cameras, etc., with various recording capabilities. Also, in terms of context, other types were included, such as online gaming, synthetic voice, Tiktok, Zoom meetings, online classes, online games, advertisements, marketing, walkie-talkie conversations, text-to-speech, synthetic voice, children rhymes, comedy videos, etc. Further, we explicitly distorted the clean/good quality clips using compression, clipping, gain, filtering, different annoying background noises at different signal-to-noise ratios, reverberation, and their combinations.

Initially, we collected around six hundred multimedia samples from various websites such as YouTube, Flickr, Tiktok, etc. Next, the clips were filtered in order to have a diverse yet equal representation of various context types. Further, these were clipped to an average of 8 sec time. For synthetically distorted clips, the distortion types and strengths were manually adjusted so as to make sure that the samples were separated by different perceptual levels of distortion. These impairments significantly affect the quality of UGM clips, which subsequently change their perceptual quality. The final IIT-JMU-UGM Audio Dataset-2 consists of 2075 audio samples containing various degrees and types of real-world distortions enveloping wide content variations.

In order to have the subjective scores corresponding to these 2075 audio clips, 26 subjects took part in the subjective testing with the age range between 18 to 40 years, and these subjects did not have any known hearing disability. To retain compatibility, we mimic the testing methodology of the previous IIT-JMU-UGM Audio Dataset. An initial session for training was organized where a small collection of audio samples with different perceptual quality was presented to the subjects. This helped the candidates get familiar with the task and train them for subjective scoring. Additionally, their demographic information was collected. Finally, each candidate was presented with the audio clip and instructed to access the quality as per the Absolute Category Rating (ACR) scale of 1 (very poor) to 5 (excellent). In order to reduce the effect of fatigue, the test was conducted in a series of sessions. For each clip, the mean opinion score (MOS) or final subjective score is obtained by taking the average of the subjective rating given by the subjects over each clip.



(a) (b)

Figure 1: *(a) Scatterplot depicting MOS, and (b) histogram representing the quantized MOS over five bins representing the different audio quality in the IIT-JMU-UGM Audio Dataset-2.*

The Fleiss' Kappa inter-rated reliability score for the subjective testing on IIT-JMU-UGM Audio Dataset was reported as 0.405; on the newly added audio samples, it was 0.442, and on the whole database, i.e., IIT-JMU-UGM Audio Dataset-2 it is 0.425. These scores are transcribed as there being a "Moderate agreement among the subject". Additionally, figure 1(a) depicts the scatter plot of the MOS of the various audio samples in the dataset Fig. 1(b) represents the histogram obtained by quantizing the MOS over five bins pertaining to different audio quality. From these figures, it can be observed that the proposed database covers the complete spectrum of perceptual quality scores.

2.2. Feature Extraction

In order to evaluate the audio signals, we need to extract the most prominent features that represent the perceptual qualities of these clips. Hand-crafted descriptors have the advantage of being compact, computationally efficient, and do not require a large training set. With this view, we made use of three hand-crafted acoustics features, which include Mel-Frequency Cepstral Coefficients (MFCCs), chroma, and Mel-scaled spectrogram. The Mel spectrogram represents a spectrogram where the frequencies are converted to the Mel scale, which provides a linear scale for the human auditory system. The Mel spectrogram has been immensely used in applications such as [19]. The MFCCs features are known to show a good correlation with the human auditory system. Consequently, they have been employed in tasks such as speaker recognition, quality assessment, etc. [20, 21]. The chroma features project the semitones of the music octave for each time frame. These have shown to give good results in music and audio structure processing, and analysis applications [22, 23]. More details about these features can be obtained from [24]. In the proposed model, for each audio clip, 12 chroma features, 40 MFCC coefficients, and 128 Mel spectrogram features were extracted. All these features were then concatenated to a single 1-D feature vector with size equivalent to 180 corresponding to each audio sample. For feature extraction, Python's Librosa library was used, the sample rate was set to default, hop length and the window length (fftsize) were set to 512 and 2048, respectively.

2.3. Model Architecture

In this work, we have introduced transformer-based learning for audio quality assessment. The detailed architecture of the proposed model is shown in Fig. 2. It takes as input a vector consisting of the three concatenated hand-crafted features along with the corresponding ground-truth label for each clip for training the network. The transformer-based deep learning architectures [25] are modeled to process sequential input data as done by the recurrent neural networks (RNNs). However, instead of processing the input data in sequential order, they employ the "attention mechanism" to give context to each part of the data. Such a mechanism helps to provide more parallelization and give better performance. These models have recently shown superior performance in the application of computer vision [26] and natural language processing [27–29]. The transformer model mainly consists of an Encoder-Decoder structure. First, the input sequence of the given data is embedded into an n-dimensional vector space. Next, the input embedding is passed through a positional encoder which keeps track of the order or relative position of each part of the sequence. These positions are concatenated with the existing n-dimensional vector and fed into the encoder. The encoder is mainly composed of two segments, the first is the Multi-Head Attention (MHA), and the second is the Feed Forward layers. The MHA mechanism consists of a number of scaled dot-product attention units. Given a sequence vector, the attention unit computes the embeddings containing the contextual information about the particular token as well as the weighted combination with respect to similar tokens. During the training, three weight matrices, corresponding to key weights W_K, value weights W_V, and query weights W_Q, are learned by the attention unit. Finally, the $Attention$ for all tokens is obtained as [25]:

$$Attention(Q, K, V) = softmax\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right) \times V \quad (1)$$

The matrices Q, K and V are the matrices where the i^{th} rows are vectors q_i, k_i, and v_i respectively. In order to stabilize gradients during training, the attention weights are divided by the square root of the dimension of the key vectors $\sqrt{d_k}$. The softmax function is used for weight normalization for better optimization. Next, every output encoding is separately processed by the feed-forward neural network. Subsequently, the output encodings are fed to the next encoder and the decoders.

In this paper, we propose the use of only the encoder module of the transformer, wherein instead of feeding the output of the encoder to the decoder, it is directly fed to a fully connected layer. This mainly helps to optimize the feature vector while taking 'attention' into consideration. We configured the model for four layers of the encoder (Layer 1 to 4), the number of heads (h) in each MHA was set to four, and the number of neurons equal to 2048. In addition, we applied Adam optimizer, which exploits the benefits of both AdaGrad and RMSProp and achieves optimal results at a fast rate. We used the mean square error as the loss function, and the initial learning rate of the Adam optimizer equivalent to $3{\times}10^{-6}$. The output vector from the last encoder (Layer 4) was reshaped before sending it to the final dense layer. The dense layer has a single output unit without an activation function as the prediction of the audio quality is required to be a continuous numerical value in the range of 1 to 5 (bad to excellent). Also, in order to have generalized scores, we performed an 80/20 training/test split and

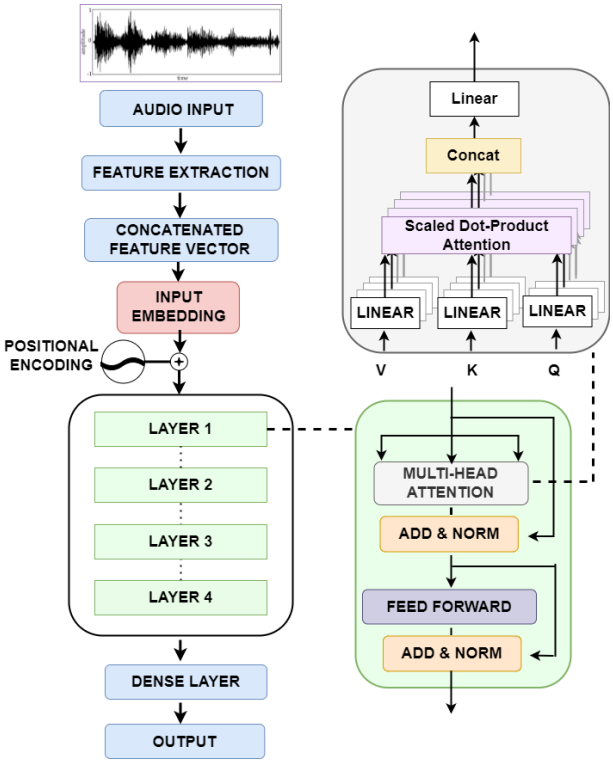


Figure 2: *Architecture of the proposed model.*

k-fold cross-validation over the database, where $k = 5$. The architecture was coded in Python using the Tensorflow and Keras libraries on Google Colab Nvidia GPU.

3. Results

To assess the performance of the proposed method with the available state-of-the-art quality assessment techniques, we make use of some of the well-known co-relation evaluation metrics. These include Spearman's Rank Order Correlation Coefficients (SRCC), Kendall's Correlation Coefficients (KRCC), and Pearson Linear Correlation Coefficients (PLCC). A higher value of PLCC, SRCC, and KRCC (upper range of 1) depicts a higher correlation between the subjective and the objective scores. We compared the proposed model with various existing audio quality assessment algorithms whose source codes were openly available, such as WAWEnets [30], MOSNET [6], SRMR [5], NIST-STNR [31], WADA [32], SNRVAD [33], NISQA [34], and UGM-GRU [17]. For a fair comparison, the previously proposed UGM-GRU model was re-trained on the IIT-JMU-UGM Audio Dataset-2. Since the proposed database consists of many real-world implicitly distorted clips whose pristine reference clips were not available, we compared the proposed model with only non-intrusive audio quality assessment techniques. As illustrated from Table 1, the proposed metric outperforms the existing quality metrics with the PLCC, SRCC, and KRCC is equivalent to 0.816, 0.812, and 0.613 respectively on the proposed IIT-JMU-UGM Audio Dataset-2. Conversely, the next best-performing model i.e. UGM-GRU had a PLCC equal to 0.754.

Next, an F-Test was conducted to determine the statistical significance of the proposed metric with respect to the other techniques. This test is based upon the variance hypothesis,

Table 1: *Comparison of performance (in terms of PLCC, SRCC, and KRCC) of the proposed model against other quality techniques.*

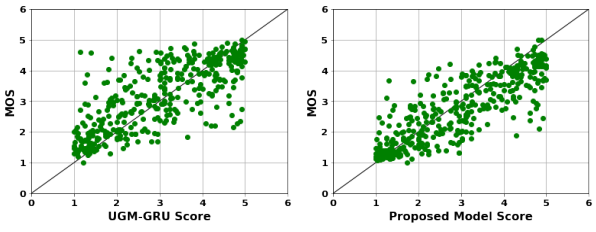
Metric	PLCC	SRCC	KRCC
Proposed Model	**0.816**	**0.812**	**0.613**
UGM-GRU [17]	0.754	0.746	0.554
WAWEnets [30]	0.433	0.414	0.287
NISQA [34]	0.415	0.383	0.268
NIST-STNR [31]	0.323	0.280	0.190
MOSNET [6]	0.322	0.312	0.209
SRMR [5]	0.272	0.285	0.193
SNRVAD [33]	0.270	0.269	0.180
WADA [32]	0.284	0.258	0.175

whose result, when equal to '0' means that the proposed method is statistically comparable to the other metric, '+1' represents that the proposed method is better, and '-1' denotes that proposed method is worse than the other metric [35]. Table 2 gives the F-score between the proposed metric and the other metrics. As analyzed from this table, the F-score is equivalent $+1$ with respect to all the other metrics. Thus, denoting that our proposed method is statistically better than the rest of the available metrics.

Table 2: *Statistical significance comparison of the proposed metric with respect to the other metrics.*

Metric	[17]	[30]	[34]	[31]	[6]	[5]	[33]	[32]
F-Score	+1	+1	+1	+1	+1	+1	+1	+1

Figure 3 illustrates the scatter plots between the MOS and the objective scores obtained from the (a) second-best performing UGM-GRU metric and (b) the proposed model. In order to have better visualization of the results, we plot only 400 random samples from the database. The diagonal line depicts the ideal situation wherein the objective scores are equivalent to the MOS. As depicted from these plots, there is better linearity between the MOS and the scores obtained from the proposed metric compared to the score obtained from the UGM-GRU metric. This depicts that the proposed model better correlates with human perception.



(a) (b)

Figure 3: *Scatterplot between the mean opinion score (MOS) and the objective scores predicted by the (a) UGM-GRU [17] and, (b) proposed model, on the IIT-JMU-UGM Audio Dataset-2.*

Moreover, we conducted an ablation study wherein we utilized different RNN models such as RNN, LSTM, and GRU. These RNN networks were fed with the concatenated features consisting of chroma, Mel spectrogram, and MFCC. In each of these models, we used a stacked architecture along with intermediate dropout layers. The final result was obtained by k-fold validation. For all these models, the Adam optimizer was used. As shown in Table 3, the GRU and LSTM models have comparable performance, however, the proposed transformer-based model outperforms all of these models. This ablation study also validates the applicability of transformers in the realm of UGM audio quality assessment.

Table 3: *Ablation study of the proposed algorithm against various RNN models.*

Metric	PLCC	SRCC	KRCC
Proposed Model	**0.816**	**0.812**	**0.613**
GRU	0.771	0.764	0.565
LSTM	0.780	0.778	0.578
Simple RNN	0.435	0.440	0.306

In order to show that the transformer-based proposed model doesn't significantly depend upon the number of layers of encoder, we have also conducted an ablation study. The performance of the metric when the number of layers in transformers is varied is shown in Table 4. The number of heads in all the ablation was set to 4 and all the other parameters were kept constant. From this table, it can be seen that the proposed model gives the best results with the number of encoder layers equal to 4. Also, the overall performance of the transformer on any of the different number of layers is still producing better performance as compared to the second-best performing technique UGM-GRU. This confirms that the model is robust to hyperparameter optimization.

Table 4: *Ablation study of the proposed algorithm against various number of encoder layers.*

Metric	PLCC	SRCC	KRCC
Transformer (1 layer)	0.802	0.797	0.598
Transformer (2 layers)	0.808	0.803	0.603
Transformer (4 layers)	**0.816**	**0.812**	**0.613**
Transformer (6 layers)	0.801	0.796	0.598

4. Conclusion

User-generated content is one of the most prevalent multimedia types in the present age. Consequently, assessing the perceptual quality assessment of UGM in an automated and non-intrusive manner could benefit different stakeholders, starting with the content producers to streaming companies, and also the end users/audience. Compared to the previous quality metrics, our contribution is two-fold; first, we created an extended IIT-JMU-UGM Audio Dataset-2 comprising 2075 audio samples (with subjective scores), which are more diverse in terms of content, context, and types of distortions to mimic real-world scenarios. We have also included the implicit distortions in the proposed dataset. Secondly, we then propose an end-to-end non-intrusive transformer-based metric that objectively determines the perceptual quality of UGM audio clips. The results show significant improvements compared to existing state-of-the-art algorithms developed for speech quality assessment and correlate well with human perception.

5. References

[1] 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 *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 2001, pp. 749–752.

[2] J. 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 II-Perceptual Model," *AES: Journal of the Audio Engineering Society*, vol. 61, pp. 385–402, 06 2013.

[3] 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. 1, pp. 1–18, 2015.

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[5] T. H. Falk, C. Zheng, and W. Chan, "A Non-Intrusive Quality and Intelligibility Measure of Reverberant and Dereverberated Speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1766–1774, 2010.

[6] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning based objective assessment for voice conversion," in *Proc. Interspeech 2019*.

[7] T. Kastner and J. Herre, "An efficient model for estimating subjective quality of separated audio source signals," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 95–99.

[8] J. Serrà, J. Pons, and S. Pascual, "Sesqa: Semi-supervised learning for speech quality assessment," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 381–385.

[9] Q. Huang and T. Hain, "Exploration of Audio Quality Assessment and Anomaly Localisation Using Attention Models," in *Proc. Interspeech 2020*, pp. 4611–4615.

[10] N. Nessler, M. Cernak, P. Prandoni, and P. Mainar, "Non-Intrusive Speech Quality Assessment with Transfer Learning and Subject-Specific Scaling," in *Proc. Interspeech 2021*, pp. 2406–2410.

[11] M. Yu, C. Zhang, Y. Xu, S.-X. Zhang, and D. Yu, "MetricNet: Towards Improved Modeling For Non-Intrusive Speech Quality Assessment," in *Proc. Interspeech 2021*, pp. 2142–2146.

[12] X. Dong and D. S. Williamson, "A Pyramid Recurrent Network for Predicting Crowdsourced Speech-Quality Ratings of Real-World Signals," in *Proc. Interspeech 2020*, pp. 4631–4635.

[13] M. Torcoli, T. Kastner, and J. Herre, "Objective measures of perceptual audio quality reviewed: An evaluation of their application domain dependence," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1530–1541, 2021.

[14] J.-H. Fleßner, T. Biberger, and S. D. Ewert, "Subjective and objective assessment of monaural and binaural aspects of audio quality," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1112–1125, 2019.

[15] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 331–335.

[16] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 626–630.

[17] D. Mumtaz, V. Jakhetiya, K. Nathwani, B. N. Subudhi, and S. C. Guntuku, "Non-intrusive perceptual audio quality assessment for user-generated content using deep learning," *IEEE Transactions on Industrial Informatics*, pp. 1–1, 2021.

[18] B. M. Fazenda, P. Kendrick, T. J. Cox, F. Li, and I. Jackson, "Perception and automated assessment of audio quality in user generated content: An improved model," in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, 2016, pp. 1–6.

[19] G. Mittag and S. Möller, "Deep Learning Based Assessment of Synthetic Speech Naturalness," in *Proc. Interspeech 2020*, pp. 1748–1752.

[20] A. Chowdhury and A. Ross, "Fusing mfcc and lpc features using 1d triplet cnn for speaker recognition in severely degraded audio signals," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1616–1629, 2020.

[21] Y. Shan, J. Wang, X. Xie, L. Meng, and J. Kuang, "Non-intrusive speech quality assessment using deep belief network and backpropagation neural network," in *International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 11 2018, pp. 71–75.

[22] F. Zalkow and M. Müller, "Ctc-based learning of chroma features for score–audio music retrieval," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2957–2971, 2021.

[23] K. O'Hanlon and M. B. Sandler, "Comparing cqt and reassignment based chroma features for template-based automatic chord recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 860–864.

[24] G. Sharma, K. Umapathy, and S. Krishnan, "Trends in audio signal feature extraction methods," *Applied Acoustics*, vol. 158, p. 107020, 2020.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, 2017, p. 6000–6010.

[26] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal transformer with multi-view visual representation for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4467–4480, 2020.

[27] Y. Shi, Y. Wang, C. Wu, C. Fuegen, F. Zhang, D. Le, C.-F. Yeh, and M. L. Seltzer, "Weak-Attention Suppression for Transformer Based Speech Recognition," in *Proc. Interspeech 2020*, pp. 4996–5000.

[28] C. Wu, Z. Xiu, Y. Shi, O. Kalinli, C. Fuegen, T. Koehler, and Q. He, "Transformer-Based Acoustic Modeling for Streaming Speech Synthesis," in *Proc. Interspeech 2021*, pp. 146–150.

[29] R. Zhang, H. Wu, W. Li, D. Jiang, W. Zou, and X. Li, "Transformer based unsupervised pre-training for acoustic representation learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6933–6937.

[30] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 331–335.

[31] NIST. [Online]. Available: https://labrosa.ee.columbia.edu/ dpwe/tmp/nist/doc/stnr.txt

[32] C. Kim and R. Stern, "Robust signal-to-noise ratio estimation based on waveform amplitude distribution analysis," in *Proc. Interspeech*, 2008, pp. 2598–2601.

[33] SNRVAD. [Online]. Available: https://labrosa.ee.columbia.edu/projects/snreval/

[34] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proc. Interspeech*, 2021, pp. 2127–2131.

[35] Sadbhawna, V. Jakhetiya, S. Chaudhary, B. N. Subudhi, W. Lin, and S. C. Guntuku, "Perceptually unimportant information reduction and cosine similarity-based quality assessment of 3d-synthesized images," *IEEE Transactions on Image Processing*, vol. 31, pp. 2027–2039, 2022.