Interspeech 2022
18-22 September 2022, Incheon, Korea

Soft-label learn for No-Intrusive Speech Quality Assessment

Junyong Hao, Shunzhou Ye, Cheng Lu, Fei Dong, Jingang Liu, Dong Pi

junyong.hao@unisoc.com, jingang.liu@unisoc.com,
dong.pi@unisoc.com

Abstract

Mean opinion score (MOS) is a widely used subjective metric to assess the quality of speech, and usually involves multiple human to judge each speech file. To reduce the labor cost of MOS, no-intrusive speech quality assessment methods have been extensively studied. However, due to the highly subjective bias of speech quality label, the performance of models to accurately represent speech quality scores is difficult to be trained. In this paper, we propose a convolutional self-attention neural network (Conformer) for MOS score prediction of conference speech to effectively alleviate the disadvantage of subjective bias on model training. In addition to this novel architecture, we further improve the generalization and accuracy of the predictor by utilizing attention label pooling and soft-label learning. We demonstrate that our proposed method achieves RMSE cost of 0.458 and PLCC score of 0.792 on evaluation test datasets of Conferencing Speech 2022 Challenge.

Index Terms: speech quality, conformer, subjective bias

1. Introduction

With the rapid development of hands-free devices and online conferencing applications, we can easily carry out real-time voice communication. However, speech quality is often affected by environmental noise, room reverberation, interference, network packet loss etc., which has caused obstacles to voice communication. In order to monitor the speech quality of the communication network, online system provider needs a tool that can detect the communication quality [1]. Therefore, how to assess the speech quality of an online system in real time and objectively is an urgent problem to be solved.

Speech assessment metrics are currently mainly divided into two categories. One is intrusive speech quality evaluation, and these metrics need clean reference signals. Earlier, some assessment metrics such as signal-to-noise ratio (SNR) [2], signal-to-interference ratio (SIR), segmental signal-to-noise ratio (SSNR) [3], and signal-to-distortion ratio (SDR) were used as an assessment metrics to evaluate speech quality. Although these methods can well show the difference between degraded speech and clean speech, they cannot fully demonstrate speech quality at the sense of hearing level. Hence, perceptual evaluation of speech quality (PESQ) [4, 5, 6] proposed by the International Union as an end-to-end speech quality assessment metrics. Subsequently, the perceptual objective listening quality assessment (POLQA) [7] was introduced as an iterative version of PESQ, and showed strong correlation with subjective assessment. Nevertheless, the application of intrusive methods is restricted because of harsh conditions requiring a clean reference, especially in real-time speech communication scenarios. Another type of metrics is non-intrusive speech quality assessment, which is aimed at scenarios where pure reference cannot be obtained. It only needs degraded audio for speech quality assessment, which is more suitable for practical scenarios. The P.563 [8] was proposed by ITU-T to assess speech quality in scenarios where only degraded audio is unavailable, but has limitation which is used for narrowband (NB) speech signals only. ITU-T P.808 [9] is a preferred solution to break through the limitations of P.563. Unfortunately, the inherent high cost of this method is often unacceptable.

In order to alleviate the high cost of subjective speech assessment, many deep learning methods have been applied to Non-Intrusive Speech Assessment. These models are trained to predict subjective assessment scores or objective evaluation scores. AutoMOS [10] achieves correlations approaching those of human raters with a long-short-term-memory (LSTM). Quality-Net [11] uses a bidirectional-long-short-term-memory (BLSTM) and proposes a method of frame constraints and sensible initializations to predict MOS. MOSA-Net [12] composed of a convolutional neural network and a bidirectional long short-term memory (CNN-BLSTM) with an attention mechanism, show their capability to model human perception. In [13], Mittag proposes a deep learning model NISQA that predicts speech quality of super-wideband (SWB) speech samples. In addition, the multi-task learning criterion that simultaneously optimizes multiple metrics is used to train the assessment model. They also have improved speech assessment performance by a CNN-Self-Attention-Attention-Pooling (CNN-SA-AP) structure.

As stated above, the mainstream of no-intrusive speech quality assessment approaches in academia and industry is to use the mean opinion score as the training target. In the evaluation process of speech sample MOS, in order to ensure the evaluation quality of speech dataset, different speech samples are evaluated by multiple judges. MOS of a speech sample is simply the average of the scores of multiple judges, which inevitably introduces personal characteristic. Subjective bias makes it difficult for models to learn true speech quality. MBNet [14] proposes a method by learning a mean subnet and a bias subnet structure to better utilize every judge, and achieves better performance than MOSNet and its variant models [15, 16]. Moreover, DNSMOS [17] proposes a multi-stage self-teaching method to eliminate subjective bias in the process of model training.

Based on these methods, we propose a method named Soft-label Learning Conformer (SLL-Conformer) for no-intrusive speech quality assessment. Furthermore, we use attention label pooling to integrate local prediction into global prediction, and use soft-label learning to alleviate the bias caused by subjective MOS in model training.

2. Proposed Method

In this section, we introduce the proposed methods for speech quality assessment. We first describe the framework of our

Copyright © 2022 ISCA

10.21437/Interspeech.2022-10400

neural networks, and then introduce the method of attention label pooling and soft-label learning for training speech quality assessment model.

2.1. Backbone network

Conformer neural network, which consists of several cascaded convolutional layers and self-attention layers, is the representative model including speech quality assessment and other tasks. We assume that both global and local interactions are important for speech quality assessment. Therefore, Conformer architecture is adopted as the backbone network in our SLL-Conformer model. In this paper, the overall framework of our model is shown in Figure 1. We employ a Conformer with 9 convolutional layers and 2 multi-headed-self-attention layers with 8 head as backbone feature extraction network. Finally, a classifier with attention label pooling outputs speech quality prediction results of the input Mel audio features. The total parameters of our model are 421220, and model size is 1.89M.

Resnet is one of the most well-known backbone networks in deep neural networks [18]. Compared with the previous network architecture, resnet introduces a shortcut connection to address the problem of vanishing gradient, and is able to extract rich semantics from the input data to represent feature information. In this paper, residual structure with CNN is used to extract discriminative embedding from 2D representations of audio. In detail, CNN network is stacked with 3 same structure CNN blocks, and a BatchNorm [19] layer acts on the data of each frame to normalize the input data before fed to CNN network.

Audio is a sequence signal, and the relevance of context information also has an important impact on speech quality. CNN is expert in extracting local features, but its ability is weak when extracting long sequences. Self-attention has shown strong extraction ability in series tasks [20, 21]. In this paper，multi-headed-self-attention layer with 8 head is used to obtain detailed context information of audio.

2.2. Attention label pooling

The subjective score of speech quality has strong context relevance, so the attention in time dimension is not the same when obtaining the score. Therefore, attention label pooling is used after backbone network to integrate frame level prediction to clip level prediction.

We integrate the prediction results of local output at the end of network. As shown in Figure 1, attention label pooling uses a full connection layer to obtain local prediction y_T, uses two full connection layers with softmax activation function to generate attention factors w_T. Finally, clip level speech quality prediction is integrated as follows:

$$y = \sum_T w_T \odot y_T \tag{1}$$

Where w_T is attention weight. y_T is local output, and T is the length of input audio. Like [13], the attention factors computed with masked use zero-padded in time dimension to yield normalized attention factors.

2.3. Soft label

We use Mean Squared Error (MSE) as the loss function like other similar regression methods. However, in the process of speech quality label acquisition (such as MOS score), a certain subjective bias is introduced. We propose to learn the

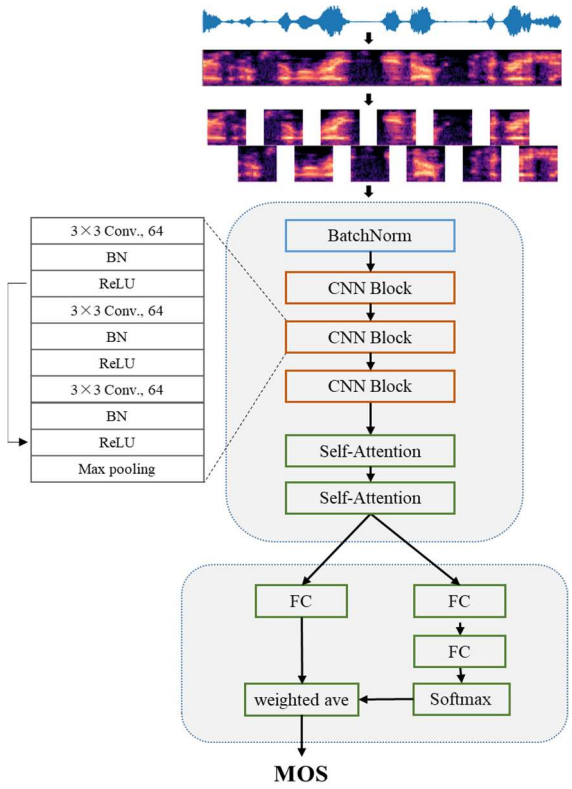


Figure 1: *The proposed framework of SLL-Conformer model for no-intrusive speech quality assessment*

subjective factors with the help of perturbations introduced by soft labels. Finally, the loss function used to train predicted MOS is:

$$loss_{mos} = MSE(p, y) + \lambda_1 MSE(p, y_{soft}) \tag{2}$$

Where p is predicted label, y is real label, y_{soft} is soft label, and we set $\lambda_1 = 0.1$ to balance the cost of real label and soft label applied to predicted label. In order to produce more accurate soft label, we adapt mean teacher [22] as the method of soft-label learning. Mean teacher is composed of two networks: teacher model and student model. Teacher model and student model have exactly the same network structure. Therefore, with the learning of network, teacher model can output the neighbor value of label predicted by student model, which brings disturbance in the process of student model training. The generated soft labels of teacher model can be in the vicinity of the real labels, so that the model can effectively learn the bias caused by subjective factors.

The mean teacher is designed for classification tasks [22], but speech quality assessment is a regression task. The prediction for classification tasks is discrete, which has large prediction tolerance; but regression task needs to predict continuous values. In order to adapt mean teacher to speech quality assessment, we made some improvements. Instead of adding a disturbance to the input of teacher model, the input data of teacher model and student model in this task are exactly the same. Particularly, the parameters of teacher model are updated by moving average and loss gradient:

$$\theta'_t = \alpha(\theta'_{t-1} - \beta \frac{\partial \mathcal{L}_{soft}}{\partial \theta'_{t-1}}) + (1 - \alpha)\theta_t \tag{3}$$

Where θ' and θ are parameters of teacher model and student model. t is current iterative step. α is the weight of moving average on student model and set in the light of [22]. $\mathcal{L}_{soft}$ is prediction cost of teacher model, which is calculated by:

$$\mathcal{L}_{soft} = MSE(y_{soft}, y) \quad (4)$$

We enable the teacher model to perform prediction loss calculation. Learning rate β set to 0.0001, which is 0.1 times that of student model. Finally, parameters of teacher model is saved and used for experimental analysis.

3. Experiment Settings

Experiments on *Conferencing Speech 2022-Non-Intrusive Speech Quality Assessment in Online Conferencing Application* challenge are conducted to demonstrate the performances of our SLL-Conformer methods. The dataset used in this challenge consists of multiple datasets.

3.1. Tencent dataset

The Tencent datasets from Conferencing Speech 2022 Challenge includes speech conditions with reverberation and without reverberation of about 14k Chinese speech samples. In without reverberation set, speech clips were artificially corrupted to simulate impairment (packet-loss, jitter, throttle) in realistic meeting scenario. In with reverberation set, 28% of speech clips created with simulated conditions to make a more diverse database. Each speech clip was processed with at least two types of impairment, meanwhile there was no same speech utterance with different impairments.

3.2. NISQA dataset

The NISQA dataset [13] contains seven sub-datasets. The NISQA_TRAIN_SIM and NISQA_VAL_SIM datasets contain 12500 speech samples with simulated speech distortions (e.g. noise, background, codes). The datasets NISQA_TRAIN_LIVE and NISQA_VAL_LIVE contain about 1,220 recorded speech samples in the scene of live phone calls and Skype calls. The NISQA_TEST_LIVETALK dataset is a collection of real-time voice samples. The datasets of NISQA_TEST_P501 (British) and NISQA_TEST_FOR (Australian) contain 480 speech samples simulating different codecs distortion, such as background noise, packet loss and newspaper clipping.

3.3. PSTN dataset

The PSTN dataset [23] (58709 speech samples) is degraded speech samples which are mixture of clean reference and background noise. Clean reference files derived from the public audiobook dataset Librivox, and background noise added to the clean files with an SNR between the clean files with an SNR between 0-40dB by using the DNS Challenge [24]. Noise clips from Audioset [25], DEMAND corpus [26], and Freesound.

By transforming one-dimensional time domain signal of sound into two-dimensional time-frequency domain signal through FFT, Mel feature of 48 dimensions are extracted as input to neural network. We use the raw audio sampling rate instead of the uniform sampling rate for all samples, and maximum sampling rate is set to half of the raw sampling rate when extracting Mel features. Model is trained with a learning rate of 0.001 of Adam optimizer, and mini-batch size of 25. Tencent, PSTN, and NISQA dataset is used in this paper. In order to better analyze the performance of our methods, we split 80% of Tencent and PSTN dataset into training set, and split NISQA dataset following the original way. Finally, training set group has 67237 files, and verification set group has 17467 files. The best model performance after multiple training runs is presented in Table 2 and Table 3, which report the averaged metrics including RMSE, spearman's rank correlation coefficient (SRCC) [27] and pearson's linear correlation coefficient (PLCC) [28]. All metrics calculated by the method provided by competition organizers.

4. Results

We conduct experiments to verify the effectiveness of our SLL-Conformer methods. First of all, an ablation experiment is implemented to verify the feasibility of using soft-label learning in speech quality assessment. Then, the performances of our SLL-Conformer model is compared with some other representative approaches.

Table 1 shows the results of verification set for soft-label learning. It can be seen that Conformer with soft-label learning (SSL-Conformer) achieves the best results. SSL-Conformer model is superior to Conformer model in all metrics, especially RMSE metric. This shows that soft-label learning effectively improves the ability of fitting MOS score.

Table 1: *Soft-label comparison results of verification set*

	SLL-Conformer	**Conformer**
RMSE	0.425	0.500
PLCC	0.887	0.866
SRCC	0.881	0.851

To further illustrate the performance of our SLL-Conformer model, we compare the speech quality assessment results of our proposed approach with several representative systems training set and verification set in Table 2. Baseline2 is a simple adaptation of the NISQA [13] network for this competition dataset. It should be pointed out that baseline2 does not split training set and verification set on Tencent dataset and PSTN dataset, and uses all datasets as train set for model training. In particularly, the NISQA model is trained on a larger set of 59 train sub-datasets (72903 files).

Our SLL-Conformer model and baseline2 carried out experimental analysis on the same competition dataset of Conferencing Speech 2022 Challenge. Compared to baseline2, our SLL-Conformer model outperforms better in almost all training and validation datasets. From the comparison between baseline2 and NISQA can be seen that NISQA model achieved better results with more training data. Therefore, more training data can improve the performance of speech quality assessment. As can been seen from the comparison between SLL-transformer model and NISQA model, SLL-transformer model can still surpass even if SLL-transformer model uses less data for training.

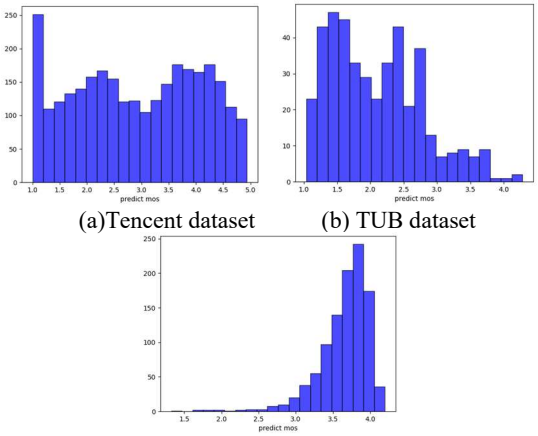
From the prediction performance of each dataset, it can be seen that TEST LIVETALK dataset is the worst result on all three models. TEST LIVETALK dataset is a recording of real telephone call, which makes its distribution different from other datasets. Even so, our SLL-Conformer model still shows strong robust on TEST LIVETALK dataset.

Table 2: *Results of each split set in terms of RMSE, PLCC, and SRCC*

Dataset	SLL-Conformer			Baseline2			NISQA[13]		
	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC	RMSE	PLCC	SRCC
VAL LIVE	**0.402**	**0.830**	**0.826**	0.475	0.813	0.809	0.403	0.820	0.823
VAL SIM	**0.456**	**0.913**	**0.910**	0.526	0.902	0.900	0.494	0.903	0.902
TEST P501	**0.429**	**0.928**	**0.926**	0.631	0.904	0.905	0.588	0.904	0.906
TEST FOR	**0.365**	**0.914**	**0.904**	0.553	0.863	0.854	0.411	0.891	0.876
TEST LIVETALK	**0.608**	0.778	**0.759**	0.710	**0.863**	0.713	0.618	0.778	0.755
VAL_With_Rever	0.268	0.958	0.947	/	/	/	/	/	/
VAL_Without_Rever	0.369	0.954	0.952	/	/	/	/	/	/
VAL_PSTN	0.501	0.822	0.825	/	/	/	/	/	/
TRAIN_LIVE	**0.272**	**0.921**	**0.920**	0.392	0.880	0.880	0.347	0.868	0.865
TRAIN_SIM	**0.279**	**0.967**	**0.967**	0.396	0.947	0.948	0.387	0.939	0.941
TRAIN_With_Rever	**0.183**	**0.98**	**0.976**	0.277	0.963	0.955	/	/	/
TRAIN_Without_Rever	**0.212**	**0.985**	**0.983**	0.424	0.963	0.960	/	/	/
TRAIN_PSTN	**0.474**	0.839	0.839	0.514	**0.843**	**0.843**	/	/	/

Table 3: *Results of all evaluation test datasets*

Dataset	SLL-Conformer			Baseline1[29]		
	PLCC	RMSE	RMSE after 3rd mapping	PLCC	RMSE	RMSE after 3rd mapping
MS	0.620	0.372	0.252	0.361	0.585	0.293
TUB	0.745	0.722	0.470	0.348	1.094	0.649
Tencent	0.956	0.349	0.344	0.881	0.624	0.55
average	0.792	0.458	0.329	0.530	0.768	0.497



(a)Tencent dataset (b) TUB dataset

(c) MS dataset

Figure2: *Predicted MOS distribution of evaluation test datasets*

Table 3 shows the results of our SLL-Conformer model and baseline1 on evaluation test datasets. The backbone network of baseline1 is CNN-BLSTM structure, which is also proposed in [13]. We can see that our proposed SLL-Conformer method is significantly better than baseline1 on all evaluation test datasets.

Figure 2 shows the predicted MOS distribution of evaluation test datasets. We can see that the predicted MOS distribution of Tencent dataset is relatively balanced, TUB dataset is lower and MS dataset is higher. The distribution difference among evaluation test datasets may also be an important factor in the generalization performance of SLL-Conformer model.

5. Conclusions

In this paper, we propose a Soft-label Learning Conformer (SLL-Conformer) method for no-intrusive speech quality assessment based on deep learning. In addition, we use attention label pooling to integrate local prediction into global prediction, and use soft-label learning to alleviate the bias caused by subjective MOS. Experiments on *Conferencing Speech 2022-Non-Intrusive Speech Quality Assessment in Online Conferencing Application* challenge show our proposed SLL-Conformer has strong robust on evaluation test datasets. We also find that the amount of data affects the performance of speech quality assessment, and the distribution difference between datasets poses a great challenge to the generalization ability of models. For future works, we will consider data augmentation and domain adaptation among heterogeneous datasets.

6. References

[1] V.Grancharov and W. B. Kleijn, "Speech quality assessment," pp. 83–100, 2008.

[2] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, 629–632.

[3] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. of ICSLP*, 1998, vol. 7, pp.2819–2822.

[4] ITU-T Rec. P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs."

[5] 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. Proceedings*, 2001, pp.749–752.

[6] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs."

[7] ITU-T Rec. P863, "Perceptual objective listening quality assessment."

[8] ITU-T Rec. P.563, "Single-ended method for objective speech quality assessment in narrow-band telephony applications."

[9] ITU-T Rec. P.808, "Subjective evaluation of speech quality with a crowdsourcing approach."

[10] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," in *Proc.of NIPS 2016 End-to-end Learning for Speech and Audio Processing Workshop*, 2016.

[11] 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 *Proc.of Interspeech*, 2018, pp. 1873–1877.

[12] 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.of Interspeech*, 2019, pp. 1541–1545.

[13] G. Mittag, B. Naderi, A. Chehadi, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *arXiv preprint arXiv*: 2104.09494, 2021.

[14] Y. Leng, X. Tan, S. Zhao, F. Soong, X.Y. Li, and T. Qin, "MBNET:Mos prediction for synthesized speech with mean-bias network," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 391–395.

[15] Y. Choi, Y. Jung, and H. Kim, "Neural mos prediction for synthesized speech using multi-task learning with spoofing detection and spoofing type classification," *arXiv preprint arXiv*: 2007.08267, 2020.

[16] Y. Choi, Y. Jung, and H. Kim, "Deep mos predictor for synthetic speech using cluster-based modeling," in *Proc.of Interspeech*, 2020, pp.1743-1747.

[17] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors." in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[19] S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, 2015, pp. 448-456.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *arXiv preprint arXiv*: 1706.03762, 2017.

[21] J.B. Cordonnier, A. Loukas, M. Jaggi, "On the relationship between self-attention and convolutional layers," in a*rXiv preprint arXiv*: 1911.03584, 2019.

[22] Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *In Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1195-1204.

[23] G. Mittag, R. Cutler, Y. Hosseinkashi, "DNN no reference PSTN speech quality prediction". arXiv: 2007.14598, 2020.

[24] C. K. A. Reddy, E. Beyrami, H. Dubey, V. Gopal, R. Cheng, R. Cutler, S. Matusevych, "The Interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," in *arXiv preprint arXiv*:2005.13981, 2020.

[25] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[26] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Proceedings of Meetings on Acoustics*, vol. 19, no. 1, p. 035081, 2013.

[27] C. Spearman, "The proof and measurement of association between two things," in *International Journal of Epidemiology*, vol. 39, no. 5, pp. 1137–1150, 2010.

[28] K. Pearson, "Notes on the history of correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 1920.

[29] G. Yi, W. Xiao, Y. Xiao and B Naderi,"ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications", https://arxiv.org/pdf/2203.16032.pdf, 2022.