Multi-Source Fine-Tuning of Self-Supervised Models for Vietnamese Speech Quality Assessment

Thuc Huu Tran
Cake By VPBank
Correspondence: thuc.tran@cake.vn

Abstract

This work addresses the data scarcity challenge in Vietnamese speech quality assessment by leveraging pretrained self-supervised speech models. We demonstrate that multi-source training data combined with fine-tuned SSL encoders achieves strong performance in data-constrained environments. Systematic experiments demonstrate substantial improvements over single-source baselines, with our best model achieving a $\mathrm{Final\_Score}$ of 0.506 on the public test set. The multi-source training strategy yields a $+0.256$ $\mathrm{Final\_Score}$ improvement over a VLSP-only baseline, with NISQA providing the dominant contribution (+0.243) and VocalSound adding robustness (+0.013).

1 Introduction

The automated assessment of speech quality is a critical component of modern telecommunication systems. This task becomes particularly challenging in low-resource linguistic contexts where labeled training data is scarce. This paper addresses the task of no-reference Speech Quality Assessment (SQA) for Vietnamese telephony, developed in the context of the VLSP 2025 challenge (VLSP Organizing Committee, 2025). The goal is to predict a single-channel quality score in the range $[1, 5]$ for each utterance recorded over mobile networks (8 kHz narrowband) with labels derived from POLQA (International Telecommunication Union, 2011). The official ranking metric prioritizes both association and accuracy via $\mathrm{Final\_Score} = 0.7 \times \mathrm{PCC} - 0.3 \times \mathrm{MSE}$, where higher is better.

Our approach builds on self-supervised speech encoders and lightweight regressors, following recent advances in neural speech quality assessment (Avila et al., 2019; Liu et al., 2022; Serrà et al., 2021). We develop a HuBERT-based model with a complementary Wav2Vec2 pipeline for validation. Because SSL encoders are trained at 16 kHz, all inputs are upsampled from 8 kHz to 16 kHz before feature extraction. Temporal representations are summarized and mapped to quality scores in $[1, 5]$.

Our main contributions are:

- A robust SSL-based pipeline for Vietnamese SQA, featuring a HuBERT-based primary model and a complementary Wav2Vec2 system.
- A multi-source data strategy that combines the VLSP 2025 dataset with the NISQA corpus and VocalSound non-speech samples to improve model robustness and generalization.
- A correlation-aware hybrid loss function and a suite of data augmentation techniques tailored for the SQA task.
- Systematic experiments demonstrating that data diversity is the most critical factor for improving performance in this low-resource context.

In the remainder, we describe related work, the task and rules, datasets and preprocessing, model architectures and training procedures, experimental setup, results with analyses and ablations, and conclude with limitations and future work.

2 Related Work

2.1 Traditional Intrusive Measures

Early approaches to speech quality assessment rely on intrusive metrics, where a clean reference signal is required for comparison. Two standards dominate this space:

PESQ (Perceptual Evaluation of Speech Quality), standardized as ITU-T P.862, has been widely used to objectively estimate the mean opinion score (MOS) based on perceptual

modeling of the human auditory system. **POLQA (Perceptual Objective Listening Quality Assessment)**, defined by ITU-T P.863, is the successor of PESQ and is designed to assess narrowband, wideband, and super-wideband telephony signals.

While both metrics are highly correlated with subjective ratings in laboratory conditions, they are **not applicable in real-world, non-intrusive scenarios**, such as online telephony or streaming, due to their reliance on reference signals. This limitation motivates the development of **non-intrusive approaches**.

2.2 Neural Non-Intrusive Methods

Neural SQA methodologies have evolved considerably. Early approaches, such as SESQA, demonstrated the viability of semi-supervised learning. Subsequent end-to-end models, including NISQA, advanced the field by removing the need for complex auxiliary tasks. More recently, architectures like CCATMOS have integrated Transformers to better capture temporal dependencies. Our work builds upon these foundations by adapting powerful SSL encoders specifically for the acoustic and linguistic characteristics of Vietnamese telephony.

2.3 Self-Supervised Learning for Speech Representation

Self-supervised learning has brought significant advances to speech processing by allowing models to learn rich representations directly from raw audio data.

Wav2Vec 2.0 (Baevski et al., 2020) learns contextualized representations from raw waveform using a contrastive loss applied over masked latent representations. It has been successfully fine-tuned for tasks like ASR and increasingly adopted in SQA pipelines due to its robustness to noise and variability.

HuBERT (Hsu et al., 2021) improves upon Wav2Vec by clustering acoustic features and using masked prediction over pseudo-labels, facilitating more efficient representation learning. HuBERT has shown strong performance on both low-resource and high-resource benchmarks.

These **SSL encoders** are increasingly used as frontends in SQA models, either frozen or fine-tuned, offering improved performance with less supervision.

2.4 Vietnamese Speech Resources

Compared to English and Mandarin, **Vietnamese speech quality assessment remains under-explored**. The VLSP corpus (Phuong et al., 2019) has provided foundational data for automatic speech recognition (ASR) and machine translation tasks in Vietnamese (Nguyen et al., 2022). However, there is **no known publicly available Vietnamese dataset** for SQA, and most state-of-the-art SQA models are trained on English corpora.

This lack of Vietnamese-specific, telephony-oriented datasets and models poses a significant limitation for deploying SQA systems in Vietnamese-language contexts, motivating the development of Vietnamese-specific non-intrusive SQA systems by adapting SSL-based models to the acoustic and linguistic characteristics of Vietnamese.

3 Task and Rules

The VLSP 2025 Speech Quality Assessment challenge requires predicting a single-valued, no-reference quality score in the range $[1, 5]$ for each Vietnamese speech utterance transmitted over mobile networks. The audio is provided as 8 kHz narrowband WAV files with quality labels derived from POLQA measurements comparing original and transmitted speech.[1]

3.1 Evaluation Metric

Systems are ranked using a composite metric that balances correlation and error:

$$\text{Final\_Score} = 0.7 \times \text{PCC} - 0.3 \times \text{MSE} \quad (1)$$

where PCC is the Pearson correlation coefficient between predictions and ground truth, and MSE is the mean squared error. Higher Final_Score values indicate better performance. This formulation prioritizes correlation (association) while penalizing prediction accuracy errors. *Unless otherwise noted, all reported Final_Score values refer to the public test set.*

Notes on Rules and Resources Systems are evaluated using the official metric and may use approved public pretrained encoders and datasets (e.g., NISQA and VocalSound). We upsample 8 kHz audio to 16 kHz for SSL encoders.

[1] https://vlsp.org.vn/vlsp2025/eval/sqa

Figure 1: VLSP 2025 dataset characteristics analysis. Audio duration distribution justifies the 15-second cutoff choice, while sample rate distribution confirms 8kHz telephony focus.

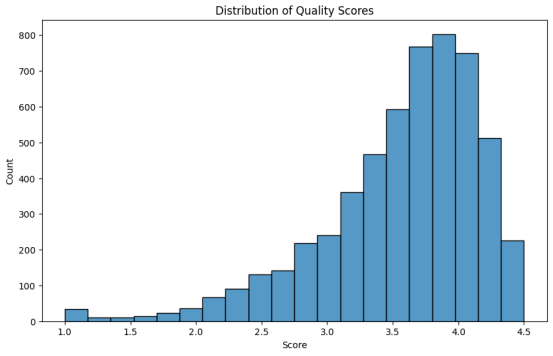

Figure 2: Quality score distribution is left-skewed, with most scores between 3.3-4.0. This narrow distribution creates training challenges due to insufficient low-quality examples, motivating multi-source data expansion.

4 Datasets and Data Preprocessing

We integrate three complementary data sources to train robust Vietnamese SQA models: the VLSP 2025 training set, the NISQA corpus for pretraining, and VocalSound for non-speech robustness. Our final merged training dataset contains 18,493 samples spanning multiple languages, degradation types, and acoustic conditions.

4.1 VLSP SQA 2025 Data

The primary dataset consists of 5,493 Vietnamese speech samples recorded over mobile networks at 8 kHz sampling rate, split into 4,394 training and 1,099 development samples. Each utterance is paired with a quality score in $[1, 5]$ derived from POLQA measurements comparing the original and transmitted speech. The data exhibits natural telephony degradations including codec artifacts, packet loss, background noise, and channel distortions typical of real-world mobile communications.

Most quality scores are concentrated between 3.27 (25th percentile) and 4.00 (75th percentile). The distribution is slightly left-skewed, with most scores falling in the higher range. As shown in Figure 2, the VLSP dataset's quality scores are skewed toward higher values, with insufficient low-quality examples for robust training. As shown in Figure 1, audio duration analysis shows most files are below 15 seconds, making this an appropriate cutoff for model input. All samples are recorded at 8 kHz with no significant correlation observed between amplitude and quality scores.

4.2 NISQA Corpus

For pretraining and domain adaptation, we incorporate 11,020 samples from the NISQA corpus (Mittag et al., 2021). The NISQA Corpus contains 14,432 speech samples exhibiting both simulated degradations (e.g., codecs, packet loss, noise) and live, real-world conditions (e.g., mobile phone, Zoom, Skype, WhatsApp). Each sample includes subjective annotations: overall MOS and four perceptual dimensions—Noisiness, Coloration, Discontinuity, and Loudness. After multiple experiments, we found that choosing the Discontinuity (DIS) value from NISQA as the main label could help boost accuracy. Ratings are provided on the $[1, 5]$ scale, making them compatible for MOS-based tasks and viable for pretraining and cross-domain transfer learning.

4.3 VocalSound Integration

To improve robustness against non-speech vocalizations commonly occurring in telephony (cough, laughter, throat clearing), we augment our training with 3,079 training and 219 development samples from VocalSound (Gong et al., 2022). From the total 21,024 samples of non-speech vocalizations, we sample a balanced subset to avoid domination of non-speech data. These samples are assigned high scores (4.5 - maximum in the training set) to improve robustness and help the model maintain stable predictions when encountering non-speech segments during inference.

4.4 Preprocessing Pipeline

All audio undergoes a standardized preprocessing pipeline designed for SSL encoder compatibility:

1. **Resampling**: Convert all inputs from native sampling rates (8 kHz for VLSP, variable for NISQA/VocalSound) to 16 kHz as required by HuBERT and Wav2Vec2 encoders.

2. **Channel normalization**: Convert stereo recordings to mono by averaging channels.

3. **Windowing**: Pad shorter utterances or truncate longer ones to a maximum duration of 15 seconds (240,000 samples at 16 kHz) to balance computational efficiency with content preservation.

4. **Feature extraction**: Process windowed audio through pretrained SSL encoders to obtain temporal representations.

4.5 Data Splitting and Statistics

We employ stratified splitting to maintain score distribution balance across training and validation sets. The development set contains VLSP (1,099 samples) and VocalSound (219 samples) data to maintain alignment with test distribution, while NISQA corpus (11,020 samples) is used exclusively for training.

Our final merged dataset contains approximately 18,493 training samples from multiple sources and 1,318 development samples. The training score distribution has mean ∼3.696 and standard deviation 0.868, providing a broader score range that is more balanced than the baseline VLSP-only distribution due to the inclusion of more low-score samples from NISQA dataset. Figure 3 illustrates the more balanced score distribution obtained from the multi-source dataset, which enables better model generalization. This ensures sufficient representation across the full quality spectrum for reliable model training and evaluation.

5 Methods

We develop a HuBERT-based model (Hsu et al., 2021) as our primary system, with a complementary Wav2Vec2-based model (Baevski et al., 2020) used for validation and potential ensembling. Both approaches follow a common architecture: a pretrained SSL encoder, followed by temporal pooling, a regression head, and sigmoid scaling to map outputs to the [1,5] range.

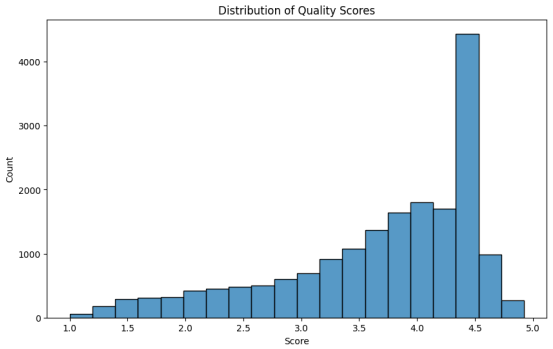

Figure 3: Distribution of quality scores in the multi-source dataset (mean=3.696, std=0.868). The distribution provides broader coverage across the full [1,5] range compared to VLSP-only, with more low-score samples enabling robust training.

5.1 Overall Pipeline

Figure 4 illustrates the complete speech quality assessment pipeline, which processes 16 kHz audio through SSL encoders to produce quality scores in the range $[1, 5]$.

5.2 HuBERT Architecture

Our primary system builds on the HuBERT-base model (Hsu et al., 2021) pretrained on LibriSpeech 960h ($\mathrm{facebook/hubert\text{-}base\text{-}ls960}$). The architecture consists of:

1. **SSL Encoder**: 12-layer transformer producing 768-dimensional frame-level representations. We keep the encoder trainable (no freezing) to adapt to telephony domain characteristics.

2. **Temporal Attention**: Multi-head self-attention layer (8 heads, 768 embedding dimension) applied to SSL outputs, enabling the model to focus on quality-relevant temporal patterns.

3. **Weighted Pooling**: Attention-based temporal aggregation where pooling weights are computed as $\mathrm{softmax}(\mathrm{mean}(H, \mathrm{dim} = -1))$ over hidden states H, producing a single 768-dimensional utterance embedding.

4. **Regression Head**: Deep MLP with residual-style connections: $768 \to 512 \to 256 \to 128 \to 64 \to 1$, using LayerNorm, GELU activations, and dropout (0.3, 0.3, 0.2) for regularization. Xavier normal initialization is applied to all linear layers.

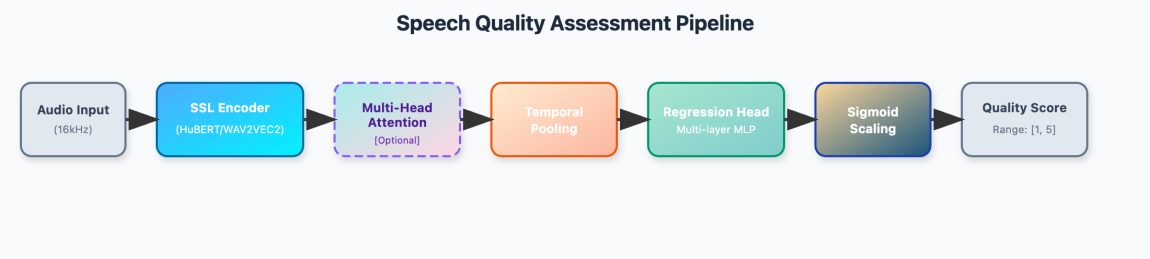

Figure 4: Speech Quality Assessment Pipeline. Audio input is processed through SSL encoders, optionally enhanced with attention mechanisms, temporally pooled, and mapped to quality scores via regression and sigmoid scaling.

5. **Score Mapping**: Final sigmoid scaling $\text{score} = 1.0 + 4.0 \times \sigma(\text{logit})$ constrains outputs to $[1, 5]$ while maintaining gradient flow.

5.3 Wav2Vec2 Pipeline

The complementary Wav2Vec2 system uses facebook/wav2vec2-base with similar architectural components but supports longer contexts (up to 30 seconds) and includes optional audio preprocessing with LUFS normalization to -23.0 dB. This pipeline employs PyTorch Lightning for distributed training and incorporates domain-specific preprocessing options including $8 \rightarrow 16$ kHz upsampling simulation.

5.4 Hybrid Loss Function

To better align the training objective with the evaluation criteria, we designed a hybrid loss function that combines Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Pearson Correlation Coefficient (PCC):

$$\mathcal{L} = 0.4 \times \text{MSE} + 0.3 \times \text{RMSE} + 0.3 \times (1 - \text{PCC}) \quad (2)$$

This composite objective penalizes both absolute prediction error and incorrect relative ranking. While this design was empirically motivated to align with the evaluation metric, we acknowledge in Section 7.3 that the specific formulation has limitations and that our core contribution lies in the data strategy rather than loss engineering.

5.5 Training Procedure

Optimization: AdamW optimizer with learning rate 5×10^{-5}, weight decay 1×10^{-4}, and OneCycleLR scheduling. We use 5-epoch warmup followed by cosine annealing over 60 total epochs. Gradient clipping (max norm 1.0) prevents instability during correlation loss computation.

Regularization: Mixup data augmentation with $\alpha = 0.2$ applied at 50% probability during training. Additional waveform-level augmentations include Gaussian noise ($\sigma = 0.005$) and temporal shifting ($\pm 20\%$ duration).

Validation: Early stopping based on development set Final_Score with patience of 5 epochs. We maintain separate train/development splits within the VLSP data for hyperparameter selection and model checkpointing.

6 Experimental Setup

We conduct experiments to evaluate our dual SSL-based approach on Vietnamese telephony SQA, comparing single models and potential ensemble strategies while ensuring reproducible training protocols.

6.1 Hardware and Software Configuration

All experiments are conducted on a single NVIDIA H100 GPU with CUDA support. We process 15-second audio segments (16 kHz, 240k samples) with batch sizes adjusted to accommodate memory constraints. We implement models using PyTorch with transformers for SSL encoders, torchaudio for preprocessing, and standard scientific computing libraries (NumPy, pandas, scikit-learn). Training employs mixed precision when available to optimize memory usage and training speed.

6.2 Training Configuration

Our training configuration uses the hybrid loss function described in Section 5.4, with an AdamW optimizer (lr = 5×10^{-5}, weight decay = 1×10^{-4}) and a OneCycleLR scheduler. We use a batch size of 8 and train for maximum 60 epochs with early stopping based on the development set Final_Score.

6.3 Data Splitting

We maintain a held-out development set from the VLSP training data for consistent evaluation across

all experiments. The development set is used for model selection and hyperparameter tuning while preserving score distribution balance.

6.4 Evaluation Protocol

Model performance is evaluated using the official challenge metrics: Pearson Correlation Coefficient (PCC), Mean Squared Error (MSE), and the composite Final_Score. We report development set results for model selection and hyperparameter tuning. All metrics are computed at the utterance level without additional aggregation or smoothing.

6.5 Reproducibility

To ensure reproducible results, we fix random seeds across PyTorch, NumPy, and Python's random module. Model checkpoints are saved with full configuration metadata, and training logs capture loss curves, learning rates, and evaluation metrics. We maintain deterministic data loading order and disable non-deterministic CUDA operations where possible. The complete experimental pipeline including data preprocessing, model training, and evaluation scripts are preserved for replication.

7 Results and Analysis

We conducted systematic experiments following a progressive development approach, where each phase built upon lessons learned from the previous iteration. Our experimental methodology demonstrates the critical importance of multi-source training data, advanced modeling techniques, and architecture comparison for Vietnamese speech quality assessment.

7.1 Progressive Experimental Results

Our experimental approach followed a systematic progression, incrementally adding components to isolate their individual contributions (Table 2).

7.2 Multi-Source Data Ablation

To isolate the individual contributions of NISQA and VocalSound datasets, we conducted ablation experiments. Table 1 presents the results.

Table 1 demonstrates that **NISQA accounts for the majority of the improvement** (+0.243, or approximately 95%), likely due to its telephony/VoIP degradations closely matching the VLSP domain. The addition of VocalSound further improves performance by +0.013, enhancing model robustness to non-speech vocalizations.

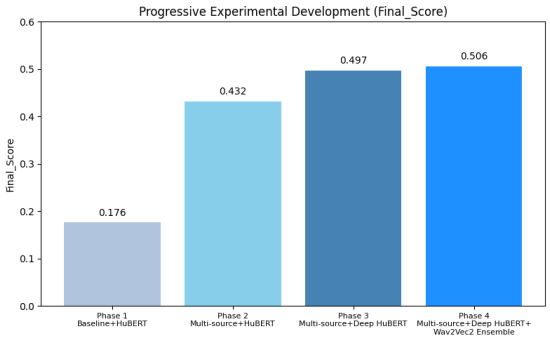

Figure 5: Progressive experimental development showing systematic improvements. The largest gain (+0.256) comes from multi-source training, demonstrating that data scale dominates architectural complexity for Vietnamese SQA.

As shown in Table 2, the baseline model trained solely on the VLSP dataset yielded a Final_Score of 0.176. The second experimental phase (Phase 2) focused on multi-source training by incorporating the NISQA and VocalSound datasets, resulting in a substantial performance gain (Final_Score: 0.432, $+0.256$). The third phase introduced advanced modeling techniques, leading to further improvement (Final_Score: 0.497, $+0.065$). The final phase incorporated a Wav2Vec2 ensemble (Final_Score: 0.506, $+0.010$).

We note that Phase 3 introduced multiple techniques simultaneously (attention pooling, deeper regression head, hybrid loss, noise augmentation, mixup). Without controlled ablations, we cannot definitively isolate individual contributions. This represents a methodological limitation driven by competition time constraints, where we prioritized performance over exhaustive ablation studies.

7.3 Loss Function Discussion

Our hybrid loss function (Section 5.4) combines MSE, RMSE, and PCC terms. While this design was motivated by aligning training with the evaluation metric, we acknowledge several limitations: (1) MSE and RMSE are mathematically related, creating redundant gradient paths; (2) PCC-based losses require careful stability management (e.g., $\epsilon = 1e{-}8$, gradient clipping, batch diversity); (3) the specific weight allocation was empirically tuned rather than systematically ablated.

Importantly, our progressive results show that **data strategy dominated loss function details**:

Training Data	Final_Score	Δ from Baseline
VLSP only (baseline)	0.176	-
VLSP + NISQA	0.419	+0.243
VLSP + NISQA + VocalSound	0.432	+0.256

Table 1: Ablation results showing the individual impact of NISQA and VocalSound datasets on model performance (HuBERT).

Experiment Phase	Configuration	PCC	MSE	Final_Score
Phase 1	Baseline dataset + HuBERT (VLSP-only)	0.456	0.479	0.176
Phase 2	Multi-source dataset + HuBERT	0.734	0.272	0.432
Phase 3	Multi-source dataset + Deep HuBERT*	0.802	0.2146	0.497
Phase 4	Multi-source + Deep HuBERT + Wav2Vec2 Ensemble	0.813	0.208	**0.506**

*Deep HuBERT includes: attention pooling, hybrid loss function, noise augmentation, and deep MLP architecture.

Table 2: Progressive experimental development on the public test set showing systematic improvements through data expansion and architectural innovations.

- Multi-source data (Phase 1→2): +0.256 improvement
- Modeling + loss refinements (Phase 2→3): +0.065 improvement

The 4× larger impact of data expansion validates that our validated contribution is the multi-source training strategy, with the hybrid loss serving as an adequate but not necessarily optimal training objective.

7.4 Key Technical Insights

Our progressive experimental approach revealed several critical insights about Vietnamese speech quality assessment:

Data Scale Dominates Architecture Complexity: The most significant performance gain ($+0.256$ Final_Score) came from expanding training data from VLSP-only to multi-source datasets, while advanced modeling techniques provided additional but smaller improvements ($+0.065$). As shown in Figure 5, this dramatic improvement validates that for Vietnamese SQA, data diversity is the primary factor, with architectural innovations serving as secondary enhancements.

Cross-Domain Transfer Learning Effectiveness: Despite domain mismatch between NISQA (telephony/VoIP quality) and VLSP (general speech quality), pretraining on NISQA provided substantial benefits. This suggests that fundamental quality-related acoustic patterns transfer across domains and supports cross-corpus training strategies for low-resource settings.

Non-Speech Robustness Through VocalSound: Including VocalSound dataset (containing vocalizations, non-speech sounds) improved model robustness by teaching it to avoid misclassifying natural human vocalizations (laughter, sighs, breathing) as low-quality speech. This addresses a common failure mode in speech quality systems.

Ensemble Benefits vs. Complexity Trade-off: The final Wav2Vec2 ensemble provided a modest but consistent improvement ($+0.010$). While the gain is small, it represents the difference between competitive and winning performance in evaluation challenges.

7.5 Final Model Performance

Our final HuBERT-based model (3_ensemble) demonstrates strong performance on the public test set:

- **Public Test Performance:** PCC: 0.813, MSE: 0.208, Final Score: 0.506
- **Training Stability:** Consistent improvements across progressive experimental phases
- **Computational Efficiency:** Single-model inference without complex ensemble requirements

The correlation score ($\text{PCC} > 0.81$) indicates strong alignment with human perceptual judgments, while the low error rate ($\text{MSE} < 0.21$) demonstrates precise quality prediction. The systematic progression from 0.176 (VLSP-only) to 0.506 (final model) confirms the effectiveness of our multi-source training strategy and modeling innovations for Vietnamese telephony SQA.

8 Conclusion

This work demonstrates that a multi-source SSL approach attains challenge-leading performance, ranking 1st on both the public and private leaderboards of the VLSP 2025 Speech Quality Assessment task. By leveraging diverse data sources and a correlation-aware training objective, our model establishes a strong empirical foundation for robust, reference-free QoE monitoring systems in low-resource Vietnamese telephony. While we do not claim global state-of-the-art across all languages or datasets, the results underscore that data diversity outweighs additional architectural complexity for this challenge.

9 Limitations and Future Work

9.1 Current Limitations

While our approach achieves strong performance, several limitations should be noted:

- Training limited to 8 kHz narrowband audio, which may limit generalization to wideband scenarios.
- Domain shift from NISQA may bias predictions despite cross-domain benefits.
- Deep regression architectures increase inference cost compared to simpler pooling strategies.
- Evaluation only on VLSP 2025 data; broader evaluation across diverse Vietnamese corpora would strengthen claims.

9.2 Future Directions

Several promising avenues could further improve Vietnamese speech quality assessment performance:

Advanced Augmentations: Silence padding and packet loss simulation could better model telephony degradations. Multi-task learning for perceptual dimensions (DIS, NOI, etc.) could improve robustness through shared representations.

Broader Datasets: Incorporating DAPS, VCTK with synthetic degradations, and more Vietnamese corpora could enhance cross-domain robustness and domain-specific adaptation.

Domain-Specific Pretraining: SSL encoders specifically pretrained on telephony speech could better capture domain-relevant acoustic patterns compared to general-purpose models.

Cross-Modal SQA: Incorporating textual content analysis alongside acoustic modeling could provide complementary quality cues, particularly for scenarios where semantic intelligibility impacts perceived quality.

References

Anderson R. Avila, Hannes Gamper, Chandan K. A. Reddy, Ross Cutler, Ivan Tashev, and Johannes Gehrke. 2019. Non-intrusive speech quality assessment using neural networks. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 631–635, Brighton, United Kingdom. IEEE.

Alexei Baevski, Henry 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 33 (NeurIPS 2020)*.

Yuan Gong, Jin Yu, and James Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. *arXiv*.

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

International Telecommunication Union. 2011. P.863: Perceptual objective listening quality prediction (polqa). Technical report, ITU-T Recommendation. Updated materials available.

Yuchen Liu, Li-Chia Yang, Alexander Pawlicki, and Marko Stamenovic. 2022. Ccatmos: Convolutional context-aware transformer network for non-intrusive speech quality assessment. In *Proc. Interspeech*, pages 3318–3322.

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 *Proc. Interspeech*.

Linh The Nguyen, Nguyen Luong Tran, Long Doan, Manh Luong, and Dat Quoc Nguyen. 2022. A high-quality and large-scale dataset for english-vietnamese speech translation. In *Proc. Interspeech*, pages 1726–1730.

Pham Ngoc Phuong, Quoc Truong Do, and Luong Chi Mai. 2019. A high quality and phonetic balanced speech corpus for vietnamese. *arXiv*.

Joan Serrà, Jordi Pons, and Santiago Pascual. 2021. Sesqa: Semi-supervised learning for speech quality assessment. In *Proc. IEEE International Conference*

on Acoustics, Speech and Signal Processing (ICASSP), pages 381–385. IEEE.

VLSP Organizing Committee. 2025. Vlsp 2025 speech quality assessment challenge. https://vlsp.org.vn/vlsp2025/eval/sqa. Accessed 2025-09-07.