Quality-Aware Contrastive Self-Supervised Learning for Speech Quality Assessment

Minh Tu Le[1*], Bao Thang Ta[1,2*], Phi Le Nguyen[2], Van Hai Do[1,3]
[1]Viettel AI, Viettel Group [2]Hanoi University of Science and Technology [3]Thuyloi University
tulm7@viettel.com.vn; thangtb3@viettel.com.vn; lenp@soict.hust.edu.vn; haidv@tlu.edu.vn
*These authors contributed equally to this work.

***Abstract*—Speech Quality Assessment (SQA) plays a critical role in evaluating the performance of communication systems, yet the scarcity of labeled data significantly limits the development of robust SQA models. In this paper, we propose Contrastive learning for Quality Awareness (CQA), a novel self-supervised learning (SSL) framework specifically designed for speech quality assessment. Unlike conventional SSL models that prioritize content modeling, CQA focuses on learning quality-discriminative features by constructing contrastive pairs based on shared distortion types. Notably, two segments with different content can still be considered a positive pair if they share the same type of distortion. This design allows the model to focus solely on distortion characteristics, rather than the content of the original speech samples, effectively disentangling quality-related features from semantic information. Additionally, we introduce PESQ score prediction as an auxiliary task to align the pretrained SSL model more directly with the downstream SQA objective. Extensive experiments on the NISQA benchmark demonstrate that CQA significantly outperforms existing state-of-the-art methods across five quality dimensions: Mean Opinion Score (MOS), Noisiness, Coloration, Discontinuity, and Loudness. Specifically, our approach achieves a 2.1–6.2% improvement in Pearson's Correlation Coefficient (PCC) and a 7.5–14.4% reduction in Root Mean Squared Error (RMSE) compared to the best existing methods on the TEST_FOR and TEST_P501 datasets, setting a new state-of-the-art in speech quality assessment.**

***Index Terms*—Speech Quality Assessment, Contrastive Learning, Self-Supervised Learning.**

I. Introduction

The rapid advancement of information and communication technology has made voice calls and online meetings integral to daily communication. However, speech transmitted over networks is susceptible to various forms of degradation, including packet loss, background noise, reverberation, and filtering, which can significantly impact user experience and lead to communication failures. Consequently, Speech Quality Assessment (SQA) has become a critical component for evaluating and optimizing communication systems' performance, attracting substantial interest from telecom companies and internet service providers.

Recent developments in deep learning have demonstrated remarkable success in addressing the SQA problem. Various neural network architectures have been explored, utilizing diverse feature representations extracted from speech signals to enhance quality prediction accuracy [1]–[3]. Furthermore, several studies have tailored their approaches specifically to SQA by designing customized objective functions that better capture the perceptual nature of speech quality [4]–[8]. Despite these significant advancements in architectural design and learning objectives, the development of robust SQA models remains fundamentally constrained by the scarcity of labeled training data. Obtaining high-quality labeled data for SQA, particularly Mean Opinion Scores (MOS), is inherently labor-intensive and expensive, requiring extensive human evaluation sessions with multiple trained assessors—a process that severely limits the scalability of data collection efforts.

To overcome the limitations imposed by labeled data scarcity, researchers have increasingly turned to self-supervised learning (SSL) approaches. Large-scale SSL models such as XLS-R [9], wav2vec2 [10], and HuBERT [11], which have demonstrated exceptional performance across various speech processing tasks, have been successfully adapted for speech quality assessment [12]–[14]. Recent investigations have explored the utilization of XLS-R and BYOL-S [15] as feature extractors for SQA applications [16], while comprehensive comparative studies have evaluated the effectiveness of various SSL models including CPC [17], TERA [18], APC [19], and wav2vec2 in the context of quality assessment [20]. Additionally, [21] demonstrated promising results by employing pretrained Automatic Speech Recognition (ASR) Conformer models for SQA tasks.

While these SSL-based approaches have shown considerable success, they face a fundamental limitation: these models were originally designed and optimized for content-focused tasks such as ASR, where the primary objective is to understand and transcribe speech content. When adapted to SQA through fine-tuning, these models may exhibit suboptimal performance due to the inherent mismatch between their pretraining objectives and the discriminative nature of speech quality assessment [22]. Unlike content understanding tasks, SQA requires the model to focus on quality-related acoustic characteristics and distinguish between subtle variations in speech rather than semantic content.

To address this fundamental challenge, we propose **Contrastive learning for Quality Awareness (CQA)**, a novel self-supervised learning framework explicitly designed and optimized for speech quality assessment. Our approach leverages contrastive learning principles to learn quality-discriminative representations by constructing positive and negative sample pairs based on shared distortion characteristics rather than content similarity. Specifically, we introduce a quality-aware

pair construction strategy where speech segments subjected to same distortions form positive pairs, while segments with different distortions serve as negative samples. Positive and negative pairs can include speech segments extracted from different original speech samples. This helps our method focus only on the distortion, rather than on the content of the original speech samples. This design enables the model to focus on quality-related features and develop representations that are inherently suitable for discriminating between speech distortions.

Furthermore, we enhance the alignment between pretraining and downstream objectives by incorporating PESQ score prediction [23] as an auxiliary task. PESQ, being a widely adopted intrusive speech quality metric, provides direct supervision for quality-related learning and strengthens the connection between the pretext task and the ultimate SQA goal.

Our experimental evaluation demonstrates that CQA significantly outperforms existing state-of-the-art methods across multiple quality dimensions on the NISQA benchmark [24]. The proposed approach achieves substantial improvements of 2.1–6.2% in Pearson Correlation Coefficient (PCC) and 7.5–14.4% reduction in Root Mean Squared Error (RMSE) compared to current best-performing methods, establishing new state-of-the-art performance for speech quality assessment.

The main contributions of this work are summarized as follows:

- We propose CQA, the first contrastive learning framework specifically tailored for speech quality assessment, which addresses the mismatch between content-focused SSL pretraining and quality-discriminative downstream tasks.
- We introduce a novel quality-aware pair construction strategy that creates positive and negative samples based on distortion characteristics, enabling effective learning of quality-discriminative representations.
- We incorporate PESQ score prediction as an auxiliary task to strengthen the alignment between pretraining and downstream SQA objectives.
- We demonstrate significant performance improvements over existing state-of-the-art methods across five quality dimensions on standard benchmarks, achieving new state-of-the-art results.

The remainder of this paper is organized as follows: Section II details the proposed CQA framework, including the construction of quality-aware positive and negative sample pairs for contrastive learning. Section III presents comprehensive experimental results and ablation studies, highlighting the effectiveness of each component in our approach. Finally, Section IV summarizes our findings and discusses potential directions for future research.

II. Proposed Method

In this section, we present Contrastive learning for Quality Awareness (CQA), a novel framework tailored for the SQA problem. Leading SSL models for speech representation, such as wav2vec2 [10], WavLM [25], and HuBERT [11], focus on predicting the hidden structure of speech, such as masked frames, to capture content-related information. In contrast, SSL models in image processing, like SimCLR [26] and MoCo [27], adopt a different strategy. These models compare augmented versions of the same image (positive pairs) with unrelated images (negative samples), organizing them in an embedding space based on similarity.

We argue that SSL approaches focusing on predicting hidden structures, though effective for content-based tasks like ASR, may not be optimal for SQA, which is inherently more discriminative in nature. SQA requires distinguishing between subtle speech distortions. In this sense, contrastive learning—which excels at capturing discriminative features (as seen in tasks like image classification and object detection)—better aligns with the needs of SQA. Previous work [22] further supports the idea that the choice of pretext task significantly affects downstream task performance.

Motivated by these insights, we propose to adapt the principles of image-based contrastive learning to speech quality assessment. Instead of focusing on content understanding, our approach emphasizes the representation of quality, allowing the model to capture and differentiate between various types of speech distortions.

CQA forms positive and negative pairs of speech samples based on shared distortions. For each training batch, CQA takes clean speech samples as input and randomly extracts two segments from each sample. Subsequently, CQA randomly generates a set of D distortions and applies this same set of distortions to all extracted segments. For example, for a speech sample x^n in the training batch, we randomly extract two segments, x_1^n and x_2^n, which naturally share similar frequency and volume properties. We then apply the same set of D random distinct distortions to both segments, producing two sets of degraded speech: $X_1^n = [x_{1,1}^n, x_{1,2}^n, \ldots, x_{1,D}^n]$ and $X_2^n = [x_{2,1}^n, x_{2,2}^n, \ldots, x_{2,D}^n]$, as illustrated in Figure 1.

When the same distortion d is applied, the resulting degraded samples are treated as a positive pair, including those extracted from different original speech samples. This helps our method focus only on the distortion, rather than on the content of the original speech samples. In contrast, pairs with different distortions are considered negative pairs.

These speech samples are then fed into an encoder, for which we adopt the Conformer architecture, known for its superior performance in speech processing tasks. The features extracted by the encoder, which may vary in length, are processed by an Attentive Pooling layer [21] to produce fixed-size embedding vectors. Since certain time frames may disproportionately influence overall speech quality, attentive pooling helps the model focus on quality-relevant frames, capturing key characteristics of speech distortions.

Following the approach in [26], a non-linear projection head is employed to map the embedding vectors into a lower-dimensional space, facilitating effective contrastive learning. Cosine similarity is used to compute similarity between the

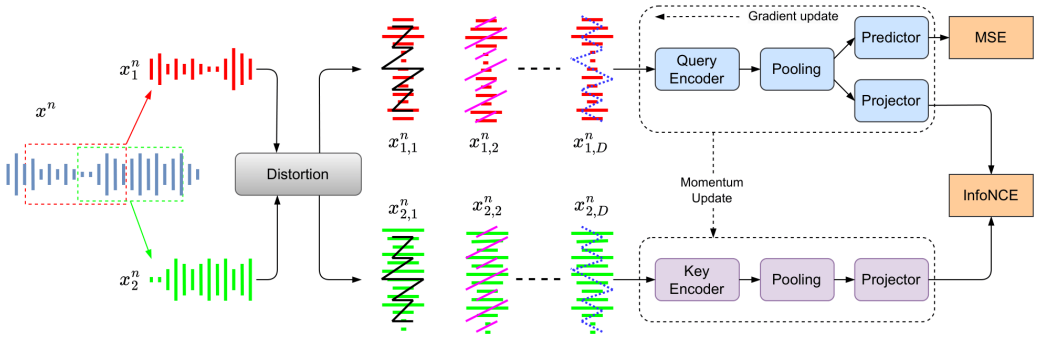

Fig. 1. **Overview of the proposed method**: Two segments extracted from the same clean speech are subjected to different distortions, generating two sets of degraded speech. These sets are fed into the query and key branches, and their projected vectors are compared using the InfoNCE loss to measure similarity. Additionally, the embeddings from the query branch are used to predict PESQ scores.

projected vectors, which are then used to calculate the contrastive loss.

Figure 1 illustrates the proposed CQA built on the MoCo framework [27], which implements two branches: a query branch updated by gradient descent, and a key branch updated by momentum. For each sample x^n, we define $x^n_{q,d}$ and $x^n_{k,d'}$ as segments selected randomly from x^n, where they are applied distortions d and d', respectively. The notation q indicates that the segment is fed to the query branch, and k indicates that it is fed to the key branch. $z^n_{q,d}$ denotes the latent vector after the projector module of segment $x^n_{q,d}$. In the case of CQA built on SimCLR [26], the model employs only one branch for both query and key roles.

Additionally, CQA maintains a queue of K projected vectors, providing additional negative samples for each batch. The contrastive loss is then computed as:

$$L_{\text{cons}}(n,i,d) = -\log \frac{\exp\left(\frac{z^n_{q,d}\cdot z^i_{k,d}}{\tau}\right)}{\sum_{j=1}^{N}\sum_{d'=1}^{D}\exp\left(\frac{z^n_{q,d}\cdot z^j_{k,d'}}{\tau}\right) + \sum_{j=1}^{K}\sum_{\substack{d'=1\\ d'\neq d}}^{D}\exp\left(\frac{z^n_{q,d}\cdot z^j_{k,d'}}{\tau}\right)} \tag{1}$$

Here, N is the batch size, K is the queue size (following [27]), and τ is a temperature parameter controlling the sharpness of the softmax distribution.

The complete contrastive loss across all samples and distortions in a batch is:

$$L_{\text{cons}} = \frac{1}{N^2 \times D}\sum_{n=1}^{N}\sum_{i=1}^{N}\sum_{d=1}^{D} L_{\text{cons}}(n,i,d) \tag{2}$$

To further strengthen the alignment between the pretext task and the downstream speech quality assessment objective, we introduce an auxiliary task: predicting the PESQ score [23]. For each degraded speech sample $x^i_{k,d}$, PESQ estimates its perceptual quality by comparing it with the original clean speech x^i_k. We use Mean Squared Error (MSE) as the loss function for this regression task. The overall training loss is the combination of the contrastive loss L_{cons} and the PESQ prediction loss L_{PESQ}, weighted by a hyperparameter α:

$$L_{\text{train}} = L_{\text{cons}} + \alpha L_{\text{PESQ}} \tag{3}$$

This combination ensures that the model not only learns to differentiate between distortions but also aligns its representations with a perceptual speech quality metric, enhancing downstream performance.

In the fine-tuning phase, the encoder and pooling layers from the pretrained model serve as the backbone, providing quality-aware embeddings for each input. These embeddings are passed to a Feedforward Network (FFN) designed to predict the MOS as well as four additional quality dimensions: Noisiness, Coloration, Discontinuity, and Loudness. The model is trained in a multi-dimensional regression setting, with MSE as the loss function across all five dimensions, yielding a comprehensive assessment of speech quality.

III. Experiments

A. Dataset

Pretraining: Since we require high-quality clean speech to generate degraded samples, the pretraining dataset is obtained from the clean subset of LibriSpeech [28], which contains audiobook recordings. We utilize the train-clean-100 and train-clean-360 subsets for training, totaling 460 hours of audio, while the test-clean and dev-clean subsets, comprising 11 hours, are used for validation.

We apply seven types of degradations: Background Noise, Room Impulse Response (RIR), Filtering, Packet Loss, Clipping, Gaussian Noise, and Codec variations. The background noise and RIR data are sourced from the DNS-Challenge dataset [29], with Signal-to-Noise Ratios (SNRs) ranging from -10 to 40 dB, and RIR durations of 0.3, 0.5, 0.7, and 0.9 seconds. The filtering category includes Low Pass, High Pass, Band Pass, and Band Reject filters. Gaussian noise is applied with SNRs from -5 to 25 dB. Packet loss is simulated using lossy noise files from the PLC Challenge dataset [30]. To create a distortion type, we randomly apply one or combine multiple types.

Fine-tuning: The NISQA corpus [24], recognized as a benchmark dataset for speech quality assessment [31], is used for fine-tuning and evaluation. It includes two training subsets, TRAIN_SIM and TRAIN_LIVE, comprising 12,500 samples with a total duration of 30.7 hours. The validation set consists of VAL_SIM and VAL_LIVE, which together

contain 1,220 samples spanning 3.1 hours. For testing, we use two datasets: TEST_FOR and TEST_P501, comprising 480 samples over 1.1 hours. VAL_SIM and VAL_LIVE represent seen conditions, while TEST_FOR and TEST_P501 correspond to unseen conditions. Each sample is labeled with five quality dimensions: Mean Opinion Score (MOS), Noisiness (NOI), Coloration (COL), Discontinuity (DIS), and Loudness (LOUD).

B. Setup

In our experiments, we adopt the Conformer architecture with the medium configuration proposed by [32]: 16 Conformer layers with a decoder dimension of 320, 4 attention heads, and a convolution kernel size of 32. The input features consist of 80-channel filterbanks. The attentive pooling layer produces embeddings of size 2048 and uses a dropout rate 0.1.

During pretraining, we set the loss temperature to 0.07 and use a queue of 8192 key values, following the default hyperparameters of MoCo [27], and assign a prediction task weight of $\alpha = 0.3$. The model is trained for 50 epochs with a batch size of 64. The projection layer output dimension is set to 256, and the size of D is 4. Fine-tuning is then performed for 200 epochs, with a learning rate of 0.001. All experiments are conducted on an NVIDIA A100 GPU with 40GB of memory.

We evaluate model performance using two metrics: Pearson's Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE), where higher PCC and lower RMSE values indicate superior performance.

C. Ablation Study

This section examines the contributions of our proposed CQA method and the auxiliary PESQ prediction task. All experiments use the Conformer architecture, with training conducted on TRAIN_SIM and TRAIN_LIVE, and evaluation performed on VAL_LIVE and VAL_SIM datasets.

We test five versions: *Conformer (fs)*: trained from scratch. ***Conformer (pt PESQ)***: pretrained on PESQ prediction. ***CQA (MoCo)***: implementing CQA following MoCo. ***CQA (SimCLR) + PESQ*** and ***CQA (MoCo) + PESQ***: full versions differing only in using SimCLR or MoCo for CQA training

Table I reports the average PCC and average RMSE on all quality dimensions for each model. Some key findings are drawn as follows:

Impact of Contrastive Learning: *CQA (MoCo)* outperforms both the scratch method *Conformer (fs)* and the PESQ prediction pretrained model *Conformer (PESQ)*.

Impact of Auxiliary Task (PESQ Prediction): Although *Conformer (PESQ)* performs similarly to the scratch method *Conformer (fs)*, the PESQ prediction task significantly benefits contrastive learning. Due to its relation to the downstream task, *CQA (MoCo) + PESQ* improves upon *CQA (MoCo)* in both datasets.

Comparison of SimCLR and MoCo: *CQA (SimCLR) + PESQ* not only underperforms compared to *CQA (MoCo) + PESQ* but also fails to show improvement over *CQA (MoCo)*, highlighting MoCo's advantage in this task.

D. Comparison with SoTA methods

We compare our proposed method against several state-of-the-art models, evaluating both pretrained and trained-from-scratch approaches: **Trained-from-scratch models**: We include the baseline *NISQA* model from [24] (Exp 1) and a Conformer model trained from scratch, referred to as *Conformer (fs)* (Exp 2). **Additional labeled data**: We also evaluate the *NISQA+* model (Exp 3), a version of NISQA trained on additional private labeled data. Results for this model are publicly provided in [24]. **Pretrained models**: We consider Impairment Representation Learning [33] (*IRL*, Exp 4), a prior SSL-based SQA approach. We also include *wav2vec2 (pt)* (Exp 5), the pretrained ASR model *Conformer (pt ASR)* (Exp 6), with results provided by [21], and a recent approach *MultiGauss* (Exp 7) representing current state-of-the-art methods. **Proposed method**: Our proposed model, **CQA + PESQ** (Exp 8), uses a contrastive loss following MoCo combined with an MSE loss for PESQ prediction.

All models are trained on two labeled datasets, with the exception of *NISQA+*, which leverages additional private data. Additionally, the *wav2vec2 (pt)* and *Conformer (pt ASR)* models are pretrained on 960 hours of LibriSpeech, while our proposed method uses a smaller subset of 460 hours.

The average PCC and RMSE values for the five quality dimensions across each dataset are presented in Table II. Overall, models leveraging pretraining or additional labeled data (Exps 3-8) consistently outperform those trained from scratch (Exps 1-2).

While the baseline *IRL* shows improvement over scratch models, it performs worse than *NISQA+* and other pretrained models. Our proposed method achieves superior performance compared to *NISQA+* despite utilizing significantly less labeled data, highlighting the efficiency of our contrastive learning approach. The proposed *CQA + PESQ* framework demonstrates exceptional performance across all evaluation datasets, achieving the highest PCC values on three out of four datasets: 0.699 on VAL_SIM, 0.908 on TEST_FOR, and 0.903 on TEST_P501. On the remaining dataset, VAL_LIVE, our method achieves a competitive PCC of 0.865, ranking among the top-performing approaches. In terms of RMSE, our method delivers the lowest error rates on two datasets while showing competitive results on the remaining datasets.

Notably, although models such as *wav2vec2 (pt)* (Exp 5) and *Conformer (pt ASR)* (Exp 6) perform strongly due to extensive pretraining, **CQA + PESQ** surpasses them, particularly on the test datasets. For instance, in TEST_P501 and TEST_FOR, RMSE improves by approximately 9.18–13.8% relative to *wav2vec2 (pt)*, while PCC increases by about 2.7–6.2%. Compared to *Conformer (pt ASR)*, our method increases 2.1 to 2.3% of PCC values while reducing 7.5 to 14.4% of RMSE values.

MultiGauss (Exp 7) achieves the highest PCC (0.866) on VAL_SIM, but exhibits severe overfitting with substantial performance degradation on real-world datasets: PCC drops to 0.694 on VAL_LIVE (19.8% decrease), 0.821 on TEST_FOR

TABLE I
COMPARISON WITH OTHER PRETRAINING METHODS

Model	VAL_LIVE		VAL_SIM	
	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓
Conformer (fs)	0.652	0.500	0.843	0.524
Conformer (pt PESQ)	0.651	0.509	0.831	0.551
CQA (simCLR) + PESQ	0.671	0.506	0.860	0.500
CQA (MoCo)	0.687	0.469	0.854	0.496
CQA (MoCo) + PESQ	**0.699**	**0.465**	**0.865**	**0.486**

TABLE II
AVERAGE PCC AND RMSE ON ALL QUALITY DIMENSIONS ON VALIDATION AND TEST DATASET

Exp	Model	Data		VAL_SIM		VAL_LIVE		TEST_FOR		TEST_P501	
		unlabeled	labeled	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓
1	NISQA [24]	X	small	0.683	0.477	0.851	0.522	0.879	0.429	0.876	0.509
2	Conformer (fs)	X	small	0.652	0.500	0.843	0.524	0.853	0.443	0.856	0.567
3	NISQA+ [24]	X	large	0.679	0.498	0.852	0.520	0.899	0.372	0.878	0.517
4	IRL [33]	small	small	0.695	**0.460**	0.846	0.509	0.862	0.434	0.869	0.521
5	wav2vec2 (pt) [12]	large	small	0.685	0.467	0.867	**0.482**	0.884	0.392	0.850	0.540
6	Conformer (pt ASR) [21]	large	small	0.695	0.478	**0.869**	0.496	0.888	0.385	0.884	0.543
7	MultiGauss [34]	small	small	**0.866**	0.516	0.694	0.523	0.821	0.457	0.854	0.523
8	**CQA + PESQ**	small	small	0.699	0.465	0.865	0.486	**0.908**	**0.356**	**0.903**	**0.465**

(9.6% decrease), and 0.854 on TEST_P501 (5.4% decrease). We hypothesize this occurs because MultiGauss optimizes for a probabilistic objective that may overfit to the simulated data distribution. In contrast, our CQA framework demonstrates consistent improvements (higher PCC and lower RMSE) on VAL_LIVE, TEST_FOR, and TEST_P501 (unseen conditions), highlighting the effectiveness of our quality-aware contrastive learning approach for real-world SQA applications.

Detailed PCC and RMSE results for all quality dimensions on TEST_FOR and TEST_P501 are provided in Tables III and IV. These tables do not include *Conformer (pt ASR)* (Exp 6) and *MultiGauss* (Exp 7) results, as they are not publicly available.

While *wav2vec2 (pt)* achieves a PCC of 0.912 for MOS, its RMSE remains higher than that of **CQA + PESQ**. Notably, **CQA + PESQ** is pretrained on only 460 hours of data, compared to 960 hours used by other models such as *wav2vec2*. Moreover, across most quality dimensions, **CQA + PESQ** consistently delivers superior results—particularly in *Loudness*, where it improves PCC by 9.2–15.7% and reduces RMSE by 25.6-41.5% relative to *wav2vec2 (pt)*.

IV. CONCLUSION

This paper introduced Contrastive learning for Quality Awareness (CQA), a self-supervised framework tailored for Speech Quality Assessment (SQA). CQA addresses the limitations of content-centric pretraining by focusing on quality-discriminative features through a novel distortion-aware contrastive learning strategy. By forming positive pairs from segments sharing the same type of distortion—even across different utterances—the model learns to isolate distortion characteristics from speech content. Additionally, incorporating PESQ prediction as an auxiliary task strengthens alignment with downstream SQA objectives. Experimental results on the NISQA benchmark demonstrate that CQA consistently outperforms existing state-of-the-art methods across five quality dimensions.

For future work, we plan to evaluate CQA on more diverse and real-world datasets and explore dynamic weighting strategies between the contrastive and auxiliary loss components, replacing the current fixed-weight approach to allow the model to better adapt during different training stages. These efforts will further enhance the generalization and applicability of CQA in real-world communication systems.

ACKNOWLEDGEMENT

Bao Thang Ta was funded by the PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), VinUniversity, code VINIF.2025.TS65.

REFERENCES

[1] Danilo de Oliveira, Julius Richter, Jean-Marie Lemercier, Simon Welker, and Timo Gerkmann, "Non-intrusive Speech Quality Assessment with Diffusion Models Trained on Clean Speech," in *Interspeech 2025*, 2025, pp. 2330–2334.

[2] Wafaa Wardah, Robert P. Spang, Vincent Barriac, Jan Reimes, Anna Llagostera, Jens Berger, and Sebastian Möller, "SQ-AST: A Transformer-Based Model for Speech Quality Prediction," in *Interspeech 2025*, 2025, pp. 2335–2339.

[3] Imran E Kibria and Donald S. Williamson, "AttentiveMOS: A Lightweight Attention-Only Model forSpeech Quality Prediction," in *Interspeech 2025*, 2025, pp. 2340–2344.

[4] Minh Tu Le, Bao Thang Ta, Nhat Minh Le, Phi Le Nguyen, and Van Hai Do, "A gaussian distribution labeling method for speech quality assessment," in *International Conference on Computational Data and Social Networks*. Springer, 2023, pp. 27–38.

[5] Xinyu Liang, Fredrik Cumlin, Christian Schüldt, and Saikat Chatterjee, "DeePMOS: Deep Posterior Mean-Opinion-Score of Speech," in *INTERSPEECH 2023*, 2023, pp. 526–530.

[6] Bao Thang Ta, Minh Tu Le, Van Hai Do, and Huynh Thi Thanh Binh, "Enhancing no-reference speech quality assessment with pairwise, triplet ranking losses, and asr pretraining," in *INTERSPEECH 2024*, 2024, pp. 2700–2704.

[7] Yu-Fei Shi, Yang Ai, and Zhen-Hua Ling, "Universal Preference-Score-based Pairwise Speech Quality Assessment," in *Interspeech 2025*, 2025, pp. 2345–2349.

[8] Enjamamul Hoq, Nikhil Gupta, Danielle Omondi, and Ifeoma Nwogu, "FUSE-MOS: Fusion of Speech Embeddings for MOS Prediction with Uncertainty Quantification," in *Interspeech 2025*, 2025, pp. 2350–2354.

[9] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *INTERSPEECH 2022*, 2022, pp. 2278–2282.

TABLE III
RESULTS OF COMPARED METHODS ON EACH QUALITY DIMENSION IN TEST_FOR

Exp	Model	Data		MOS		NOI		DIS		COL		LOUD	
		unlabeled	labeled	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓
1	NISQA [24]	X	small	0.850	0.482	0.858	0.435	0.886	0.509	0.866	0.387	0.934	0.331
2	Conformer (fs)	X	small	0.816	0.520	0.840	0.446	0.858	0.521	0.824	0.420	0.926	0.309
3	NISQA+ [24]	X	large	0.891	0.418	0.892	0.339	0.899	0.458	0.887	0.352	0.928	**0.293**
4	IRL [33]	small	small	0.818	0.503	0.873	0.411	0.840	0.537	0.847	0.399	0.931	0.322
5	wav2vec2 (pt) [12]	large	small	**0.911**	**0.377**	0.910	0.310	**0.925**	**0.375**	0.813	0.497	0.860	0.403
8	**CQA + PESQ**	small	small	0.893	0.406	**0.915**	**0.302**	0.898	0.443	**0.896**	**0.330**	**0.939**	0.300

TABLE IV
RESULTS OF COMPARED METHODS ON EACH QUALITY DIMENSION IN TEST_P501

Exp	Model	Data		MOS		NOI		DIS		COL		LOUD	
		unlabeled	labeled	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓	PCC ↑	RMSE ↓
1	NISQA [24]	X	small	0.899	**0.542**	0.864	0.543	0.851	0.597	0.830	0.503	0.937	0.359
2	Conformer (fs)	X	small	0.851	0.678	0.869	0.584	0.806	0.642	0.824	0.512	0.931	0.418
3	NISQA+ [24]	X	large	0.904	0.560	0.900	0.440	0.834	0.624	0.840	0.548	0.914	0.412
4	IRL [33]	small	small	0.870	0.655	0.883	0.548	0.823	0.600	0.833	**0.473**	0.935	0.331
5	wav2vec2 (pt) [12]	large	small	**0.912**	0.573	0.897	0.472	0.793	0.630	0.834	0.492	0.816	0.533
8	**CQA + PESQ**	small	small	0.911	0.588	**0.925**	**0.431**	**0.865**	**0.511**	**0.868**	0.482	**0.944**	**0.312**

[10] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[11] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[12] Helard Becerra, Alessandro Ragano, and Andrew Hines, "Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction," in *INTERSPEECH 2022*, 2022, pp. 4088–4092.

[13] Bastiaan Tamm, Rik Vandenberghe, and Hugo Van Hamme, "Analysis of xls-r for speech quality assessment," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023, pp. 1–5.

[14] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8442–8446.

[15] Gasser Elbanna, Neil Scheidwasser-Clow, Mikolaj Kegler, Pierre Beckmann, Karl El Hajal, and Milos Cernak, "Byol-s: Learning self-supervised speech representations by bootstrapping," in *HEAR: Holistic Evaluation of Audio Representations*. PMLR, 2022, pp. 25–47.

[16] Karl El Hajal, Zihan Wu, Neil Scheidwasser-Clow, Gasser Elbanna, and Milos Cernak, "Efficient speech quality assessment using self-supervised framewise embeddings," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[17] Luyu Wang, Kazuya Kawakami, and Aäron van den Oord, "Contrastive predictive coding of audio with an adversary.," in *INTERSPEECH*, 2020, pp. 826–830.

[18] Andy T Liu, Shang-Wen Li, and Hung-yi Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, 2021.

[19] Gene-Ping Yang, Sung-Lin Yeh, Yu-An Chung, James Glass, and Hao Tang, "Autoregressive predictive coding: A comprehensive study," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1380–1390, 2022.

[20] Wei-Cheng Tseng, Chien yu Huang, Wei-Tsung Kao, Yist Y. Lin, and Hung yi Lee, "Utilizing Self-Supervised Representations for MOS Prediction," in *INTERSPEECH 2021*, 2021, pp. 2781–2785.

[21] Bao Thang Ta, Minh Tu Le, Nhat Minh Le, and Van Hai Do, "Probing Speech Quality Information in ASR Systems," in *INTERSPEECH 2023*, 2023, pp. 541–545.

[22] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola, "What makes for good views for contrastive learning?," *Advances in neural information processing systems*, vol. 33, pp. 6827–6839, 2020.

[23] 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.

[24] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," *INTERSPEECH 2021*, pp. 2127–2131, 2021.

[25] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[27] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[28] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[29] Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 deep noise suppression challenge," in *INTERSPEECH*, 2021.

[30] Lorenz Diener, Sten Sootla, Solomiya Branets, Ando Saabas, Robert Aichner, and Ross Cutler, "INTERSPEECH 2022 Audio Deep Packet Loss Concealment Challenge," in *INTERSPEECH 2022*, 2022, pp. 580–584.

[31] Gaoxiong Yi, Wei Xiao, Yiming Xiao, Babak Naderi, Sebastian Möller, Wafaa Wardah, Gabriel Mittag, Ross Culter, Zhuohuang Zhang, Donald S. Williamson, Fei Chen, Fuzheng Yang, and Shidong Shang, "ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications," in *INTERSPEECH 2022*, 2022, pp. 3308–3312.

[32] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *INTERSPEECH 2020*, 2020, pp. 5036–5040.

[33] Lianwu Chen, Xinlei Ren, Xu Zhang, Xiguang Zheng, Chen Zhang, Liang Guo, and Bing Yu, "Impairment representation learning for speech quality assessment," in *INTERSPEECH 2022*, 2022, pp. 3323–3327.

[34] Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan K.A. Reddy, Christian Schüldt, and Saikat Chatterjee, "Multivariate Probabilistic Assessment of Speech Quality," in *Interspeech 2025*, 2025, pp. 5413–5417.