Interspeech 2024
1-5 September 2024, Kos, Greece

Enhancing No-Reference Speech Quality Assessment with Pairwise, Triplet Ranking Losses, and ASR Pretraining

Bao Thang Ta[1,2], *Minh Tu Le*[1,2], *Van Hai Do*[1,3*], *Huynh Thi Thanh Binh*[2]

[1]Viettel AI, Viettel Group, Vietnam
[2]Hanoi University of Science and Technology, Vietnam
[3]Thuyloi University, Vietnam

{thangtb3, tulm7}@viettel.com.vn, haidv@tlu.edu.vn, binhht@soict.hust.edu.vn

Abstract

Speech Quality Assessment (SQA) without reference signals has garnered attention due to its wide applications. Current SQA methods often rely on the Mean Square Error (MSE) loss to approximate human subjective ratings. However, MSE treats all deviations from the ground truth symmetrically, ignoring their direction and relative quality distinctions among speech samples. Therefore, predictions learned through MSE have limited correlations. This paper introduces a novel approach that leverages the relative quality distinctions among speech samples. By enforcing relative ranking using Pairwise and Triplet Ranking Losses, our method encourages the SQA model to learn not only the absolute quality of individual speech samples but also their quality in comparison to others, addressing the limitations of MSE-based approaches. Additionally, we suggest pretraining the SQA encoder with an ASR task to enhance generalization. Experiments on NISQA test sets confirm our approach's effectiveness.

Index Terms: Speech Quality Assessment, Triplet Ranking Loss, No-Reference Evaluation, Pairwise Loss

1. Introduction

Speech is a fundamental mode of human communication, and its quality plays a pivotal role in determining the effectiveness of various speech-based applications, spanning from telecommunications to video conferencing, voice assistants, and voice-over-IP services [1, 2]. Recognizing its crucial role, accurate assessment of speech quality receives significant attention in both research and industry.

The objective of speech quality assessment (SQA) methods is to rate the quality of an input speech signal on a scale from 1 to 5, where a higher score indicates better signal quality. Traditional methods, such as PESQ [3] and POLQA [4], rely on reference signals, comparing transmitted or recorded speech with a pristine version. However, in real-world scenarios, obtaining or using reference signals may be impractical or impossible, particularly in telecommunication applications where only the degraded signal at end devices is available.

The emergence of No-Reference Speech Quality Assessment (NR-SQA) addresses this limitation by aiming to evaluate speech quality without relying on reference signals [5, 6, 1]. NR-SQA methods seek to replicate the human perceptual process, yielding objective quality scores that exhibit strong correlations with human subjective judgments. This branch of research has garnered significant attention from the speech processing community due to its wide-ranging applications, which include enhancing user experiences in telecommunication systems and optimizing voice-controlled devices [7, 8].

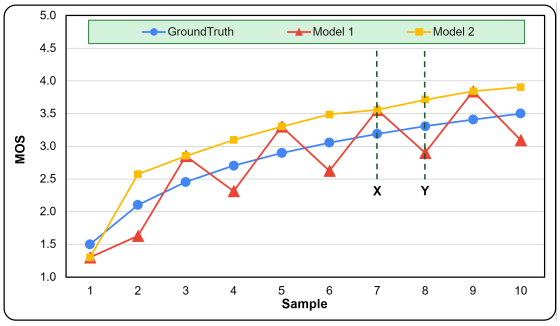

Figure 1: *Correlation of two models with the same MSE loss.*

In recent years, deep learning approaches have demonstrated impressive performance in the NR-SQA. A variety of neural network architectures have been investigated, leveraging different features extracted from speech signals. For instance, Soni et al. [7] employed an autoencoder, while Catellier et al. [6] utilized a CNN with waveform input and Avila et al. [8] with Q spectrum. Additionally, Cauchi et al. [9] proposed a modulation energy-based LSTM network, and Mittag et al. [5] presented a combination of CNN and LSTM models. Mittag et al. [1] introduced the use of self-attention with CNNs, while Shu et al. [2] proposed the Subband Adaptive Attention Temporal Convolutional Neural Network (SAA-TCN) model. Transformer architectures were also explored by Liu et al. [10] and Jayesh et al. [11]. Moreover, Ta et al. [12] introduced the Conformer architecture, a recent state-of-the-art model in speech processing, for NR-SQA tasks.

However, current NR-SQA methods frequently rely on deep learning models trained with the mean square error (MSE) loss to minimize error between predicted quality scores and human-rated quality scores [13, 14, 15]. While these models have shown promise, they often grapple with significant limitations in their capacity to consistently capture human perceptual judgments. One notable limitation is that MSE-based approaches treat samples independently during the training process, failing to account for the relative quality distinctions among them.

This issue becomes strikingly evident when observing an example in Figure 1. It can be seen that despite having the same MSE loss value, Model 2 exhibits a notably superior correlation with ground truth data compared to Model 1. Model 1 only relies on MSE, which treats all deviations from the ground truth (actual quality) symmetrically, without considering the direction of the error. Consequently, predicted quality scores can de-

*Corresponding Author

10.21437/Interspeech.2024-2527

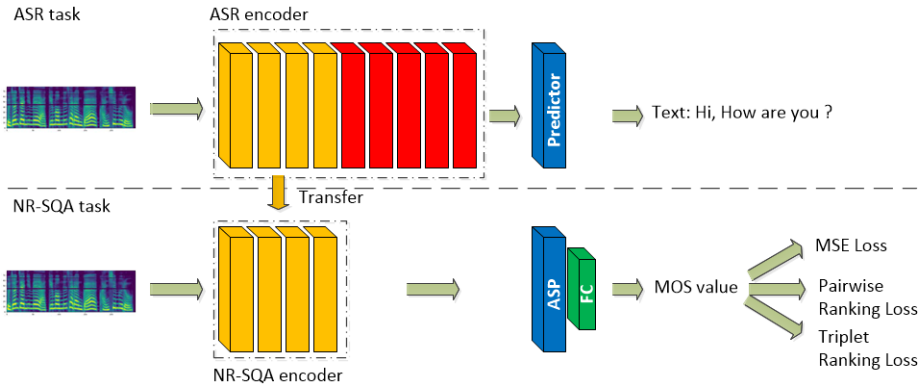

Figure 2: *Proposed No-Reference Speech Quality Assessment model.*

viate both larger and smaller than the ground truth quality. This can lead to situations where a lower-quality speech signal (X in Figure 1) can receive a higher predicted quality score than a better signal (Y). As a result, the predicted quality scores often fail to consistently demonstrate robust monotonic correlations with human evaluation scores.

To address this limitation, NR-SQA loss functions need to explicitly consider the relative ranking of quality scores. Some recent works by Manocha et al. [16, 17, 18] have made strides in handling these concerns. The authors proposed sampling N fixed pairs of inputs during each training step and utilized an auxiliary task to recognize which input is better, aiding the NR-SQA model in learning the relative quality ranking between them. In this paper, we introduce a novel approach for NR-SQA. During the training process, we sample a batch with N utterances and consider the relative ranking of any pair of random utterances in the batch. The relative ranking of $N \times (N-1)/2$ pairs of utterances is calculated to train the model. This enables our model to glean more relative ranking information from the data. Furthermore, we enhance the model's ability to capture the relative quality differences between samples by proposing an additional loss called the triplet ranking loss, introducing new definitions for positive and negative instances. This loss function aids the model in learning the relative ranking between triplets of training data samples, as opposed to pairs. This departure from traditional MSE-based methods addresses the limitations that have hindered the robustness of NR-SQA models.

Furthermore, we propose enhancing the generalization of our NR-SQA model by pretraining the encoder with an Automatic Speech Recognition (ASR) task. ASR pretraining equips our model with the ability to extract relevant features and contextual information from speech signals, which can significantly improve its ability to assess speech quality accurately in diverse real-world scenarios.

To validate the effectiveness of our proposed approach, we conducted comprehensive experimental evaluations on the NISQA dataset [1], a widely recognized benchmark for NR-SQA. Our results demonstrate substantial improvements over existing NR-SQA methods across diverse test sets. These findings underscore the potential of our approach to the field of NR-SQA and open doors to more accurate and reliable speech quality evaluation in practical applications.

2. Proposed Model

This section introduces our proposed approach, which combines Pairwise and Triplet Ranking Losses with the traditional MSE loss and incorporates an ASR pretraining stage for NR-SQA tasks. This approach addresses the limitations of existing NR-SQA methods, ensuring robust and precise speech quality assessment.

Our NR-SQA model, illustrated in Figure 2, consists of several key architectural components:

2.1. NR-SQA Encoder

Our encoder is designed with a stack of Conformer layers [19], a state-of-the-art architecture widely recognized in ASR and various speech processing tasks [20, 21, 22]. This architectural choice empowers our encoder to adeptly capture high-quality features from input data.

Moreover, a recent work [12] demonstrated the effectiveness of pretrained ASR models in extracting rich and compact speech quality information from mel-spectrograms. Surprisingly, these ASR models can compete with quality information derived from self-supervised models such as Wav2Vec2 [23] and HuBERT [24] while maintaining a substantially higher inference speed. This work also highlighted that speech quality attributes like noise, mean opinion score (MOS), discontinuity, coloration, and loudness are well concentrated in the early layers of the ASR encoder.

Building upon these findings, we propose utilizing ASR pretraining for the NR-SQA encoder in this work. The process unfolds as follows: Initially, we train an end-to-end ASR model based on the Conformer architecture, renowned as a state-of-the-art choice for ASR tasks [19, 20]. This ASR pretraining enables the NR-SQA model to leverage knowledge from ASR tasks. Subsequently, we extract and fine-tune some of the initial layers of the pretrained ASR encoder for the NR-SQA task. This adaptation empowers the model to tailor its learned representations to the specific nuances of NR-SQA, all while capitalizing on the foundational knowledge acquired during ASR pretraining.

2.2. NR-SQA Predictor

Following the encoding stage, the embedded representations undergo further processing through a dedicated predictor layer. This layer incorporates Attentive Statistical Pooling (ASP) [25] and Fully Connected (FC) layers to generate a precise predic-

tion of speech quality. By synthesizing the acquired representations with contextual information, we hope the proposed predictor can ensure highly accurate assessment.

2.3. Loss Functions

In our proposed model, we employ three distinct loss functions to enable effective learning of both absolute and relative quality distinctions among speech samples.

2.3.1. Mean Square Error (MSE) Loss

The Mean Square Error (MSE) loss quantifies the squared differences between predicted quality scores ($\bar{y}_i$) and actual quality scores (y_i). Lower MSE values indicate reduced error between predicted and actual quality scores:

$$L_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2 \tag{1}$$

It can be seen that MSE treats samples independently and does not consider relative quality distinctions among samples.

2.3.2. Pairwise Ranking Loss

The Pairwise Ranking Loss overcomes the limitations of MSE by prompting the model to distinguish the relative order among speech samples within each batch. Consider a training batch containing N samples. We evaluate the order of any two samples, denoted as x_i and x_j, with corresponding actual and predicted quality values y_i, y_j, $\bar{y}_i$, and $\bar{y}_j$. If $y_i > y_j$, it implies that x_i exhibits better quality than x_j, and vice versa. To facilitate the model's understanding of the relative order between these two samples, we formulate the Pairwise Ranking Loss inspired by a Cross Entropy Loss function as follows:

$$\text{L}_P(x_i, x_j) = -\frac{e^{y_i}}{e^{y_i}+e^{y_j}} log(\frac{e^{\bar{y}_i}}{e^{\bar{y}_i}+e^{\bar{y}_j}}) - \frac{e^{y_j}}{e^{y_i}+e^{y_j}} log(\frac{e^{\bar{y}_j}}{e^{\bar{y}_i}+e^{\bar{y}_j}}) \tag{2}$$

The total Pairwise Ranking Loss:

$$L_{pairwise} = \frac{2}{N(N-1)}\sum_{x_i,x_j} L_P(x_i, x_j) \tag{3}$$

2.3.3. Triplet Ranking Loss

To further enhance the model's ability to capture the relative quality differences between samples, we propose using an additional loss called the triplet ranking loss [26]. This loss function helps the model learn the relative ranking between triplets of training data samples, as opposed to pairs. The triplets consist of an anchor sample, denoted as x_a, a positive sample, x_p, and a negative sample, x_n.

Let y_{max}, y'_{max}, y_{min}, and y'_{min} represent the highest, second-highest, lowest, and second-lowest actual quality of samples, respectively. Additionally, $\bar{y}_{max}$, $\bar{y}'_{max}$, $\bar{y}_{min}$, and $\bar{y}'_{min}$ correspond to the predicted quality scores for these samples. For the NR-SQA task, we define anchor, negative, and positive samples as follows:

Case 1: The sample with the highest quality is chosen as an anchor, x_a. The second-highest quality sample is selected as the positive sample, x_p. Meanwhile, the negative sample is the one with the lowest quality. The triplet loss function aims to minimize the distance between samples with the highest and second-highest quality while maximizing the distance between the highest-quality sample and the lowest quality sample. The distance is defined as $d(x_a, x_n) = |\bar{y}_{max} - \bar{y}_{min}|$, and $d(x_a, x_p) = |\bar{y}_{max} - \bar{y}'_{max}|$. The objective is to minimize $\max(0, d(\bar{y}_{max}, \bar{y}'_{max}) - d(\bar{y}_{max}, \bar{y}_{min}) + margin_1)$.

Case 2: The anchor, positive, and negative samples are the samples with the lowest, second-lowest, and highest quality, respectively. The triplet loss function aims to minimize the distance between samples with the lowest and second-lowest quality while maximizing the distance between the lowest quality sample and the highest quality sample. The objective is to minimize $\max(0, d(\bar{y}_{min}, \bar{y}'_{min}) - d(\bar{y}_{min}, \bar{y}_{max}) + margin_2)$.

For a perfect situation, where the estimated quality scores match the subjective scores, we have $margin_1 + |y_{max} - y'_{max}| \leq |y_{max} - y_{min}|$, which implies that $margin_1 \leq y'_{max} - y_{min}$. Similarly, we have $margin_2 \leq y_{max} - y'_{min}$. For simplicity, we define $margin_1 = y'_{max} - y_{min}$ and $margin_2 = y_{max} - y'_{min}$. Overall, the triplet ranking loss function is calculated as follows:

$$\begin{aligned} L_{\text{triplet}} &= L_T(\bar{y}_{max}, \bar{y}'_{max}, \bar{y}_{min}) + L_T(\bar{y}_{min}, \bar{y}'_{min}, \bar{y}_{max}) \\ &= \max(0, d(\bar{y}_{max}, \bar{y}'_{max}) - d(\bar{y}_{max}, \bar{y}_{min}) + margin_1) \\ &\quad + \max(0, d(\bar{y}_{min}, \bar{y}'_{min}) - d(\bar{y}_{min}, \bar{y}_{max}) + margin_2) \end{aligned} \tag{4}$$

2.4. Training Strategy

Our model is trained using a multitask setting, simultaneously incorporating MSE, Pairwise, and Triplet Ranking Losses. This approach ensures that the model asseses both the absolute and relative qualities of speech samples, leading to robust and accurate speech quality assessment.

$$L_{total} = L_{\text{MSE}} + \lambda_1 \times L_{\text{pairwise}} + \lambda_2 \times L_{\text{triplet}} \tag{5}$$

where λ_1 and λ_2 are scale factors.

3. Experiment

3.1. Setup

The NR-SQA encoder, derived from a pretrained ASR model with seven Conformer layers, follows the configuration of a Transducer ASR model with 15 Conformer layers, a 320-sized embedding, four attention heads, and a 31-sized convolution kernel [12, 19]. The ASR model was trained on LibriSpeech for 200 epochs with 80-dimensional Mel-spectrograms as input, using AdamW optimization with a learning rate of 0.1. Our NR-SQA model, featuring approximately 19.5 million parameters, adopts the same input as the pretrained ASR. Training spans 200 epochs with AdamW (learning rate: 0.1, batch size: 16). To maintain comparable dynamic ranges for constitutive losses, we set $\lambda_1 = 10$ and $\lambda_2 = 0.1$.

3.2. Dataset

We conducted experimental assessments using NISQA [1], a well-known datasets in the NR-SQA. For training, we employed two datasets: TRAIN_SIM and TRAIN_LIVE. Validation and testing utilized four datasets: VAL_LIVE, VAL_SIM, TEST_LIVETALK, and TEST_FOR. Our evaluations were based solely on Mean Opinion Score (MOS) data.

TRAIN_SIM and VAL_SIM were generated through simulations incorporating various speech distortions like packet loss and clipping. Conversely, TRAIN_LIVE and VAL_LIVE comprised real recordings with genuine distortions such as keyboard typing and street noise.

Table 1: *Pearson Correlation Coefficient (PCC) and Root Mean Square Error (RMSE) of compared models.*

ID	Model	Loss	VAL_SIM		VAL_LIVE		TEST_LIVETALK		TEST_FOR		Average		RTFX ↑
			PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	
0	**NORESQA-MOS [18]**	MSE + Pairwise	0.838	**0.618**	0.750	0.468	0.754	**0.706**	0.762	0.626	0.776	**0.605**	4.2
1	**Conformer (from scratch)**	MSE	0.806	0.732	0.723	0.530	0.720	0.903	0.739	0.782	0.747	0.737	**427.5 (≈ 102x)**
2	**Conformer (pretrained ASR)**	MSE	0.828	0.652	0.741	0.493	0.674	0.759	0.753	0.627	0.749	0.623	**427.5 (≈ 102x)**
3	**Conformer (pretrained ASR)**	MSE + Pairwise	0.838	0.664	0.782	0.463	**0.759**	0.753	0.756	0.729	0.784	0.652	**427.5 (≈ 102x)**
4	**Conformer (pretrained ASR)**	MSE + Pairwise + Triplet	**0.840**	0.660	**0.802**	**0.439**	**0.759**	0.750	**0.764**	**0.620**	**0.791**	0.617	**427.5 (≈ 102x)**

Notes: NORESQA-MOS utilizes n = 100 non-matching clean references. RTFX is the audio duration processed / processing time ratio on a single GPU Tesla T4 (higher is better)

TEST_FOR encompassed both simulated distortions and live VoIP calls via platforms like Zoom and Skype. During these calls, original speech samples were directly played from laptops, with subsequent distortions like packet loss, warping, and low bitrate introduced. TEST_LIVETALK featured authentic phone call recordings where speakers used devices like smartphones or laptops directly. All samples were downsampled to 16 kHz to ensure compatibility with the pretrained model.

A summary of all datasets is presented in Table 2.

Table 2: *The NISQA datasets*

Datasets	Source	#Samples	Hours
TRAIN_SIM	AusTalk [27], TSP [28]	10,000	24.7
VAL_SIM	DNS Challenge [29], UK-Ireland [30]	2,500	6.0
TRAIN_LIVE	Live phone and Skype	1,020	2.6
VAL_LIVE		200	0.5
TEST_FOR	Forensic speech dataset	240	0.6
TEST_LIVETALK	Real phone and VoIP calls	232	0.6

3.3. Results

To validate the effectiveness of the proposed method, we implemented three variant models, and a recent approach NORESQA-MOS utilizing relative ranking information proposed by Manocha et al. [18] is used as a baseline, as described in Table 1. Conformer, a state-of-the-art architecture in various speech processing tasks, including NRSQA [12], was selected as the core for all models. All models were trained under identical settings for fair comparisons. Our evaluation metrics included the Pearson Correlation Coefficient (PCC) and Root Mean Square Error (RMSE), where higher PCC values indicate a stronger correlation, and lower RMSE values signify reduced error between the predicted and actual quality values. Key insights from our experiments include:

Impact of ASR Pretraining: The results clearly demonstrate the substantial impact of ASR pretraining on the models' performance. Models 2, 3, and 4, which leverage pretrained ASR models as a starting point, consistently outperform the model trained from scratch (Model 1). This finding underscores the importance of leveraging ASR models for NR-SQA tasks.

Benefit of Ranking Loss Functions: The inclusion of ranking loss functions (pairwise and triplet) in models 3 and 4 results in significantly improved PCC values compared to model 2, which only utilizes MSE loss. This indicates that incorporating ranking loss functions helps the model better capture the relative ranking between different samples, making prediction scores more correlated with human subjective evaluations.

The combination of pairwise and triplet losses also helps Model achieve the highest PCC values across all test cases while maintaining competitive RMSE when compared with all models. Notably, the baseline NORESQA-MOS achieves the best RMSE value, but it is very slow due to needing a large number of references (about 100 as suggestion by authors [18]) to reduce variance in prediction. This makes it slower than our models by about 102 times. Meanwhile, our Model 4 still achieves a very competitive RMSE, higher PCC, and is much faster. This confirms that the combination of pairwise and triplet losses is suitable for speech quality assessment tasks.

Generalization Across Real-World Tests: Notably, our enhancements in model performance, achieved through pretraining and advanced loss functions, are consistent across various test sets. The improvements in both PCC and RMSE values are particularly significant in the more challenging test scenarios, such as LIVETALK and VAL_LIVE. This suggests that our proposed enhancements lead to better generalization capabilities, especially in real-world, live speech scenarios.

4. Conclusion

This paper introduces a novel NR-SQA approach, overcoming MSE-based limitations through Pairwise and Triplet Ranking Losses. Achieving robust correlation and competitive RMSE, our model's performance is significantly boosted by integrating ASR pretraining, enhancing generalization across diverse real-world scenarios.

For future improvements, we suggest optimizing computational efficiency by considering pair order only above a quality threshold. Additionally, exploring relative rankings among more than three samples could further narrow the gap between machine and human-based assessments.

5. References

[1] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *INTERSPEECH*, pp. 2127–2131, 2021.

[2] X. Shu, Y. Chen, C. Shang, Y. Zhao, C. Zhao, Y. Zhu, C. Huang, and Y. Wang, "Non-intrusive speech quality assessment with a multi-task learning based subband adaptive attention temporal convolutional neural network," *Proc. Interspeech 2022*, pp. 3298–3302, 2022.

[3] ITU-T Recommendation 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," 2001.

[4] ITU-T Recommendation P.863, "Perceptual objective listening quality assessment," 2011.

[5] G. Mittag and S. Möller, "Quality degradation diagnosis for voice networks-estimating the perceived noisiness, coloration, and discontinuity of transmitted speech." in *INTERSPEECH*, 2019, pp. 3426–3430.

[6] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *ICASSP 2020*. IEEE, 2020, pp. 331–335.

[7] 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.

[8] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019*. IEEE, 2019, pp. 631–635.

[9] B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-intrusive speech quality prediction using modulation energies and LSTM-network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, pp. 1151–1163, 2019.

[10] Y. Liu, L.-C. Yang, A. Pawlicki, and M. Stamenovic, "CCATMos: Convolutional Context-aware Transformer Network for Non-intrusive Speech Quality Assessment," in *Proc. Interspeech 2022*, 2022, pp. 3318–3322.

[11] M. K. Jayesh, M. Sharma, P. Vonteddu, M. A. B. Shaik, and S. Ganapathy, "Transformer Networks for Non-Intrusive Speech Quality Prediction," in *Proc. Interspeech 2022*, 2022, pp. 4078–4082.

[12] B. T. Ta, M. T. Le, N. M. Le, and V. H. Do, "Probing Speech Quality Information in ASR Systems," in *Proc. INTERSPEECH 2023*, 2023, pp. 541–545.

[13] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022*. IEEE, 2022, pp. 8442–8446.

[14] W.-C. Tseng, C. yu Huang, W.-T. Kao, Y. Y. Lin, and H. yi Lee, "Utilizing Self-Supervised Representations for MOS Prediction," in *Proc. Interspeech 2021*, 2021, pp. 2781–2785.

[15] H. Becerra, A. Ragano, and A. Hines, "Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction," in *Proc. Interspeech 2022*, 2022, pp. 4088–4092.

[16] P. Manocha, B. Xu, and A. Kumar, "Noresqa: A framework for speech quality assessment using non-matching references," *Advances in Neural Information Processing Systems*, vol. 34, pp. 22 363–22 378, 2021.

[17] P. Manocha, Z. Jin, and A. Finkelstein, "Sqapp: No-reference speech quality assessment via pairwise preference," in *ICASSP 2022*, 2022, pp. 891–895.

[18] P. Manocha and A. Kumar, "Speech Quality Assessment through MOS using Non-Matching References," in *Proc. Interspeech 2022*, 2022, pp. 654–658.

[19] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Proc. Interspeech 2020*, 2020, pp. 5036–5040.

[20] S. Kim, A. Gholami, A. Shaw, N. Lee, K. Mangalam, J. Malik, M. W. Mahoney, and K. Keutzer, "Squeezeformer: An efficient transformer for automatic speech recognition," *arXiv preprint arXiv:2206.00888*, 2022.

[21] Y. Zhang, Z. Lv, H. Wu, S. Zhang, P. Hu, Z. Wu, H.-y. Lee, and H. Meng, "Mfa-conformer: Multi-scale feature aggregation conformer for automatic speaker verification," in *INTERSPEECH*. ISCA, 2022.

[22] B. T. Ta, X. V. Dang, Q. T. Duong *et al.*, "Improving vietnamese accent recognition using asr transfer learning," in *2022 25th Conference of the Oriental COCOSDA (O-COCOSDA)*. IEEE, 2022.

[23] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[24] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 29, p. 3451–3460, oct 2021.

[25] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Proc. Interspeech 2018*, 2018, pp. 2252–2256.

[26] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson, "Ranked list loss for deep metric learning," in *Proceedings of the IEEE CVPR 2019*, 2019, pp. 5207–5216.

[27] D. Burnham, D. Estival, S. Fazio, J. Viethen, F. Cox, R. Dale, S. Cassidy, J. Epps, R. Togneri, M. Wagner, Y. Kinoshita, R. Göcke, J. Arciuli, M. Onslow, T. Lewis, A. Butcher, and J. Hajek, "Building an audio-visual corpus of Australian English: large corpus collection with an economical portable and replicable black box," in *Proc. Interspeech 2011*, 2011, pp. 841–844.

[28] P. Kabal, "Tsp speech database," *McGill University, Database Version*, vol. 1, no. 0, pp. 09–02, 2002.

[29] C. K. Reddy, V. Gopal, R. Cutler, E. Beyrami, R. Cheng, H. Dubey, S. Matusevych, R. Aichner, A. Aazami, S. Braun, P. Rana, S. Srinivasan, and J. Gehrke, "The INTERSPEECH 2020 Deep Noise Suppression Challenge: Datasets, Subjective Testing Framework, and Challenge Results," in *Proc. Interspeech 2020*, 2020, pp. 2492–2496.

[30] I. Demirsahin, O. Kjartansson, A. Gutkin, and C. Rivera, "Open-source multi-speaker corpora of the english accents in the british isles," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6532–6541.