VLSP 2025 Shared Task: Speech Quality Assessment for Vietnamese Telecommunication

Bao Thang Ta[1], Minh Tu Le[2], Quang Trung Le[3], Van Hai Do[4],
[1]Hanoi University of Science and Technology, [2]WorldQuant, [3]Torilab, [4]Thuyloi University
Correspondence: haidv@tlu.edu.vn

Abstract

This paper presents the VLSP 2025 shared task on Speech Quality Assessment (SQA) for Vietnamese telecommunication. The task addresses the challenge of predicting perceptual quality scores for speech transmitted over mobile networks without requiring reference signals. Participants were provided with a Vietnamese dataset containing speech samples recorded over real mobile networks, each annotated with POLQA-based quality scores ranging from 1 to 5. The shared task attracted 16 registered teams, with 7 teams submitting results on the public test set and 5 teams on the private test set. This paper describes the task definition, dataset creation methodology, evaluation metrics, participating systems, and analysis of results.

1 Introduction

Speech Quality Assessment (SQA) plays a critical role in modern telecommunication systems, enabling service providers to monitor and maintain high-quality user experiences. With the increasing reliance on mobile networks and Voice over IP (VoIP) services, the ability to automatically assess speech quality has become essential for network optimization and quality control (Mittag et al., 2021).

Traditional speech quality assessment methods, such as PESQ (Perceptual Evaluation of Speech Quality) (ITU-T Recommendation P.862, 2001) and POLQA (Perceptual Objective Listening Quality Assessment) (ITU-T Recommendation P.863, 2011), require access to both the original and degraded signals, making them impractical for real-time monitoring in production environments. Non-reference (or no-reference) (Le et al., 2023; Liang et al., 2023; Jayesh et al., 2022; Ta et al., 2023) SQA methods address this limitation by predicting quality scores directly from degraded speech signals, enabling scalable and cost-effective quality monitoring.

The VLSP 2025 Speech Quality Assessment shared task focuses on developing non-reference SQA systems specifically for Vietnamese telecommunication data. Vietnamese, a tonal language with unique phonetic characteristics, presents distinct challenges for speech quality assessment compared to well-studied languages like English and Mandarin (Yu et al., 2021; Manocha et al., 2022; Ta et al., 2024). The lack of large-scale Vietnamese SQA datasets and specialized models motivates the need for dedicated shared tasks and resources.

1.1 Task Description

The VLSP 2025 SQA shared task requires participants to develop models that predict quality scores in the range [1, 5] for Vietnamese speech samples transmitted over mobile networks. The audio data consists of 8 kHz narrowband recordings with quality labels derived from POLQA measurements comparing original and transmitted speech. Systems are evaluated using a composite metric that balances correlation and prediction accuracy:

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

where PCC is the Pearson Correlation Coefficient and MSE is the Mean Squared Error. This formulation prioritizes correlation with human perception while penalizing prediction errors.

1.2 External Resources

The VLSP SQA 2025 competition allows participants to use external resources, including pre-trained models and publicly available datasets. Teams were required to propose their intended external resources prior to the competition. These proposals were then reviewed and voted on by all participating teams, and finally approved by the organizers based on criteria such as accuracy, popularity, size, and fairness. The approved resources included:

- Pretrained self-supervised speech models: HuBERT (Hsu et al., 2021), Wav2Vec2 (Baevski et al., 2020), and WavLM (Chen et al., 2022) (base version)
- NISQA corpus models (Mittag et al., 2021) for cross-domain pretraining
- Unlabeled data for self-supervised pretraining: LibriSpeech (Panayotov et al., 2015); noisy data: MUSAN noise (Snyder et al., 2015) and VocalSound (Gong et al., 2022)
- SQA tools such as TORCHAUDIO-SQUIM (Kumar et al., 2023) and pretrained NISQA (Yi et al., 2022)

1.3 Contributions

The main contributions of this shared task include:

- This is the first time the SQA task has been organized for Vietnamese [1].
- We build a comprehensive Vietnamese telecommunication speech quality dataset consisting of 9,431 samples annotated using professional POLQA-based measurements. The dataset is publicly available to facilitate future research and promote this task in the Vietnamese language.
- We provide an analysis of state-of-the-art approaches from participating teams, demonstrating the effectiveness of multi-source training and self-supervised learning.

2 Task Definition and Rules

2.1 Problem Formulation

The VLSP 2025 SQA task is formulated as a regression problem where systems must predict a continuous quality score $\hat{y} \in [1, 5]$ for each input speech sample x. The ground truth labels y are derived from POLQA measurements, which provide objective estimates of perceived speech quality by comparing reference and degraded signals.

2.2 Data Format

Input audio is provided as WAV files with the following specifications:

- Sampling rate: 8 kHz (narrowband telephony)
- Format: Single-channel (mono)
- Encoding: PCM 16-bit

Output predictions must be submitted in tab-separated format:

`utterance_name<TAB>SQA_score`

where `utterance_name` excludes file extensions (e.g., `0001` rather than `0001.wav`).

2.3 Evaluation Metrics

System performance is evaluated using three metrics:

Pearson Correlation Coefficient (PCC) measures linear correlation between predicted and ground truth scores:

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

Mean Squared Error (MSE) measures average prediction error:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \quad (3)$$

Final Score is the official ranking metric combining PCC and MSE:

$$\text{Final\_Score} = 0.7 \times \text{PCC} - 0.3 \times \text{MSE} \quad (4)$$

This composite metric ensures systems achieve both high correlation with human judgments (PCC) and low absolute prediction errors (MSE).

3 Dataset

3.1 Data Collection Methodology

The VLSP 2025 SQA dataset was created using a systematic approach to capture realistic telecommunication degradations. The methodology, adapted from professional network quality assessment practices, involves the following steps:

3.1.1 Reference Speech Preparation

High-quality Vietnamese speech samples were recorded as reference material, featuring:

- Diverse speakers (male and female) across various ages ranging from 20 to 45
- Multiple Vietnamese dialects, covering all three major regions: Northern, Central, and Southern
- Natural conversations that include various vocal sounds such as laughter, coughing, and more

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

3.1.2 Network Transmission

Reference speech was transmitted through real mobile networks using the Nemo Handy professional measurement platform developed by Keysight Technologies. This platform provides:

- Full multi-technology support (2G to 5G networks)
- POLQA implementation compliant with ITU-T P.863
- Real-time RF parameter logging
- Professional-grade measurement accuracy

3.1.3 Quality Measurement

The Nemo Handy system computed POLQA scores by comparing transmitted audio against original references. Data collection was performed across diverse conditions:

- Urban environments with varying signal conditions
- Indoor/outdoor locations
- Moving scenarios (vehicles, elevators)
- Different times of day and network congestion levels

3.2 Dataset Characteristics

The final dataset contains 9,431 Vietnamese speech samples with the following properties:

Property	Value
Total samples	9,431
Training set	5,493
Public test set	1,717
Private test set	2,221
Sampling rate	8 kHz
Quality score range	[1, 5]
25th percentile	3.27
75th percentile	4.00

Table 1: VLSP 2025 SQA dataset statistics

3.3 Quality Score Distribution

The quality score distribution exhibits characteristics typical of real-world telecommunication systems, with most samples concentrated in the mid-to-high quality range (3.25–4.00). This distribution reflects modern mobile network performance, where severe degradations are relatively rare but subtle quality variations are common. The limited number of low-quality samples (scores below 2.5) presents a challenge for model training and motivates the use of external data augmentation strategies.

3.4 Degradation Types

The dataset captures various degradation types commonly encountered in real mobile telecommunication:

- Codec artifacts (AMR, EVS codecs)
- Packet loss and jitter
- Background noise and environmental interference
- Signal attenuation and fading
- Handover effects
- Echo and reverberation

4 Participating Systems

The shared task attracted 16 registered teams, with 7 teams submitting results on the public test set and 5 teams on the private test set. This section summarizes the approaches adopted by teams that submitted technical reports. Both teams demonstrated sophisticated architectures leveraging self-supervised learning and multi-source training strategies.

4.1 Team 1 (Cake by VPBank): Multi-Source SSL Approach

The first-place team developed a HuBERT-based system enhanced with multi-source training data. Their approach consists of:

4.1.1 Architecture

- **SSL Encoder:** HuBERT-base (facebook/hubert-base-ls960) with 12 transformer layers producing 768-dimensional frame-level representations
- **Temporal Attention:** Multi-head self-attention (8 heads) to focus on quality-relevant temporal patterns
- **Weighted Pooling:** Attention-based aggregation producing utterance-level embeddings
- **Regression Head:** Deep MLP ($768 \to 512 \to 256 \to 128 \to 64 \to 1$) with LayerNorm, GELU, and dropout

- **Score Mapping:** Sigmoid scaling to constrain outputs to [1, 5]

4.1.2 Multi-Source Training Strategy

The team combined three complementary data sources:

1. **VLSP 2025 training data (5,493 samples):** Primary Vietnamese telecommunication data
2. **NISQA corpus (11,020 samples):** Pretraining on English telephony/VoIP data with Discontinuity (DIS) labels
3. **VocalSound (3,079 samples):** Non-speech vocalizations assigned high scores (4.5) for robustness

Total training data: 18,493 samples with broader score distribution (mean=3.696, std=0.868).

4.1.3 Training Configuration

- **Loss Function:** Hybrid loss combining MSE, RMSE, and PCC: $\mathcal{L} = 0.4 \times \text{MSE} + 0.3 \times \text{RMSE} + 0.3 \times (1 - \text{PCC})$
- **Optimizer:** AdamW with learning rate 5×10^{-5}, weight decay 1×10^{-4}
- **Scheduler:** OneCycleLR with 5-epoch warmup, 60 total epochs
- **Augmentation:** Mixup ($\alpha = 0.2$), Gaussian noise, temporal shifting
- **Preprocessing:** Upsampling from 8 kHz to 16 kHz, 15-second windowing

4.1.4 Results

Results on the public test set for Team 1 are shown in Table 2. The ablation study revealed that incorporating NISQA contributed an improvement of +0.243 (accounting for 95% of the total gain), while adding VocalSound provided an additional +0.013 improvement, enhancing robustness to non-speech segments. These results confirm that all proposed components contribute to the overall model performance.

4.2 Team 2 (UET): Multi-Task Learning with ListNet

The second-place team proposed EM-VSQA, a wav2vec2+BiLSTM framework enhanced with multi-task learning and targeted external data.

4.2.1 Architecture

- **Encoder:** wav2vec2-base for frame-level acoustic representations
- **Temporal Modeling:** BiLSTM (hidden=256, dropout=0.3) for capturing temporal dependencies
- **Regression Head:** MLP (256→128→64→1) with ReLU activations

4.2.2 Multi-Task Learning

The team employed a dual-objective loss function:

$$\mathcal{L} = \lambda \mathcal{L}_{\text{MSE}} + (1 - \lambda)\mathcal{L}_{\text{ListNet}} \tag{5}$$

where ListNet (Cao et al., 2007) loss encourages ranking consistency:

$$\mathcal{L}_{\text{ListNet}} = -\sum_{i=1}^{N} P(y_i) \log P(\hat{y}_i) \tag{6}$$

with probability distributions defined as:

$$P(y_i) = \frac{\exp(y_i)}{\sum_{j=1}^{N} \exp(y_j)}, \quad P(\hat{y}_i) = \frac{\exp(\hat{y}_i)}{\sum_{j=1}^{N} \exp(\hat{y}_j)} \tag{7}$$

4.2.3 External Data Strategy

Rather than joint training, the team incorporated VocalSound selectively:

- 250 VocalSound samples (5% of VLSP size) assigned MOS=5
- Training on combined VLSP+VocalSound with balanced sampling
- Improved handling of non-speech events (laughter, breathing) without overfitting

4.2.4 Results

Results on the public test set for Team 2 are shown in Table 3. The ablation study revealed that incorporating NISQA led to a substantial improvement. Furthermore, applying ListNet in the multi-task setting significantly increased the PCC of the predicted results, resulting in an additional gain of approximately 0.05 in the final score.

Table 2: Final score comparison on the public test set for different Team 1 model configurations. *Deep HuBERT* is a version using a deeper MLP architecture.

Model Configuration	Final_Score	Improvement
HuBERT + VLSP-only	0.176	-
HuBERT + VLSP + NISQA	0.419	+0.243
HuBERT + VLSP + NISQA + VocalSound	0.432	+0.256
Deep HuBERT* + VLSP + NISQA + VocalSound	0.497	+0.321
Deep HuBERT* + Wav2Vec2 ensemble + VLSP + NISQA + VocalSound	0.506	+0.330

Table 3: Ablation results for Team 2 on the public test set.

Configuration	PCC	MSE	Score
VLSP only	-0.061	1.300	-0.433
+ External data	0.746	0.268	0.441
+ Multi-task	0.798	0.211	0.495

4.3 Comparison of Approaches

Both top teams converged on several key strategies:

- **Self-supervised encoders:** HuBERT and wav2vec2 provide robust acoustic representations
- **Multi-source training:** External data dramatically improves generalization
- **Advanced loss functions:** Combining regression and correlation/ranking objectives
- **VocalSound integration:** Improves robustness to non-speech events
- **Upsampling:** 8 kHz→16 kHz conversion for SSL encoder compatibility

Key differences:

- Team 1 used larger external datasets (11k+ NISQA samples) with joint training
- Team 2 employed more selective augmentation (250 VocalSound samples) with multi-task learning
- Team 1 focused on architectural depth (attention, deep MLP)
- Team 2 emphasized loss function design (ListNet ranking)

5 Results and Analysis

Table 4 presents the official results on the private test set.

Some key findings are as follows:

Data Scale Dominates Architecture Complexity: The most significant finding across participating teams is that data diversity and scale provide larger performance gains than architectural innovations alone. Team 1's ablation study demonstrated a +0.256 improvement from multi-source data versus +0.065 from advanced modeling techniques. This 4× larger impact validates that for low-resource Vietnamese SQA, expanding training data is the primary factor for success.

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

Non-Speech Robustness: Both top teams incorporated VocalSound data to handle non-speech vocalizations (laughter, coughs, breathing). Team 2's analysis revealed that models trained only on VLSP often misclassified high-quality speech with laughter or clean vocalizations as degraded, assigning inappropriately low scores. VocalSound integration significantly improved this failure mode.

Loss Function Design: Advanced loss functions combining MSE with correlation (PCC) or ranking (ListNet) objectives consistently outperformed simple MSE-based training. These multi-objective losses encourage models to learn both absolute score prediction and relative quality relationships, aligning better with the evaluation metric.

Table 4: Official results on the private test set. Team 1 achieved the highest score on the leaderboard. Team 2 (Wav2Vec2 Explicit Multitask*) explored an ensemble variant (EEM-VSQA) with explicit VocalSound classification, but did not submit it to the competition.

Team	PCC	MSE	Score
Team 1 (HuBERT Multi-Source)	0.795	0.227	0.489
Team 2 (Wav2Vec2 Multi-Task)	0.762	0.338	0.432
Team 2 (Wav2Vec2 Explicit Multitask)*	0.812	0.254	0.492
Team 3	0.516	0.649	0.167
Team 4	0.428	0.617	0.115
Team 5	0.143	0.946	-0.183
Pretrained NISQA (distortion score) (baseline)	0.384	0.747	0.0449

6 Discussion

6.1 Success Factors

Several factors contributed to the strong performance of top systems:

1. **Pretrained SSL Models:** HuBERT and wav2vec2 provide robust acoustic representations learned from large-scale unlabeled data, transferring effectively to Vietnamese SQA.

2. **Multi-Source Training:** Combining VLSP with NISQA and VocalSound addresses data scarcity and distribution imbalance, particularly for low-quality and edge-case samples.

3. **Correlation-Aware Training:** Loss functions incorporating PCC or ranking objectives align training directly with evaluation metrics.

6.2 Remaining Challenges

Despite strong results, several challenges remain:

Computational Efficiency Deep SSL models require significant computational resources. Developing lightweight models or distillation approaches would enable broader deployment.

Interpretability Current models function as black boxes. Incorporating perceptual dimensions (as in NISQA's multi-task approach) could provide interpretable quality diagnostics.

Real-Time Processing Most submitted systems focus on accuracy rather than latency. Streaming-capable architectures would enable real-time monitoring applications.

7 Conclusion

The VLSP 2025 Speech Quality Assessment shared task successfully established a benchmark for non-reference Vietnamese telecommunication speech quality assessment. The task attracted 16 registered teams, with 7 submissions to the public test and 5 submissions to the private test, demonstrating strong community interest. The dataset and competition report are made publicly available to encourage future research in this field.

Acknowledgments

Bao Thang TA was funded by the Master, PhD Scholarship Programme of Vingroup Innovation Foundation (VINIF), code VINIF.2024.TS.074.

References

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

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.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, and 1 others. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Y. Gong, J. Yu, and J. 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).

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.

ITU-T Recommendation P.862. 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

ITU-T Recommendation P.863. 2011. Perceptual objective listening quality assessment.

M K Jayesh, Mukesh Sharma, Praneeth Vonteddu, Mahaboob Ali Basha Shaik, and Sriram Ganapathy. 2022. Transformer Networks for Non-Intrusive Speech Quality Prediction. In *INTERSPEECH 2022*, pages 4078–4082.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

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

Xinyu Liang, Fredrik Cumlin, Christian Schüldt, and Saikat Chatterjee. 2023. DeePMOS: Deep Posterior Mean-Opinion-Score of Speech. In *INTERSPEECH 2023*, pages 526–530.

Pranay Manocha, Anurag Kumar, Buye Xu, Anjali Menon, Israel Degene Gebru, Vamsi Krishna Ithapu, and Paul Calamia. 2022. SAQAM: Spatial Audio Quality Assessment Metric. In *INTERSPEECH 2022*, pages 649–653.

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. *INTERSPEECH 2021*, pages 2127–2131.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *ICASSP 2015*, pages 5206–5210.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. ArXiv:1510.08484v1.

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

Bao Thang Ta, Minh Tu Le, Nhat Minh Le, and Van Hai Do. 2023. Probing Speech Quality Information in ASR Systems. In *INTERSPEECH 2023*, pages 541–545.

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. 2022. ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications. In *INTERSPEECH 2022*, pages 3308–3312.

Meng Yu, Chunlei Zhang, Yong Xu, Shi-Xiong Zhang, and Dong Yu. 2021. MetricNet: Towards Improved Modeling For Non-Intrusive Speech Quality Assessment. In *INTERSPEECH 2021*, pages 2142–2146.