

# Impairment Representation Learning for Speech Quality Assessment

Lianwu Chen, Xinlei Ren, Xu Zhang, Xiguang Zheng, Chen Zhang, Liang Guo, Bing Yu

Kuaishou Technology, Beijing, China

chenlianwu@kuaishou.com

## Abstract

Non-intrusive speech quality assessment has been a crucial task for speech processing. In recent years, methods based on deep neural network have achieved the start-of-the-art performance for non-intrusive speech quality assessment. However, the scarcity of annotated data is usually the main challenge for training robust speech quality assessment networks. In this paper, we proposed an impairment representation learning approach to pre-train the network on a large amount of simulated data without MOS annotation. Then we further fine-tune the pre-trained model for the MOS prediction task on annotated data. The experimental results show that the proposed pre-training methods can significantly improve the performance for speech quality assessment, especially when the annotated training data is limited. Besides, the proposed method significantly outperforms the baseline system of ConferencingSpeech 2022 Challenge.

**Index Terms:** MOS prediction, speech quality assessment, contrastive learning

## 1. Introduction

Speech quality assessment aims to evaluate the quality of speech signal and is a fundamental component for monitoring speech quality and improving user experience for teleconferencing systems. The source of perceptual speech quality impairments may be originated from many aspects, including assorted types of background noise, room reverberation, device coloration, audio compression and network packet loss. Conventional objective metrics such as Perceptual Evaluation of Speech Quality (PESQ) [1], Perceptual Objective Listening Quality Analysis (POLQA) [2] and Short-Time Objective Intelligibility (STOI) [3] are widely used to evaluate the speech quality and intelligibility. These intrusive metrics require reference clean speech to estimate the perceptual speech degradation of the input speech signals. Mean Opinion Score (MOS) is the most popular subjective metrics, which is guided by the Telecommunication Standardization Sector of the International Telecommunication Union (ITU-T) Recommendation P.800 [4].

In recent years, methods based on deep neural networks have been proposed for non-intrusive speech quality assessment. In [5, 6], neural networks were proposed to predict objective scores like PESQ. Although they can yield high correlation to PESQ by training the models with a large amount of simulated data, the predicted PESQ can not truly reflect subjective speech quality. In [7–9], systems were proposed to estimate the subjective MOS. However, most of the previous methods predicted MOS by directly learning a mapping between impaired speech signals and annotated scores where the performance of the systems largely depended on the amount of annotated data.

To overcome such limitation, pre-training on a large amount of data without annotation was utilized to improve the performance of non-intrusive MOS prediction. In [10], pre-trained models with unsupervised or self-supervised learning such as

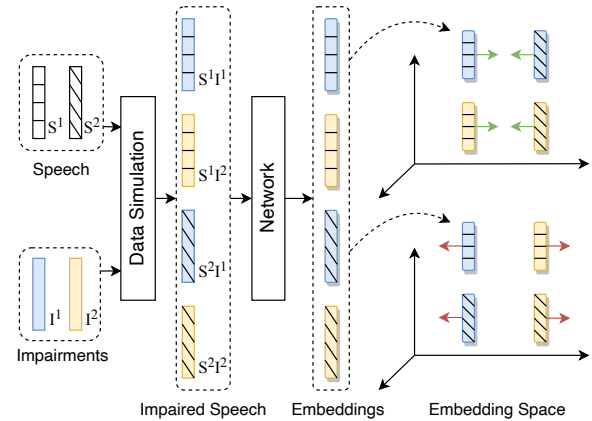


Figure 1: The proposed impairment representation learning.

wav2vec [11, 12] and Autoregressive Predictive Coding (APC) [13] were used to extract speech representations and then fine-tuned for MOS prediction. [14] also demonstrated the effectiveness of wav2vec models for generalization ability of MOS prediction. Both of these works indicated the pre-trained models with phonemic or more general speech representations can improve the MOS prediction performance for synthetic speech. However, these pre-trained models are less correlated to the impairment of speech, which is the main factor of speech quality degradation for teleconferencing systems. In [15], autoencoder is used for unsupervised feature learning on speech signals corrupted by different kinds of noises and speech enhancement methods. In [16], a two-step training method was proposed for MOS prediction of conferencing system, in which impairment classification and unsupervised deep clustering are used for the representation learning. However the representation learning by unsupervised training methods or classification for a few impairment categories can not extract discriminative representations for various kinds of impairments.

Since the degrees of impairment are highly correlated with the speech quality, we propose an impairment representation pre-training on a large amount of simulated data for non-intrusive MOS prediction. To learn a discriminative representation for various kinds of impairments, we exploit the detailed impairment information using supervised contrastive representation learning. Similar strategy has been employed in [17] to measure perceptual audio similarity for intrusive audio quality assessment. To further increase the correlation between learned representations and speech quality, we apply a multi-task learning mechanism achieved by predicting objective speech quality metrics. The pre-trained model with impairment representation is further fine-tuned for the MOS prediction task on the dataset with annotated MOS scores. Various sizes of MOS datasets are used to evaluate the relative contributions of the different pre-training stages. Self-teaching training [7] and model fusion are also applied to improve performance of the proposed system.

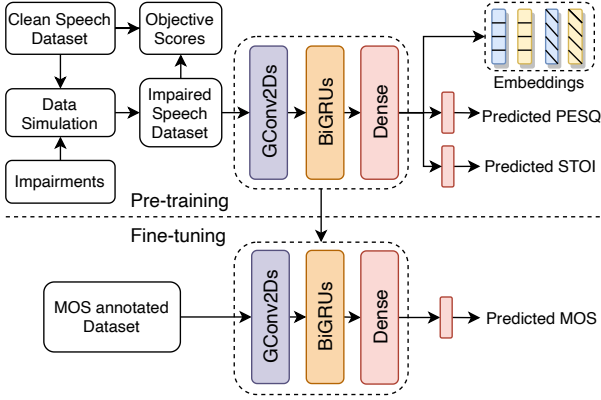


Figure 2: The proposed two-stage system with pre-training for impairment representation learning and fine-tuning for MOS prediction.

## 2. System Overview

Figure 1 illustrates the proposed impairment representation learning. Based on clean speech and impairments, speech with different impairments is generated. The impairments include various types of background noise, room reverberation, device coloration and audio compression. The impaired speech is then fed into neural network to extract corresponding embeddings. To extract impairment related representation, the speech with same impairments are pulled together and the speech with different impairments are pushed apart in embedding space.

As shown in Figure 2, the proposed two-stage speech quality assessment system consists of a pre-training stage for impairment representation learning and a fine-tuning stage for MOS prediction. During the pre-training stage, in addition to the impairment representation learning in the embedding space in Figure 1, we also propose a multi-task learning mechanism to increase the correlation between learned impairment representations and speech quality by predicting objective speech quality scores including PESQ and STOI. After learning an impairment representation on a large simulated dataset, we further fine-tune the network on a much smaller dataset with annotated MOS scores by adding a dense layer on top of the learned representations for MOS prediction. Log Mel spectrogram with 80 bins is extracted from speech signal with 16KHz sampling rate and used as input feature for pre-training and fine-tuning.

## 3. Impairment Representation Learning

### 3.1. Model Structure

The proposed neural network for impairment representation learning consists of six 2D Gated Convolution (GConv2D) blocks, three Bidirectional GRU (BiGRU) layers, one Dense layer for representation embedding extraction and one Dense layer for speech quality prediction. GConv2D has demonstrated its effectiveness in speech enhancement and audio classification tasks [18, 19]. As shown in Figure 3, in addition to the 2D convolutional (Conv2D) layer followed by a batch norm (BN) layer and dropout in the upper branch, GConv2D introduces an attention mechanism with Conv2D and sigmoid activation layer in the bottom branch. This attention mechanism makes the neural network focus on the important features and ignore the unrelated features for the representation feature learning. BiGRU [20] layers are used to extract context information from the output of GConv2D blocks. The first and the second Bi-

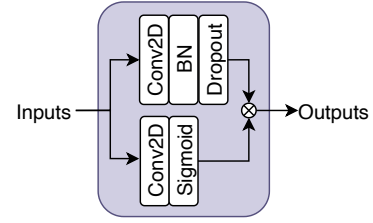


Figure 3: The structure of GConv2D block.

Table 1: Hyper-parameters of the proposed neural network.

Layer	CNN			RNN	FC
	Channel	Kernel	Stride	Units	Units
GConv2d <sup>1st</sup>	16	(3, 3)	(1, 1)		
GConv2d <sup>2nd</sup>	32	(3, 3)	(1, 2)		
GConv2d <sup>3rd</sup>	64	(3, 5)	(1, 2)		
GConv2d <sup>4th</sup>	128	(3, 7)	(1, 3)		
GConv2d <sup>5th</sup>	256	(3, 3)	(1, 1)		
GConv2d <sup>6th</sup>	512	(3, 1)	(1, 1)		
BiGRU <sup>1st</sup>				128	
BiGRU <sup>2nd</sup>				96	
BiGRU <sup>3rd</sup>				64	
Dense <sup>1st</sup>					96
Dense <sup>2nd</sup>					$P$

GRU layers return the frame sequences and the third BiGRU layer returns the features of the last frame. All of the three BiGRU layers are followed by Dropout layers.

The first Dense layer with ReLU activation is utilized to extract the impairment representation with 96 dimensions. The second Dense layer with sigmoid activation is the output layer for PESQ and STOI prediction in the pre-training stage and MOS prediction in the fine-tuning stage. For PESQ and MOS prediction, a re-scaling operation is added. The dropout rates in the network are all set to 0.2. Details of the hyper-parameters are listed in Table 1, where  $P$  equals to 2 in pre-training stage and 1 in fine-tuning stage.

### 3.2. Pre-training Objective

Contrastive loss is used for the impairment representation learning. The main idea of contrastive representation learning is to pull an anchor sample and a ‘positive’ sample together in the embedding space, while pushing the anchor sample and ‘negative’ sample apart in the embedding space. In this work, we used contrastive loss to pull audio embeddings with same impairments together and push audio embeddings with different impairments apart.

Pairs of utterances are generated for pre-training. For the  $n$ -th pair, we randomly select two clean speech ( $S_n^1, S_n^2$ ) and two types of impairments ( $I_n^1, I_n^2$ ). Then we generate four impaired speech utterances ( $S_n^1 I_n^1, S_n^1 I_n^2, S_n^2 I_n^1, S_n^2 I_n^2$ ). The ‘positive’ distance of embeddings with the same impairments but different speech in  $n$ -th pair is calculated as:

$$x_n = \frac{1}{2} (\|F_\theta(S_n^1 I_n^1) - F_\theta(S_n^2 I_n^1)\|_2 + \|F_\theta(S_n^1 I_n^2) - F_\theta(S_n^2 I_n^2)\|_2) \quad (1)$$

where  $F_\theta(\cdot)$  is the model to extract the speech embeddings,  $\theta$  is the parameters of model. The ‘negative’ distance of embeddings with different impairments but the same speech is calcu-

lated as:

$$x'_n = \frac{1}{2} (\|F_\theta(S_n^1 I_n^1) - F_\theta(S_n^1 I_n^2)\|_2 + \|F_\theta(S_n^2 I_n^1) - F_\theta(S_n^2 I_n^2)\|_2) \quad (2)$$

For  $N$  pairs of training data, the contrastive loss for impairment representation learning is calculated as:

$$\mathcal{L}_{emb} = Y * X + (1 - Y) * \text{maximum}(1 - X, 0) \quad (3)$$

where  $X = [x_1, \dots, x_N, x'_1, \dots, x'_N]$  is a  $2N$  dimensional vector with  $N$  ‘positive’ distances and  $N$  ‘negative’ distances in the embedding space,  $Y = [1, \dots, 1, 0, \dots, 0]$  is a  $2N$  dimensional vector with the first  $N$  elements equal to 1 for ‘positive’ distances and the second  $N$  elements equal to 0 for ‘negative’ distances. To further increase the correlation between the learned representations and speech quality, multi-task learning is applied during pre-training by predicting objective speech quality metrics based on the learned representations. Since we have both impaired speech and clean speech when simulating the pre-training dataset, intrusive objective speech quality metrics can be calculated. Based on the learned representations, we add one dense layer to predict wideband PESQ<sup>1</sup> and STOI<sup>2</sup>. Mean Square Error (MSE) is used to calculate the loss function  $\mathcal{L}_{pesq}$  and  $\mathcal{L}_{stoi}$  for the PESQ and STOI predictions. Combining the contrastive learning loss and objective speech quality prediction loss, the overall loss for pre-training is calculated as:

$$\mathcal{L}_{all} = \mathcal{L}_{emb} + \mathcal{L}_{pesq} + \mathcal{L}_{stoi} \quad (4)$$

## 4. MOS Prediction

After the impairment representation is trained on a large simulated dataset, we employ the impairment representation weights to initialize the model and add a dense layer for MOS prediction. The network is then fine-tuned on the MOS annotated dataset with the MSE loss:

$$\mathcal{L}_{MOS} = \text{MSE}(M, M') \quad (5)$$

where  $M$  is the annotated MOS,  $M'$  is the predicted MOS. Moreover, to eliminate the inherent noise in human rating, self-teaching training approach proposed in [7] is utilized. After getting the MOS network for the first time, the predicted MOS can be obtained for the whole training dataset. The network is trained again by combining the annotated MOS and predicted MOS by the first trained model, the new loss is calculated as:

$$\mathcal{L}_{MOS} = \text{MSE}(0.5 * M + 0.5 * r_1, M') \quad (6)$$

where  $r_1$  is the predicted MOS by the first trained model.

Model fusion can significantly improve the performance of image and video quality assessment systems [21, 22]. In this paper, to improve the generalization ability of proposed system, a simple linear fusion method is applied to fuse the results of systems with different configurations. The final output of fusion system is calculated as:

$$M'_{fusion} = \sum_j \alpha_j * M'_j \quad (7)$$

where  $J$  is the number of MOS prediction systems,  $M'_j$  is the output of  $j$ -th system,  $\alpha_j$  is the fusion weight of  $j$ -th system. In this paper we use average fusion with  $\alpha_j = \frac{1}{J}$  according to our preliminary results. Detailed information regarding the candidate models can be found in Section 5.4 and Table 3.

<sup>1</sup><https://github.com/ludlows/python-pesq>

<sup>2</sup><https://github.com/mpariente/pystoi>

## 5. Experiments

### 5.1. Dataset and Setup

**Pre-training dataset:** Clean speech datasets from LibriSpeech [23] and ST Mandarin [24], 10K stationary and non-stationary noise files from noise dataset of the ICASSP2022 DNS Challenge [25] are used for generating pre-training dataset. Four categories of impairments are simulated, including background noises, room reverberation, device coloration and audio compression. The noise impairments are applied with six kinds of SNRs [-6dB, 0dB, 6dB, 12dB, 18dB, 24dB]. Different noise files or different SNRs are considered as different impairment types, so there are 60K of noise impairment types in total. Four types of reverberation impairments are applied with simulated RIR files with RT60 in [0.3s, 0.6s, 0.9s, 1.2s]. Eight types of device coloration impairments are applied by High-Pass Filter (HPF) and Low-Pass Filter (LPF) with cut-off frequencies in [HPF: 300Hz, 1000Hz, 2000Hz and 3000Hz; LPF: 1000Hz, 2400Hz, 3600Hz and 6000Hz]. Four types of audio compression impairments are applied by Opus codec [26] with bit rates in [3kbps, 6kbps, 12kbps, 24kbps]. For each pair of data for representation learning, two impairment types are randomly selected from the four categories of [noises, reverberation, coloration, compression] with ratio [0.5, 0.2, 0.15, 0.15]. 200 thousand pairs are generated and each pair consist of four impaired utterances with 8 seconds long. So there are 800 thousand utterances with about 1800 hours in total for pre-training.

**Fine-tuning dataset:** The pre-trained model is further fine-tuned for MOS prediction on the datasets provided by the ConferencingSpeech 2022 Challenge [27]. There are three datasets we used: Tencent Corpus, PSTN Corpus and NISQA Corpus. Tencent Corpus contains 11563 speech utterances with reverberation and without reverberation. PSTN Corpus contains 58709 utterances sampled from automated phone calls. NISQA Corpus contains 14432 utterances with simulated and live conditions. Various categories of impairments are introduced by these utterances including reverberation, background noises, codec, clipping, packet-loss, jitter and so on. In our experiments, 1000 utterances of Tencent Corpus and 1000 utterances of PSTN Corpus are randomly selected for testing, and the rest of audio utterances including about 53K Tencent utterances (five times repeating of the original utterances), 58K PSTN utterances and 14K NISQA utterances are used for training and validation with a ratio of 9.5 : 0.5. All of the utterances are resampled to 16KHz audio.

**Setup:** After STFT with a frame size of 512 and hop size of 256, 80 dimensional Mel frequency bands are used to extract log power Mel spectrogram feature for pre-training and fine-tuning. The Adam optimizer is employed for model training. The initial learning rate is set to 0.001. It will decay by a factor of 0.5 when the validation loss does not decrease for 5 epochs.

### 5.2. Pre-training Evaluation

We first evaluate the pre-training systems with different size of annotated data for fine-tuning. Four MOS prediction systems with different Pre-Training (PT) methods are evaluated in Table 2. ‘None’ is the baseline system without pre-training. ‘ICC’ is the baseline pre-training method of Impairment Category Classification (ICC) as proposed in [16], a dense layer is added on the top of embedding layer to classify 4 impairment categories (noise, reverberation, device coloration and audio compression). ‘CL’ is the proposed pre-training method of Contrastive Learning (CL) as shown in Equation (3). ‘CL+OSQP’ is the proposed pre-training method with contrastive learning

Table 2: RMSE of MOS prediction for different pre-training methods and different size of annotated fine-tuning dataset.

PT-Method	FT-Size	Tencent	PSTN	All
None	1K	0.914	0.731	0.822
ICC		0.664	0.680	0.672
CL		0.614	0.644	0.629
CL+OSQP		<b>0.475</b>	<b>0.581</b>	<b>0.528</b>
None	5K	0.517	0.637	0.577
ICC		0.526	0.614	0.570
CL		0.466	0.610	0.538
CL+OSQP		<b>0.392</b>	<b>0.557</b>	<b>0.474</b>
None	120K	0.341	0.519	0.430
ICC		0.360	0.514	0.437
CL		0.334	0.514	0.424
CL+OSQP		<b>0.317</b>	<b>0.512</b>	<b>0.414</b>

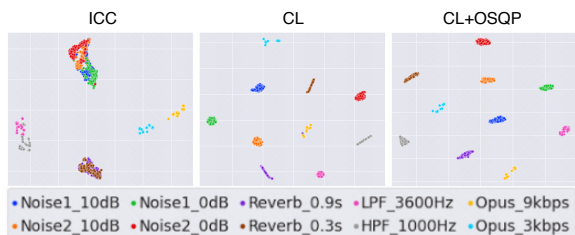


Figure 4: A visualization of learned representations for different pre-training methods.

and Objective Speech Quality Prediction (OSQP) as shown in Equation (4). Three sizes of Fine-Tuning (FT) dataset are investigated, including 1K utterances, 5K utterances and 120K utterances of the whole MOS annotated training dataset. Root Mean Square Error (RMSE) is used for evaluation.

As shown in Table 2, systems with pre-training significantly outperform the baseline system without pre-training on the small annotated dataset of 1K utterances. As the size of fine-tuning dataset increased to 5K and 120K, the proposed pre-training methods CL and CL+OSQP are still effective for improving the MOS prediction performance, while the results of the baseline pre-training method ICC is similar to the results of the system without pre-training. This indicates that pre-training with more discriminative impairment representation learning is better for MOS prediction. Besides, by adding objective speech quality metrics, CL+OSQP outperforms CL in all conditions, indicating that OSQP can increase the correlation between learned representation and subjective speech quality. It is worth noting that the proposed pre-training method CL+OSQP with 1K annotated data achieves better performance than the system directly trained on 5K annotated data, which means that the proposed method can largely reduce the amount of annotated data required for MOS prediction task.

### 5.3. Representation Analysis

To study the performance of the learned impairment representations, we extract the representations by the pre-trained models for 1600 impaired speech utterances and project them to 2D space using t-SNE [28]. The impaired speech for visualization consist of 10 types of impairments, including different settings of background noise, reverberation, device coloration and speech codec.

As shown in Figure 4, the pre-training system with ICC can not discriminate detailed impairment types such as noise impairment with different SNRs or different noise types. The

Table 3: RMSE of MOS prediction with self-teaching loss and model fusion.

PT-Method	FT-Size	DP	ST	Tencent	PSTN	All
CL+OSQP	120K	10s-z	0	0.317	0.512	0.414
CL+OSQP	120K	10s-z	1	0.309	0.509	0.409
CL+OSQP	120K	16s-r	0	0.326	0.518	0.422
CL+OSQP	120K	16s-r	1	0.324	0.512	0.418
Fusion				<b>0.293</b>	<b>0.503</b>	<b>0.398</b>

Table 4: Results on ConferencingSpeech 2022 challenge.

System	PLCC	RMSE	RMSE-Map
Baseline1	0.530	0.768	0.497
Fusion	<b>0.778</b>	<b>0.460</b>	<b>0.337</b>

proposed pre-training methods (CL and CL+OSQP) can extract discriminative representations as the representations of different impairments are well separated. Different impairments usually indicate different speech qualities. For example, speech with 10dB SNR has higher quality than the speech with 0dB SNR, speech with noise in less perceptual important frequency region (eg. 100Hz) has higher quality than speech with noise in perceptual important frequency region (eg. 2000Hz). Since the pre-trained models can extract discriminative representations for different impairments, fine-tuning based on these models can improve the performance of speech quality prediction.

### 5.4. Self-teaching and Model Fusion

We further evaluate the performance of MOS prediction with Self-Teaching (ST) as shown in Equation (6) and model fusion as shown in Equation (7) on the whole fine-tuning dataset in Table 3. Two Data Processing (DP) setups are used in the experiments. In ‘10s-z’, all speech utterances are cut to 10s, and speech utterances less than 10s are padded to 10s with zero samples. In ‘16s-r’, all speech utterances are cut to 16s, and speech utterances less than 16 are padded to 16s by repeating themselves. Note that all the systems in Table 2 are using 10s-z without ST. As shown in Table 3, self-teaching method can further improve the MOS prediction performance. Moreover, by averaging the outputs of the four systems with different data processing and training loss configurations, the final fusion system achieves the best performance.

The proposed system has ranked the 2nd place in ConferencingSpeech 2022 Challenge and significantly outperforms the Baseline1 system (a simplified version of the model in [9]) on Pearson Linear Correlation Coefficient (PLCC), RMSE and RMSE-Map (RMSE after 3rd mapping over evaluation dataset) as shown in Table 4. The proposed method has 3M parameters for single model and 12M parameters for fusion system. We evaluate real time factor (RTF) of proposed system using a privately modified version of TFLite 2.3 on Intel Core i7 (2.6 GHz) CPU (single-threaded). The RTF is 0.07 for the fusion system.

## 6. Conclusions

In this paper, we propose an impairment representation learning method based on contrastive learning and subjective speech quality prediction. The experimental results show that with large amount of simulated impaired speech, the proposed method can significantly improve the performance of MOS prediction, especially when the annotated data is limited. In the future, we would like to further explore the generalization ability of proposed method on different datasets.

## 7. References

- [1] ITU, "Itu-t: Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. rec. p.862," *ITU-T P.*, 2001.
- [2] ITU, "Itu-t: Perceptual objective listening quality prediction. rec. p.863," *ITU-T P.*, 2011.
- [3] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2010, pp. 4214–4217.
- [4] ITU, "Itu-t: Methods for subjective determination of transmission quality. rec. p.800," *ITU-T P.*, 1996.
- [5] 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. Interspeech 2018*, 09 2018, pp. 1873–1877.
- [6] 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*, 08 2021, pp. 2142–2146.
- [7] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6493–6497.
- [8] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7125–7129.
- [9] 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*, 2021, pp. 2127–2131.
- [10] W.-C. Tseng, C.-y. Huang, W.-T. Kao, Y. Lin, and H.-y. Lee, "Utilizing self-supervised representations for mos prediction," in *Proc. Interspeech 2021*, 08 2021, pp. 2781–2785.
- [11] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Proc. of Interspeech*, 2019.
- [12] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.
- [13] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," in *Proc. Interspeech 2019*, 09 2019, pp. 146–150.
- [14] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," *arXiv preprint arXiv:2110.02635*, 2021.
- [15] M. H. Soni and H. A. Patil, "Novel deep autoencoder features for non-intrusive speech quality assessment," in *2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 2315–2319.
- [16] A. Ragano, E. Benetos, and A. Hines, "More for less: Non-intrusive speech quality assessment with limited annotations," in *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, 2021, pp. 103–108.
- [17] P. Manocha, Z. Jin, R. Zhang, and A. Finkelstein, "Cdpam: Contrastive learning for perceptual audio similarity," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 196–200.
- [18] A. Li, W. Liu, X. Luo, C. Zheng, and X. Li, "Icassp 2021 deep noise suppression challenge: Decoupling magnitude and phase optimization with a two-stage deep network," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6628–6632.
- [19] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 121–125.
- [20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [21] T.-J. Liu, W. Lin, and C.-C. J. Kuo, "Image quality assessment using multi-method fusion," *IEEE Transactions on image processing*, vol. 22, no. 5, pp. 1793–1807, 2012.
- [22] J. Y. Lin, T.-J. Liu, E. C.-H. Wu, and C.-C. J. Kuo, "A fusion-based video quality assessment (fvqa) index," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–5.
- [23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [24] Surfing.ai, "St-cmds-20170001\_1, free st chinese mandarin corpus," 2017.
- [25] H. Dubey, V. Gopal, R. Cutler, A. Aazami, S. Matushevych, S. Braun, S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Icassp 2022 deep noise suppression challenge," in *Proc. of ICASSP*. IEEE, 2022.
- [26] J.-M. Valin, K. Vos, and T. Terriberry, "Definition of the opus audio codec," *IETF, September*, vol. 2, 2012.
- [27] G. Yi, W. Xiao, Y. Xiao, B. Naderi, S. Möller, G. Mittag, R. Cutler, Z. Zhang, D. S. Williamson, F. Chen *et al.*, "Conferencingspeech 2022 challenge evaluation plan."
- [28] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, 2008.