Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

SQ-AST: A Transformer-Based Model for Speech Quality Prediction

Wafaa Wardah[1], *Robert Spang*[1], *Vincent Barriac*[2], *Jan Reimes*[3], *Anna Llagostera*[4], *Jens Berger*[4], *Sebastian Möller*[1]

[1]Quality and Usability Lab, Faculty IV, Technische Universität Berlin, Germany
[2]Orange Innovation, France
[3]HEAD Acoustics GmbH, Germany
[4]Rohde & Schwarz SwissQual AG, Switzerland

wafaa.wardah@tu-berlin.de

Abstract

We present SQ-AST, a transformer-based model for non-intrusive speech quality prediction. The model predicts overall speech quality and four perceptual dimensions—noisiness, discontinuity, coloration, and loudness—using only the degraded signal. SQ-AST leverages Audio Spectrogram Transformers (AST), pretrained on large-scale audio datasets and fine-tuned on diverse speech quality corpora. It operates on short speech clips (4–12 seconds) without requiring a reference signal. Training was conducted on 106 databases comprising 165,791 samples. Independent evaluations confirm strong generalization to real-world conditions. The model is currently under consideration for ITU-T standardization, highlighting its potential for benchmarking, quality assessment, and industry adoption.

Index Terms: speech quality assessment, deep learning, audio spectrogram transformer, perceptual evaluation, machine learning

1. Introduction

Speech quality assessment is a critical component in evaluating the performance of telecommunication networks and modern speech processing systems. The measure is typically expressed as a Mean Opinion Score (MOS), derived from subjective listening tests following ITU-T Rec. P.800 [1], ITU-T Rec. P.808 [2, 3], and ITU-T Rec. P.806 [4]. Since subjective testing is time-consuming and costly, objective models have been developed to automate speech quality prediction. Traditional objective metrics, such as P.863 (POLQA) [5] and P.863.2 [6], rely on reference-based signal comparisons, making them less applicable in scenarios where reference signals are unavailable.

Advancements in non-intrusive speech quality models, such as NISQA [7] and DNSMOS [8], have demonstrated strong performance across various conditions. These models have contributed significantly to the field by improving single-ended speech quality assessment. However, as modern communication systems increasingly adopt machine learning-based codecs, new types of distortions emerge that require further adaptation in assessment models. Additionally, existing approaches often evaluate structured speech samples, which may not fully capture the variability and distortions present in live conversational speech.

To address these challenges, we introduce a new Transformer-based model designed for non-intrusive speech quality assessment. Our approach predicts overall MOS and four perceptual dimensions—Noise (NOI), Discontinuity (DIS), Coloration (COL), and Loudness (LOUD)—on a 1-5 ACR-like scale. Unlike previous models, it is specifically designed to evaluate live, conversation-style speech audio, making it better suited for real-world applications. By leveraging Transformer-based architectures, we aim to improve prediction accuracy for natural speech interactions, enhancing robustness in diverse telecommunication environments.

In this work, we leverage Audio Spectrogram Transformers (AST) [9], a deep-learning approach originally developed for audio classification, to model perceptual speech quality directly from degraded speech waveforms (4-12 seconds in length). Our proposed model operates on 48kHz fullband audio without requiring a reference signal, making it suitable for evaluating speech quality in real-world, uncontrolled environments.

The model integrates five specialized AST modules, each fine-tuned for a specific prediction task: MOS, NOI, DIS, COL, and LOUD. Each module is initialized with pretrained AST weights and further trained on speech quality datasets to optimize performance across diverse conditions. This modular structure enhances generalization across different speech processing tasks while allowing for dimension-specific improvements. The PyTorch code and trained weights are open-sourced[1] .

Our approach has been trained and validated on an extensive set of databases, incorporating new datasets that expand the scope of data-driven speech quality prediction beyond traditional rule-based and reference-based approaches. Given its performance, this work is being evaluated for potential inclusion in ITU-T standardization efforts for non-intrusive speech quality assessment methods.

2. Method

The proposed model is comprised of five AST modules that were individually trained to predict the overall speech quality, noisiness, coloration, discontinuity, and loudness. For inference, the frozen fine-tuned AST modules are integrated so that the input waveform is processed once before being passed through the five task AST modules, producing the five predicted outputs. An overview of the model architecture is shown in Figure 1.

In the proposed model, the feature extraction pipeline follows the original design proposed by the AST authors, with some modifications. While the original implementation used waveforms sampled at 16 kHz, our approach processes audio sampled at 48 kHz for higher fidelity. Log-Mel spectrograms are computed from the input waveform using a 25 ms window size, a 10 ms hop size, and 128 Mel frequency bins. For audio clips shorter than 10 seconds, an attention mask is applied to handle variable-length sequences. The resulting spectrogram is divided into overlapping patches of size 16×16, with overlaps applied in both the time and frequency dimensions. Each

[1]https://github.com/WafaaWardah/SQ-AST

10.21437/Interspeech.2025-2683

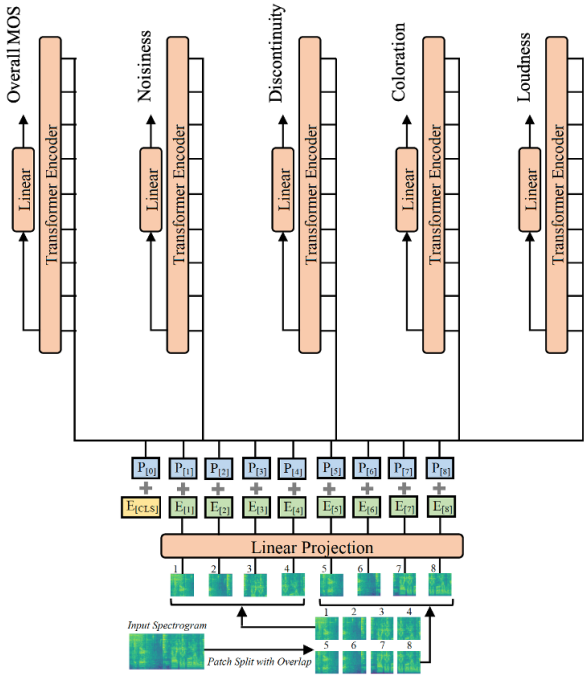


Figure 1: *The AST-based feature extraction pipeline and model architecture. The input 2D audio spectrogram is divided into overlapping* 16×16 *patches and converted into 1D embeddings with positional encoding. A classification token is appended, and the sequence is processed by five Transformer encoders, each dedicated to one of the prediction tasks: overall speech quality, noisiness, discontinuity, coloration, and loudness. The final outputs are passed through separate linear layers with Sigmoid activation for prediction. This diagram is adapted from [9].*

patch is flattened into a 1D vector of size 768 through a linear projection layer, referred to as the patch embedding layer. A classification (CLS) token is prepended to the sequence to facilitate downstream tasks. To preserve the 2D spatial structure of the spectrogram, a learnable positional embedding of size 768 is added to the patch embeddings, enabling the model to effectively capture temporal and spectral relationships within the input features. For training, the input features for each AST module were first normalized using the mean and standard deviation of the respective database, as recommended by the AST authors, to ensure consistency. After normalization, the features were passed through the Transformer encoder. To align with the model's final activation layer output, the target labels, originally in the range of 1-5, are scaled down to the range of 0-1. This scaling step is necessary to maintain compatibility with the sigmoid output of the model, ensuring effective learning during fine-tuning.

The AST model itself utilizes the Transformer encoder architecture described in [10], with an embedding dimension of 768, 12 layers, and 12 attention heads. The output of the CLS token from the Transformer encoder serves as the learned representation of the audio spectrogram. To map this representation to the target labels, a linear layer with a sigmoid activation function is employed. It leverages cross-modality transfer learning by initializing its weights using pretrained weights from a Vision Transformer (ViT) model originally designed for image-based tasks. The AST was further fine-tuned for audio classification on the Audioset dataset, enabling it to learn task-specific features for audio signal processing. This transfer learning approach enhances the model's ability to generalize effectively to audio spectrogram representations.

In our work, five AST modules were initialized with pretrained AST weights and subsequently fine-tuned on our speech quality datasets, enabling the model to adapt to the specific task requirements while leveraging the general feature representations learned during pretraining. The model was optimized using the Mean Squared Error (MSE) loss function, with the ADAM optimizer employed for efficient gradient updates. A learning rate of 1e-6 was used to ensure stable convergence during training.

Each AST module was trained for a task-specific number of epochs: MOS for 17 epochs, NOI for 11 epochs, DIS for 9 epochs, COL for 10 epochs, and LOUD for 8 epochs. Training was monitored and stopped early when the validation loss did not decrease for 20 epochs, to prevent overfitting and ensure optimal model generalization. Each of the five AST modules contains approximately 87M parameters. Model training was performed on an NVIDIA H100 SXM5 GPU for approximately one day.

3. Dataset

Altogether, 106 databases containing a total of 165,791 samples were used. The model was trained on 84 databases comprising 143,865 samples, while an additional 22 databases with 21,926 samples were used for model selection—representing approximately 87% and 13% of the total data, respectively. A significant portion of these datasets was collected as part of ITU-T Study Group 12's efforts to develop objective speech quality models. Some datasets were confidentially shared under signed non-disclosure agreements, while the publicly released NISQA Corpus [7] was also included in the training data.

Target labels, including overall quality (MOS) and perceptual dimensions, were derived from a combination of subjective and objective ratings. Objective ratings were obtained using ITU-T Rec. P.863 or ITU-T Rec. P.863.2 when reference signals were available, while the NISQA model was used for datasets without reference signals.

Additionally, 11 hidden databases were used for blind evaluation, conducted independently to ensure an unbiased assessment of the model's performance. These datasets, listed in Table 2, were curated by external industry partners and evaluation bodies to simulate real-world deployment scenarios. As the developer, we did not have access to these databases during training or fine-tuning, ensuring that the evaluation process remained completely blind. This approach allows for an objective assessment of the model's ability to generalize to previously unseen data and prevents any potential overfitting to known datasets. The results from these hidden databases provide critical insights into the robustness and practical applicability of the proposed model in real-world conditions.

4. Experiments and Results

Performance evaluation is conducted on both a validation set consisting of 22 databases and on separate, hidden test sets. For the validation set, we compute the Pearson correlation coefficient (r) per database and average the results across all 22 databases to ensure a balanced evaluation across diverse conditions. Table 1 reports the average r values per file from the

validation databases. We compare three models: the open-source NISQA model, a freshly-trained version of NISQA on our dataset (denoted NISQA*[1]), and our proposed model, SQ-AST. The results show that SQ-AST achieves higher correlations across all perceptual dimensions, indicating improved generalization and validation performance.

Table 1: *Average PCC per file across 22 validation databases for each perceptual dimension.*

Model	MOS	NOI	COL	DIS	LOUD
NISQA	0.865	0.867	0.727	0.701	0.527
NISQA*	0.785	0.850	0.689	0.771	0.606
SQ-AST	**0.901**	**0.905**	**0.854**	**0.830**	**0.749**

Hidden Dataset Evaluation

The model was evaluated on hidden databases listed in Table 2. According to ITU-T requirements [11], the overall performance for unknown databases is evaluated based on the average RMSE* across all experiments and dimensions, which must remain below 0.35. Additionally, the worst-case performance stipulates that no single RMSE* per experiment and dimension, including overall quality, should exceed 0.5.

RMSE* is an epsilon-insensitive RMSE, which accounts for the uncertainty of subjective scores by incorporating the confidence interval of individual MOS values. Unlike traditional RMSE, small deviations within an epsilon-wide band around the target value are not counted. This epsilon is defined as the 95% confidence interval of the subjective MOS value, ensuring that the uncertainty of MOS is reflected in the evaluation. RMSE* is computed in the same manner as RMSE but applied after a third-order polynomial mapping to the predicted scores. The procedure is explained in [12].

Tables 4–7 present the prediction performance for individual perceptual dimensions on DB4, DB8, and DB9—the only hidden test databases for which subjective dimension scores are available. For the remaining databases, which contain only overall quality (MOS) ratings, the corresponding performance results are summarized in Table 3. In our independent evaluation, we find that the model's overall performance remains well within the required threshold. However, the discontinuity model does not meet the worst-case requirement on DB4, where RMSE* exceeds the acceptance criteria by 0.1278 (see Table 5). While the model performs well across all dimensions generally, further investigation is warranted to identify the conditions and dimensions where improvements can be made.

[1]Trained for 11 epochs, this was the same duration of training as the proposed model, on the same dataset as SQ-AST.

Table 2: *Hidden Datasets Used for Evaluation*

ID	DB	Source
DB1	RS_P800_2018_FBliveConditions	[13]
DB2	RS_P800_2016_FBcodecs	[14]
DB3	RS_P800_2015_FBevs	[14]
DB4	RS_PAMD_SetA_SWB_SQ_2	R&S[2]
DB5	ETSI_DB01_MAN_Validation	[15]
DB6	ETSI_DB02_GER_Validation	[15]
DB7	ETSI_DB03_ENG_Validation	[15]
DB8	Orange_New2024_FR_PSAMD	[16]
DB9	Orange_FR_PAMD_2	[17]
DB10	Orange_FR_EVSvsOpus	[18]
DB11	Orange_FR_NewCoders	[19]

Table 3: *Per-condition Results for Overall Speech Quality*

DB	r	rmse(1st)	rmse*(3rd)
DB1	0.8237	0.5683	0.4697
DB2	0.9289	0.3913	0.2952
DB3	0.9413	0.3340	0.2210
DB5	0.9444	0.3713	0.2194
DB6	0.9441	0.3409	0.2232
DB7	0.9246	0.3374	0.2278
DB8	0.9390	0.0587	0.0552
DB9	0.7636	0.3040	0.3029
DB10	0.9634	0.1285	0.1252
DB11	0.9389	0.3421	0.3207
Average			**0.2460**

Table 4: *Per-condition Results for Noisiness*

DB	r	rmse(1st)	rmse*(3rd)
DB4	0.8836	0.5948	0.4601
DB8	0.8652	0.1720	0.1697
DB9	0.9460	0.1949	0.1944
Average			**0.2747**

Table 5: *Per-condition Results for Discontinuity*

DB	r	rmse(1st)	rmse*(3rd)
DB4	0.6244	0.7967	0.6278
DB8	0.8660	0.1909	0.1891
DB9	0.9177	0.1507	0.1505
Average			**0.3225**

5. Discussion

The results highlight potential dataset imbalances that may affect model learning, particularly for certain perceptual dimensions. Figure 2 illustrates the distribution of conditions and ratings across both the training and validation sets.

[2]Internal database kindly provided by Rohde & Schwarz.

Table 6: *Per-condition Results for Coloration*

DB	r	rmse(1st)	rmse*(3rd)
DB4	0.8847	0.4248	0.2909
DB8	0.9066	0.0836	0.0832
DB9	0.7612	0.2086	0.2082
Average			**0.1941**

Table 7: *Per-condition Results for Loudness*

DB	r	rmse(1st)	rmse*(3rd)
DB4	0.5120	0.3149	0.1593
DB8	0.6380	0.1732	0.1729
DB9	0.6473	0.2906	0.2898
Average			**0.2073**

For example, the loudness module was trained on a dataset in which the majority of samples exhibited optimal loudness, limiting exposure to a diverse range of conditions. Similarly, the data set contains relatively few instances with low discontinuity ratings, which can hinder the model's ability to generalize to highly degraded speech. These imbalances suggest that certain perceptual dimensions may require a more targeted dataset design to improve robustness.

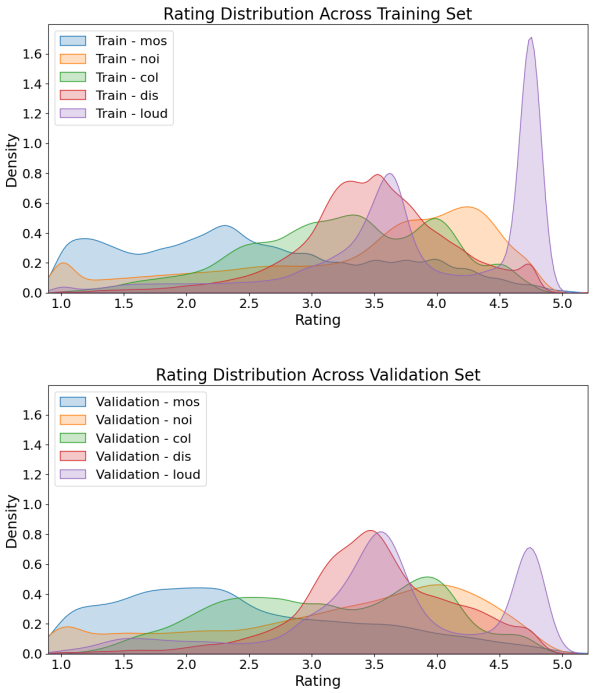


Figure 2: *Distributions of MOS and perceptual dimensions (NOI, COL, DIS, LOUD) across the training (top) and validation (bottom) datasets.*

6. Conclusion and Future Work

This work presents a Transformer-based approach for non-intrusive speech quality assessment, leveraging Audio Spectrogram Transformers to predict overall quality (MOS) and four perceptual dimensions—Noisiness (NOI), Discontinuity (DIS), Coloration (COL), and Loudness (LOUD). The model was trained and evaluated on a diverse set of databases, including both publicly available and confidential datasets. Experimental results demonstrate strong overall performance, meeting ITU-T requirements for unknown databases with one dimension and database to improve on. Analysis of dataset distributions revealed potential imbalances, particularly in the discontinuity and loudness dimensions, which may limit the model's ability to generalize to certain conditions.

To address these challenges, future work will focus on refining the training datasets to improve model robustness. Specifically, we will investigate challenging conditions for the discontinuity model, construct a new database targeting discontinuity variations, and retrain the model to enhance its performance in degraded speech conditions. Additionally, the loudness dataset will be rebalanced to ensure a more representative distribution of conditions, followed by retraining of the loudness model. Rather than training all perceptual dimension models on the same datasets, a more effective approach would be to curate dimension-specific datasets tailored to the characteristics and challenges of each perceptual task. This targeted dataset design would ensure a more balanced representation of conditions, ultimately enhancing the model's ability to generalize across real-world speech variations.

Furthermore, new hidden evaluation databases are being prepared to conduct a final assessment of the revised model, ensuring its effectiveness across a broader range of real-world speech scenarios. By improving dataset design and expanding evaluation strategies, we aim to enhance the model's generalization capabilities and contribute to the development of more reliable, standardized methodologies for perceptual speech quality assessment.

7. Acknowledgements

We thank Rohde & Schwarz SwissQual AG, HEAD Acoustics GmbH, and Orange Innovation for providing speech material and conducting independent validation tests that supported this research. This work was funded by the BIFOLD Agility Project and the Deutsche Forschungsgemeinschaft (DFG).

8. References

[1] ITU-T Recommendation P.800, *Methods for Subjective Determination of Transmission Quality*. Geneva: International Telecommunication Union, 1996.

[2] ITU-T Recommendation P.808, *Subjective Evaluation of Speech Quality with a Crowdsourcing Approach*. Geneva: International Telecommunication Union, 2018.

[3] B. Naderi and R. Cutler, "An Open Source Implementation of ITU-T Recommendation P.808 with Validation," in *Proceedings of Interspeech 2020*. ISCA, October 2020, pp. 2917–2921.

[4] ITU-T Recommendation P.806, *A Subjective Quality Test Methodology Using Multiple Rating Scales*. Geneva: International Telecommunication Union, 2014.

[5] ITU-T Recommendation P.863, *Perceptual Objective Listening Quality Assessment*. Geneva: International Telecommunication Union, 2018.

[6] ITU-T Recommendation P.863.2, *Extension of ITU-T P.863 for Multidimensional Assessment of Degradations in Telephony Speech Signals up to Full-Band*. Geneva: International Telecommunication Union, 2022.

[7] G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets," in *Proceedings of Interspeech*, 2021.

[8] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6493–6497.

[9] Y. Gong, Y. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proceedings of Interspeech 2021*. ISCA, 2021.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30. Curran Associates, Inc., 2017.

[11] ITU-T Contribution C384, *Proposal for the Requirement Specification of P.SAMD*. Geneva: International Telecommunication Union, 2019.

[12] ITU-T Recommendation P.1401, *Methods, Metrics, and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*. Geneva: International Telecommunication Union, 2020.

[13] ITU-T Contribution C286, *A Subjective ACR LOT testing Super-Wideband Speech Coding in Real Field Measurements and Prediction by P.863*. Geneva: International Telecommunication Union, 2018.

[14] ITU-T Contribution C281, *A Subjective ACR LOT testing EVS speech Coding and Prediction by P.863*. Geneva: International Telecommunication Union, 2016.

[15] ETSI, *Annex E of ETSI TS 103 558 V1.3.1: Speech and Multimedia Transmission Quality (STQ); Methods for Objective Assessment of Listening Effort*. European Telecommunications Standards Institute, July 2021, no. TS 103 558 V1.3.1.

[16] ITU-T Contribution C004, *New Subjective Database for the P.SAMD Study Item*. Geneva: International Telecommunication Union, 2025.

[17] ITU-T Contribution C565, *New unknown Subjective Database from Orange for the P.AMD and P.SAMD Projects and Validation of Candidate Models Against this Database*. Geneva: International Telecommunication Union, 2021.

[18] ITU-T Contribution C122, *Report on Investigations About the Relative Voice Quality of Telephony Services Using WB and SWB Modes of Modern Coders: Two New Subjective Tests and Predictions with FB Mode of P.863*. Geneva: International Telecommunication Union, 2023.

[19] ITU-T Contribution C222, *Comparison of Voice Quality Estimations Between Subjective and Objective Evaluations on Neural Audio Coders*. Geneva: International Telecommunication Union, 2024.