EM-VSQA: Enhancing VLSP Speech Quality Assessment with External Data and Multi-Task Learning

Tri Dung Do
University of Engineering and Technology, Vietnam National University, Vietnam
22025501@vnu.edu.vn

Abstract

Speech Quality Assessment (SQA) aims to approximate human perceptual judgments of speech quality without relying on costly and time-consuming subjective MOS tests. In this paper, we introduce **EM-VSQA**, a **wav2vec2+BiLSTM** based model enhanced with two key strategies: (i) a multi-loss learning scheme that combines Mean Squared Error with ListNet ranking loss to jointly optimize absolute prediction accuracy and relative ranking consistency, and (ii) the integration of a small subset of VocalSound data to improve robustness against non-speech vocal events. These enhancements enable the model to achieve reliable performance across diverse audio conditions. Our system ranked Top-2 in the VLSP 2025 Speech Quality Assessment shared task, highlighting the effectiveness of combining multi-loss optimization with targeted external data augmentation for robust SQA. In addition, we conducted a further exploration with an ensemble variant, **EEM-VSQA**, which incorporates a VocalSound classifier to explicitly detect vocal events and assign high-quality scores. This experimental strategy achieved the best performance on the private test set, providing deeper insights into handling challenging edge cases, although it was not part of the official submission.

1 Introduction

Speech Quality Assessment (SQA) is a fundamental problem in speech and audio processing, with applications in voice communication, speech synthesis, automatic speech recognition, and virtual assistants (Mittag et al., 2021; Shu et al., 2022). An effective SQA system ensures better user experience, enables automatic quality control, and facilitates real-time monitoring of communication systems.

Traditionally, the gold standard for evaluating speech quality relies on human subjective ratings, where listeners assess attributes such as naturalness, clarity, and intelligibility. While reliable, this approach is costly, time-consuming, and labor-intensive, making it impractical for large-scale or real-time scenarios (Manocha et al., 2021).

To address these limitations, machine learning-based methods have been developed to predict speech quality automatically. Existing approaches can be broadly categorized into *reference-based* and *non-reference* methods. Reference-based techniques (Beerends et al., 2013; Rix et al., 2001) compare degraded audio against a clean reference, which is often unavailable in real-world applications. Non-reference methods (Mittag and Möller, 2019; Catellier and Voran, 2020), on the other hand, predict speech quality directly from the degraded signal, making them more practical for deployment.

Recent progress in non-reference SQA has been driven by deep learning architectures such as Transformers (Vaswani et al., 2017; Baevski et al., 2020), CNNs (Ye et al., 2022), and Conformers (Gulati et al., 2020; Ta et al., 2024a). These models are commonly trained with a regression objective using mean squared error (MSE) loss. In some cases, auxiliary branches, such as pairwise or triplet-ranking loss, are introduced to enhance feature representation (Ta et al., 2024b).

In this paper, we propose **EM-VSQA**, a non-reference speech quality assessment framework built on a **wav2vec2+BiLSTM** backbone, enhanced with two key innovations. First, we incorporate a small but targeted subset of external VocalSound data, which improves robustness against non-speech vocal events frequently observed in real-world recordings. Second, we employ a multi-loss strategy that integrates mean squared error (MSE) with the ListNet loss, a listwise ranking objective, thereby balancing absolute score regression with relative ranking consistency. This design enables EM-VSQA to deliver more robust and gener-

alizable predictions across diverse conditions. The main contributions of this work are as follows:

- We propose **EM-VSQA**, a novel non-reference framework that integrates targeted external data to improve robustness against challenging non-speech vocal events.
- We design a multi-loss objective that jointly optimizes MSE for precise score regression and ListNet for preserving ranking relationships, thereby improving overall prediction robustness.
- We provide extensive experiments demonstrating that **EM-VSQA** consistently outperforms strong baselines, validating the effectiveness of our design choices. In addition, we further explore an ensemble variant, **EEM-VSQA**, which integrates a VocalSound classifier during inference to explicitly handle edge cases. This experimental strategy achieved the best results on the private test set, offering insights for future research directions.

2 Proposed Method

2.1 Overview

The overall framework of our approach is illustrated in Figure 1.

We employ an architecture that integrates wav2vec2 (Baevski et al., 2020) as the encoder for speech representation, BiLSTM (Graves and Schmidhuber, 2005) for feature pooling and an multi-layer perceptron (MLP) module for output prediction. The model is trained on an external non-speech vocal dataset, namely the VocalSound dataset (Gong et al., 2022), using a multi-task learning strategy with both MSE (regression loss) and ListNet Loss (Cao et al., 2007) (ranking loss). All components were integrated during the VLSP competition in response to specific issues we observed throughout the challenge.

2.2 Model Architecture

Given an input speech waveform $x \in \mathbb{R}^T$, where T denotes the number of samples, the objective of our model is to predict a Mean Opinion Score (MOS) $\hat{y}$ on a five-point scale (1 to 5).

First, we employ a wav2vec2 encoder f_θ to transform the raw waveform into frame-level acoustic representations:

$$h = f_\theta(x), \quad h \in \mathbb{R}^{L \times d}, \tag{1}$$

where L is the number of frames and d is the feature dimension.

To capture long-term temporal dependencies, we apply a bidirectional LSTM (BiLSTM) network g_ϕ on the sequence h, and obtain the final representation z by averaging over all hidden states:

$$H = g_\phi(h), \quad H \in \mathbb{R}^{T \times d}, \tag{2}$$

$$z = \frac{1}{T} \sum_{t=1}^{T} H_t, \quad z \in \mathbb{R}^d, \tag{3}$$

where T is the sequence length and d is the hidden dimension of the BiLSTM.

Finally, the aggregated feature z is passed through a MLP m_ψ to produce a scalar prediction:

$$\hat{y} = m_\psi(z), \quad \hat{y} \in [1, 5], \tag{4}$$

which corresponds to the predicted MOS of the given utterance.

During training, the model parameters are optimized using the Mean Squared Error (MSE) loss between predicted scores $\hat{y}_i$ and ground-truth MOS labels y_i:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 . \tag{5}$$

This formulation enables the model to learn a mapping from speech waveforms to continuous MOS on the 1–5 quality scale.

2.3 Training with External Data

Our model was trained on two datasets: the official public training set provided by the VLSP organizers and the external VocalSound dataset (Gong et al., 2022). Initially, the model was trained solely on the official VLSP training dataset. However, when evaluated on the public test set, the model exhibited poor performance.

Upon closer examination, we found that the public test set includes a notable number of utterances containing non-speech sounds (e.g., background laughter) as well as clean, high-quality speech segments. In such cases, the model often misinterprets these signal characteristics and assigns them low MOS scores, as if they were degraded or noisy utterances. This mismatch between training and testing distributions highlighted a limitation of using only the official public training data.

To address this issue, we incorporated the VocalSound dataset into training, which consists of

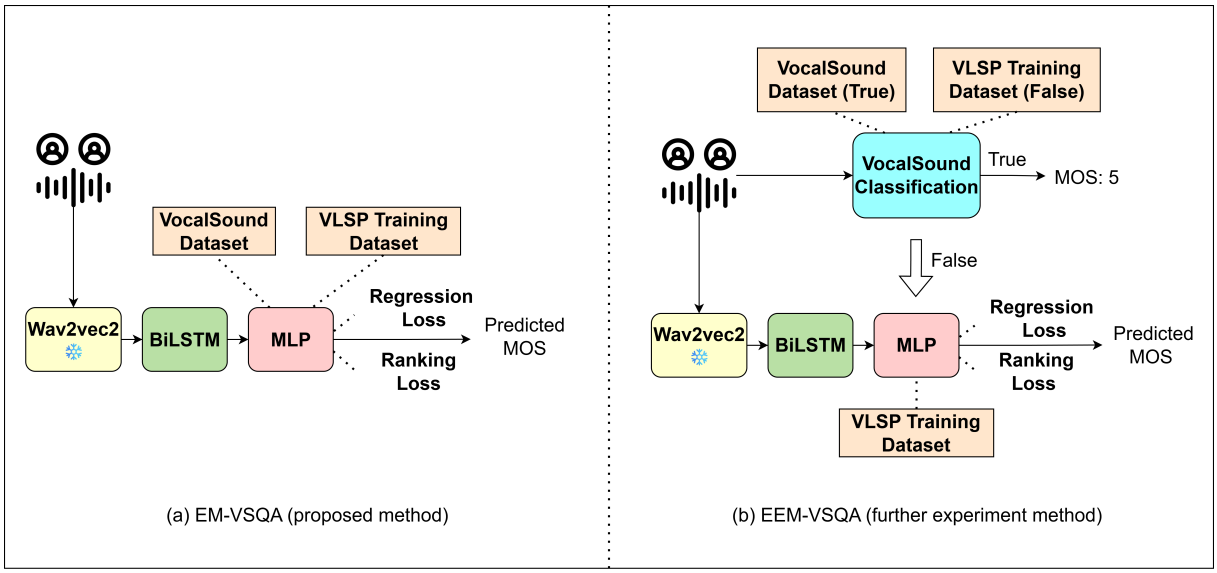

Figure 1: Overall framework of our approaches. (a) EM-VSQA: input speech is encoded by wav2vec2, followed by a BiLSTM and an MLP, trained jointly on the VocalSound and VLSP datasets with regression and ranking losses. (b) EEM-VSQA: extends EM-VSQA by adding a VocalSound classification module trained on the VocalSound dataset. At inference, if an input is detected as non-speech vocalization, a fixed MOS of 5 is assigned; otherwise, the MOS prediction is obtained from the EM-VSQA pipeline trained on the VLSP dataset.

diverse non-speech vocalizations such as laughter, coughs, and breaths. Since the recordings in VocalSound are clean and of high perceptual quality, we heuristically assigned them the maximum MOS of 5. We then performed sampling from this dataset and combined it with the original VLSP training set to construct a more diverse training distribution.

This strategy allowed the model to better distinguish between genuinely noisy utterances and clean but atypical cases such as laughter or other non-speech vocalizations. As a result, the inclusion of VocalSound significantly improved the model's prediction accuracy on the VLSP public test set.

2.4 Multi-Task Learning using Regression and Ranking Loss

While our baseline model is trained with the Mean Squared Error (MSE) loss for regression, we argue that Speech Quality Assessment (SQA) should not only be treated as an absolute scoring task, but also as a relative ranking problem: instead of only asking "how good" a single utterance is, the model should also learn "which utterances are better or worse" in comparison.

To capture this property, we introduce the ListNet loss (Cao et al., 2007), a widely used listwise ranking objective. In our setting, although the approach differs, ListNet is employed to encourage consistent rankings among speech samples within a batch, with the expectation of achieving improvements in correlation similar to those reported in (Ta et al., 2024b). Formally, given a batch of N utterances with ground-truth MOS y_i and model predictions $\hat{y}_i$, we define probability distributions using the softmax function:

$$P(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^{N} \exp(y_j)}, \quad P(\hat{y}_i) = \frac{\exp(\hat{y}_i)}{\sum_{j=1}^{N} \exp(\hat{y}_j)}. \tag{6}$$

The ListNet loss is then computed as the cross-entropy between the two distributions:

$$\mathcal{L}_{\text{ListNet}} = -\sum_{i=1}^{N} P(y_i) \log P(\hat{y}_i). \tag{7}$$

The final training objective combines both regression and ranking objectives as:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{MSE}} + (1 - \lambda)\mathcal{L}_{\text{ListNet}}, \tag{8}$$

where $\lambda \in [0, 1]$ balances the contribution of each loss.

Empirically, adding the ListNet loss leads to a significant improvement on the VLSP public test set, demonstrating the benefit of treating SQA as both regression and ranking.

3 Experiments

3.1 Baseline Models

We consider two baselines for comparison with our proposed approach (**EM-VSQA**):

wav2vec2 + BiLSTM. This baseline uses the same model architecture and training configuration as **EM-VSQA**. However, unlike **EM-VSQA**, it is trained *only* on the VLSP training set (without incorporating VocalSound) and optimizes a single MSE loss, instead of the combined multi-loss objective. This baseline highlights the contribution of both external data augmentation and the multi-loss design in our proposed approach.

NISQA (Mittag et al., 2021). The second baseline leverages the *Neural Intrusive/Non-intrusive Speech Quality Assessment (NISQA)* framework, a deep learning model for non-intrusive speech quality prediction. NISQA predicts both overall speech quality and four perceptual dimensions: *Noisiness*, *Coloration*, *Discontinuity*, and *Loudness*, which provide insights into the causes of quality degradation. In our experiments, we directly use the pre-trained NISQA model to infer speech quality on the VLSP test sets. Among the available outputs, we focus on the Distortion score, since preliminary evaluations on the training data showed that this dimension correlates best with human-annotated MOS.

3.2 Dataset

The overall dataset composition is summarized in Table 1. The training dataset provided by the VLSP organizers consists of 5,493 speech samples with varying levels of audio quality, each annotated with a Mean Opinion Score (MOS) ranging from 1 (lowest quality) to 5 (highest quality). These MOS reflect human perceptual judgments on speech quality. It is important to note that the VLSP data was originally recorded at an 8 kHz sampling rate. For consistency with our model architecture and feature extraction pipeline, all recordings are resampled to 16 kHz prior to training.

In addition, we incorporate a subset of the VocalSound dataset (Gong et al., 2022), a free crowdsourced collection of 21,024 recordings of non-speech vocal events such as laughter, sighs, coughs, throat clearing, sneezes, and sniffs, collected from 3,365 speakers with metadata (age, gender, native language, and health condition). We randomly sample 250 clips (approximately 5% of the VLSP training size) and assign them the maximum MOS of 5, so that the model learns to distinguish such events without being biased toward always predicting high scores.

For evaluation, we follow the VLSP setting, with 1,717 samples in the public test set and 2,221 samples in the private test set. During training, the data is split into 80% training and 20% validation with label-balanced partitions. To improve training stability, we adopt length-based batch sorting, where samples are grouped by duration before batching, thereby reducing padding overhead and leading to more efficient optimization.

3.3 Experimental Setup

All audio samples are resampled to 16 kHz prior to training. The proposed model architecture consists of a wav2vec2 encoder to extract frame-level acoustic representations, followed by a BiLSTM layer that captures temporal dependencies in both forward and backward directions. On top of the BiLSTM, we use a multi-layer perceptron (MLP) to predict the final MOS. The models are trained for 200 epochs to ensure sufficient convergence and stable performance. If a prediction falls outside the valid MOS range (1–5), it is clipped to remain within [1, 5]. The detailed hyperparameters are summarized in Table 2.

The choice of combining mean squared error (MSE) with ListNet loss is motivated by the dual objectives of speech quality assessment: (1) predicting an accurate MOS in the continuous scale (regression), and (2) preserving the relative ranking of speech samples in terms of quality (ranking). We set the trade-off parameter $\lambda = 0.5$ to balance these two objectives, ensuring that the model does not overfit to one aspect while neglecting the other. Empirically, this setting yielded stable training and better generalization on both public and private test sets.

3.4 Evaluation Metrics

Following the VLSP challenge setting, the final score is defined as a weighted combination of the Pearson correlation coefficient (PCC) and the Mean Squared Error (MSE):

$$\text{SCORE} = 0.7 \times \text{PCC} - 0.3 \times \text{MSE}. \quad (9)$$

The PCC between predicted scores $\hat{y}$ and ground-

Table 1: Dataset composition.

Dataset	Details
VLSP training set	5,493 speech samples (8 kHz, MOS 1–5)
VocalSound (subset)	250 clips from 21,024 recordings, MOS=5
VLSP public test set	1,717 samples
VLSP private test set	2,221 samples

Table 2: Model architecture and training hyperparameters.

Component	Configuration
Input audio	Resampled to 16kHz
Feature encoder	wav2vec2 base model
BiLSTM	Hidden = 256, Dropout = 0.3
MLP layers	$256 \rightarrow 128 \rightarrow 64 \rightarrow 1$
Activation	ReLU (between layers)
Dropout	0.3 (applied in MLP)
Optimizer	Adam
Learning rate	0.001
Training epochs	200
Loss function	$(1-\lambda) \cdot \mathcal{L}_{\text{MSE}} + \lambda \cdot \mathcal{L}_{\text{ListNet}}$
λ (weight)	0.5 (balanced training)

truth MOS y is computed as:

$$\text{PCC} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}\sqrt{\sum_{i=1}^{N}(\hat{y}_i - \bar{\hat{y}})^2}}, \tag{10}$$

where $\bar{y}$ and $\bar{\hat{y}}$ denote the means of y and $\hat{y}$, respectively.

The MSE is computed as:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2. \tag{11}$$

3.5 Results

Table 3 summarizes the experimental results on both the public and private test sets, comparing our proposed **EM-VSQA** with the baselines.

The baseline **wav2vec2+BiLSTM**, which does not use external data nor multi-loss training, performs poorly, yielding even negative correlation (PCC = -0.0612 on the public test and nearly zero on the private test). This indicates that the model trained solely on the official training data fails to generalize.

On the other hand, **NISQA (distortion score)** provides more competitive results, achieving moderate correlation (PCC = 0.3369 on the public test and 0.3842 on the private test). However, its performance is still limited in terms of both PCC and MSE, since it was not specifically adapted to the VLSP dataset.

In contrast, our proposed **EM-VSQA** significantly outperforms both baselines across all metrics. On the public test, it achieves a PCC of 0.7979 and the lowest MSE of 0.2113, while on the private test, it maintains strong performance with PCC = 0.7624 and MSE = 0.3382. These results demonstrate that the integration of external VocalSound data and multi-loss optimization leads to more robust and accurate speech quality assessment.

Furthermore, we acknowledge that in our official competition submission, the input audio was inadvertently not resampled from 8 kHz to 16 kHz prior to inference, even though the model had been trained on 16 kHz data. This oversight introduced a mismatch between training and evaluation conditions, leading to a marked degradation in performance, with the submitted scores of PCC = 0.6282, MSE = 0.4776, and SCORE = 0.2965. For the sake of accuracy and fairness, in this paper we report the corrected results obtained with proper resampling, which more faithfully represent the actual performance of our system.

3.6 Ablation Study

To better understand the contribution of each component in **EM-VSQA**, we conduct an ablation study by gradually enabling external training data and the multi-task learning strategy. The results are summarized in Table 4.

When trained only on the official dataset with a single-task objective, the system performs poorly, with PCC values close to zero (-0.0612 on the public set and 0.0125 on the private set) and even negative overall scores. This indicates that the official training data alone is insufficient to capture the variability of real-world speech quality, and the model tends to overfit.

Incorporating external training corpora yields a substantial improvement: PCC rises to 0.7455 (public) and 0.6845 (private), while MSE decreases by nearly 80% compared to the baseline. These

Table 3: Performance comparison on VLSP public and private test sets.

Method	Public Test			Private Test		
	PCC	MSE	SCORE	PCC	MSE	SCORE
wav2vec2+BiLSTM	-0.0612	1.2999	-0.4328	0.0125	1.2313	-0.3606
NISQA (distortion score)	0.3369	0.5275	0.0775	0.3842	0.7468	0.0449
EM-VSQA (Ours)	**0.7979**	**0.2113**	**0.4951**	**0.7624**	**0.3382**	**0.4322**

Table 4: Ablation study on the impact of external data and multi-task learning.

External	Multi-task	Public Test			Private Test		
		PCC	MSE	SCORE	PCC	MSE	SCORE
		-0.0612	1.2999	-0.4328	0.0125	1.2313	-0.3606
✓		0.7455	0.2679	0.4414	0.6845	0.3962	0.3603
✓	✓	**0.7979**	**0.2113**	**0.4951**	**0.7624**	**0.3382**	**0.4322**

results highlight the critical role of data scale and diversity, as the external data introduces more acoustic conditions and perceptual variations, enabling the model to better align with human ratings.

Adding the multi-task loss on top of external data further enhances robustness. The final system achieves PCC scores of 0.7979 (public) and 0.7624 (private), with consistent gains in both MSE and SCORE. This confirms that learning auxiliary objectives encourages the model to capture richer representations of perceptual quality, which are complementary to the benefits of external data.

Overall, both external data and multi-task optimization are indispensable: external data provides the necessary coverage of acoustic conditions, while multi-task learning improves generalization. Their combination leads to the best performance across all evaluation metrics.

3.7 Further Exploration

Table 5: Comparison between the proposed EM-VSQA and its ensemble variant (EEM-VSQA) on the VLSP private test set.

Method	Private Test		
	PCC	MSE	SCORE
EM-VSQA	0.7624	0.3382	0.4322
EEM-VSQA	**0.8119**	**0.2536**	**0.4922**

During the experimental analysis, we observed that **EM-VSQA** occasionally assigned unexpectedly high MOS values to distorted speech, mistakenly interpreting them as VocalSound. In addition, some samples with ground-truth MOS $= 5$ were predicted with scores below 4, indicating that certain VocalSound utterances still failed to be correctly classified.

To analyze this phenomenon, we define three types of prediction samples:

- **Over-predicted sample**: A sample for which the predicted MOS $\hat{y}$ is higher than 3.5, while the ground truth y is at least 1 point lower than the prediction, i.e.,

$$\hat{y} > 3.5 \quad \text{and} \quad y \leq \hat{y} - 1. \tag{12}$$

- **Under-predicted sample**: A sample for which the ground truth $y = 5$ but the predicted MOS is strictly lower than 4, i.e.,

$$y = 5 \quad \text{and} \quad \hat{y} < 4. \tag{13}$$

- **Normal sample**: All other cases.

The illustration of these three types is shown in Fig. 2 (a). As can be seen, many low-quality samples were clearly over-predicted, while a number of clean utterances with $y = 5$ were under-predicted. These inconsistencies indicate that the model occasionally misjudged both degraded and high-quality speech, thereby reducing robustness and limiting generalization.

We hypothesize that these inconsistencies arise because incorporating additional VocalSound data during training altered the distribution of the original dataset. Consequently, some low-quality samples were incorrectly assigned higher MOS values,

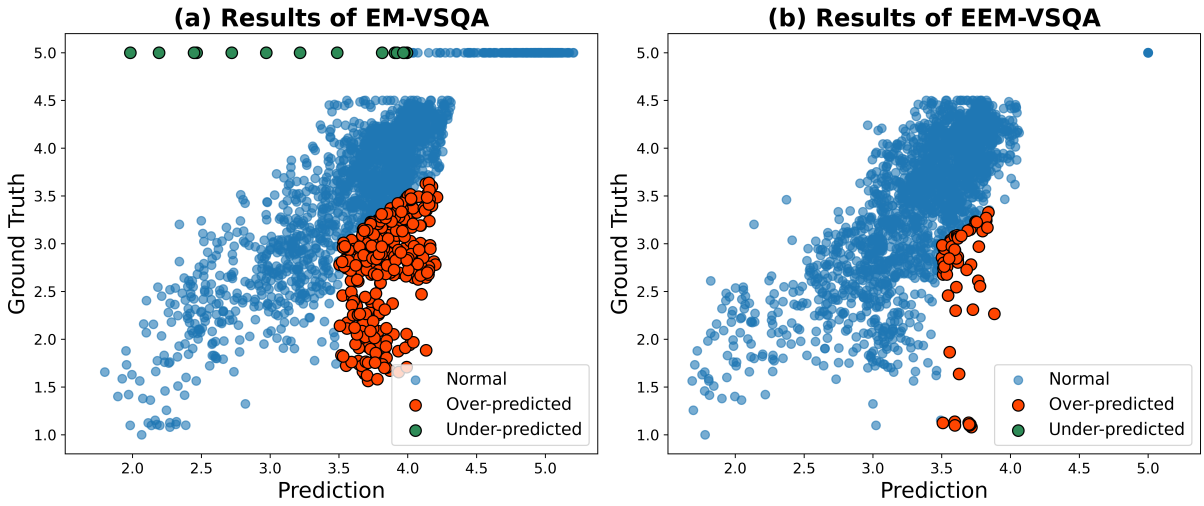

Figure 2: Scatter plots of predicted versus ground truth MOS on the VLSP private test set. Color denotes prediction types: Over-predicted ($\hat{y} > 3.5$ and $y \leq \hat{y} - 1$), Under-predicted ($y = 5$ and $\hat{y} < 4$), and Normal (otherwise). (a) **EM-VSQA**: Several distorted utterances are *Over-predicted*, while some clean utterances with MOS = 5 are *Under-predicted*. (b) **EEM-VSQA**: The ensemble strategy reduces both Over-predicted and Under-predicted cases, yielding predictions more consistent with the ground truth.

while a portion of VocalSound utterances still received lower-than-expected scores. To address this issue, it is essential to adopt a method that preserves the distribution of the official training set while explicitly detecting VocalSound utterances, ensuring that they are consistently assigned a MOS of 5 without disturbing the original distribution.

To mitigate this issue, we explored an alternative ensemble strategy, denoted as **EEM-VSQA**. Specifically, we trained a **wav2vec2+BiLSTM** model exclusively on the VLSP training data in order to preserve the distribution of the official dataset. In parallel, we introduced a VocalSound classification module, based on wav2vec2, to explicitly detect whether an input corresponds to VocalSound. At inference time, the prediction rule is defined as:

$$\hat{y} = \begin{cases} 5, & \text{if classified as VocalSound,} \\ pred, & \text{otherwise,} \end{cases} \quad (14)$$

where $pred$ denotes the MOS prediction from the **wav2vec2+BiLSTM** model.

The results on the VLSP private test set are summarized in Table 5. **EEM-VSQA** consistently outperformed **EM-VSQA**, achieving higher PCC (0.8119 vs. 0.7624), lower MSE (0.2536 vs. 0.3382), and an overall score improvement of 0.4922 compared to 0.4322. More importantly, as shown in Fig. 2 (b), the ensemble approach effectively eliminated under-predicted utterances for MOS $= 5$ samples and corrected most of the over-predicted distorted cases, thereby producing predictions more closely aligned with the ground truth.

It is worth noting that this ensemble strategy was not part of our official submission to the VLSP competition. Instead, it was conducted as a post-hoc analysis to provide deeper insights into the task and to suggest potential future directions.

4 Conclusion

In this work, we presented **EM-VSQA**, a framework for speech quality assessment that leverages multi-task learning and external data augmentation. Our approach integrates a dual-objective loss, combining MSE with ListNet ranking, which encourages the model to capture both absolute quality scores and relative ranking relationships. Furthermore, by incorporating a small subset of high-quality vocal events from the VocalSound dataset, the model learns to better handle diverse vocal expressions that commonly appear in real-world scenarios.

Experimental results on the VLSP 2025 benchmark demonstrate that **EM-VSQA** consistently outperforms strong baselines, including wav2vec2 + BiLSTM and the NISQA model, in both public and private test sets. Additional exploration further revealed that explicitly modeling VocalSound can bring additional performance gains, suggesting that targeted handling of special cases is a promising direction.

For future work, we aim to develop strategies that enhance the generalization capability of SQA systems across a broader range of real-world challenges. Instead of focusing solely on specific edge cases such as VocalSound, our goal is to design more universal frameworks that can robustly adapt to diverse noise types, recording conditions, and speaking styles, thereby improving reliability in practical deployment.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

John Beerends, Chris Schmidmer, Jens Berger, Matthias Obermann, Raphael Ullmann, Joachim Pomy, and Michael Keyhl. 2013. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i-temporal alignment. *AES: Journal of the Audio Engineering Society*, 61:366–384.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 129–136, New York, NY, USA. Association for Computing Machinery.

Andrew A. Catellier and Stephen D. Voran. 2020. Wawenets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 331–335.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 151–155. IEEE.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech 2020*, pages 5036–5040.

Pranay Manocha, Buye Xu, and Anurag Kumar. 2021. Noresqa: A framework for speech quality assessment using non-matching references. *Advances in neural information processing systems*, 34:22363–22378.

Gabriel Mittag and Sebastian Möller. 2019. Quality degradation diagnosis for voice networks — estimating the perceived noisiness, coloration, and discontinuity of transmitted speech. In *Interspeech 2019*, pages 3426–3430.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. Nisqa: A deep cnn-self-attention model for multidimensional speech quality prediction with crowdsourced datasets. In *Interspeech 2021*, pages 2127–2131.

Antony W. Rix, John G. Beerends, Mike Hollier, and Andries P. Hekstra. 2001. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, 2:749–752 vol.2.

Xiaofeng Shu, Yanjie Chen, Chuxiang Shang, Yan Zhao, Chengshuai Zhao, Yehang Zhu, Chuanzeng Huang, and Yuxuan Wang. 2022. Non-intrusive speech quality assessment with a multi-task learning based subband adaptive attention temporal convolutional neural network. In *Interspeech 2022*, pages 3298–3302.

Bao Thang Ta, Van Hai Do, and Huynh Thi Thanh Binh. 2024a. Enhancing Non-Matching Reference Speech Quality Assessment through Dynamic Weight Adaptation. In *Interspeech 2024*, pages 3859–3863.

Bao Thang Ta, Minh Tu Le, Van Hai Do, and Huynh Thi Thanh Binh. 2024b. Enhancing No-Reference Speech Quality Assessment with Pairwise, Triplet Ranking Losses, and ASR Pretraining. In *Interspeech 2024*, pages 2700–2704.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Zhe Ye, Jiahao Chen, and Diqun Yan. 2022. Residual-guided non-intrusive speech quality assessment. *arXiv preprint arXiv:2203.11499*.