FEW-SHOT AND PSEUDO-LABEL GUIDED SPEECH QUALITY EVALUATION WITH LARGE LANGUAGE MODELS

Ryandhimas E. Zezario[†], *Dyah A. M. G. Wisnu*[†,*], *Szu-Wei Fu*[‡], *Sabato Marco Siniscalchi*[§] , *Hsin-Min Wang*[†], *Yu Tsao*[†]

[†]Academia Sinica, [*]National Chengchi University, [‡]NVIDIA, [§]University of Palermo

ABSTRACT

In this paper, we introduce GatherMOS, a novel framework that leverages large language models (LLM) as meta-evaluators to aggregate diverse signals into quality predictions. GatherMOS integrates lightweight acoustic descriptors with pseudo-labels from DNSMOS and VQScore, enabling the LLM to reason over heterogeneous inputs and infer perceptual mean opinion scores (MOS). We further explore both zero-shot and few-shot in-context learning setups, showing that zero-shot GatherMOS maintains stable performance across diverse conditions, while few-shot guidance yields large gains when support samples match the test conditions. Experiments on the VoiceBank-DEMAND dataset demonstrate that GatherMOS consistently outperforms DNSMOS, VQScore, naive score averaging, and even learning-based models such as CNN-BLSTM and MOS-SSL when trained under limited labeled-data conditions. These results highlight the potential of LLM-based aggregation as a practical strategy for non-intrusive speech quality evaluation.

***Index Terms*—** speech quality, generative model, llm, pseudo label, in context learning

1. INTRODUCTION

Speech quality has long been an important indicator for evaluating various speech processing applications, including speech enhancement [1], hearing aid (HA) devices [2], and telecommunications [3]. While human-based evaluation remains the gold standard, it requires a sufficient number of listeners to obtain reliable and generalized scores. With the increasing availability of large-scale audio datasets with corresponding human-labeled scores, deep learning models have attracted significant attention for deploying non-intrusive speech quality prediction systems [4–9]. More recently, large-scale pre-trained audio models such as wav2vec 2.0 [10] and Whisper [11] have emerged as powerful acoustic feature extractors, further boosting the performance of non-intrusive quality assessment models. Nevertheless, having a sufficient number of training samples remains crucial for achieving robust and reliable model performance.

Recently, large language models (LLM) such as GPT-4o and GPT-5 have demonstrated strong reasoning capabilities and the ability to integrate multimodal information. Several works have begun exploring their use in speech-related tasks, including pathological speech detection and zero-shot intelligibility assessment [12, 13]. For example, [13] converted audio into text-based representations and prompted ChatGPT to judge naturalness. While promising, such studies rely on direct prompting and lack explicit mechanisms to incorporate acoustic features or auxiliary predictors, which limits their robustness. Moreover, our preliminary experiments indicate that ChatGPT struggles with raw audio, requiring intermediate features as proxies. However, handcrafted features (e.g., root mean square (RMS), zero-crossing rate (ZCR), and mel-frequency cepstral coefficients (MFCC)) only provide coarse summaries of the signal and often correlate weakly with perceptual judgments. This motivates the need for a framework that can integrate richer acoustic descriptors with complementary predictors, allowing LLMs to make more reliable and fine-grained quality estimations.

Given these constraints, we aim to further investigate how LLM can be leveraged for speech quality estimation. In particular, we propose GatherMOS, a framework that acts as a meta-evaluator by aggregating diverse weak signals into stronger and more reliable predictions. GatherMOS combines acoustic feature representations with pseudo-labels obtained from DNSMOS [14] and VQScore [6], enabling the LLM to reason over structured inputs rather than relying solely on coarse audio information or text proxies. We specifically choose DNSMOS and VQScore for their lightweight design and independence from large-scale models, which enhances practicality. By leveraging LLM's reasoning capabilities, GatherMOS integrates information from both descriptors and pseudo-labels to refine MOS estimation. Additionally, we explore whether incorporating more detailed acoustic features, such as MFCC and spectrogram images, along with few-shot in-context information, can further improve performance. Experimental results confirm that GatherMOS outperforms DNSMOS, VQScore, and naive score averaging, as well as learning-based models trained on limited data. We also observe that mismatched few-shot samples can reduce predic-

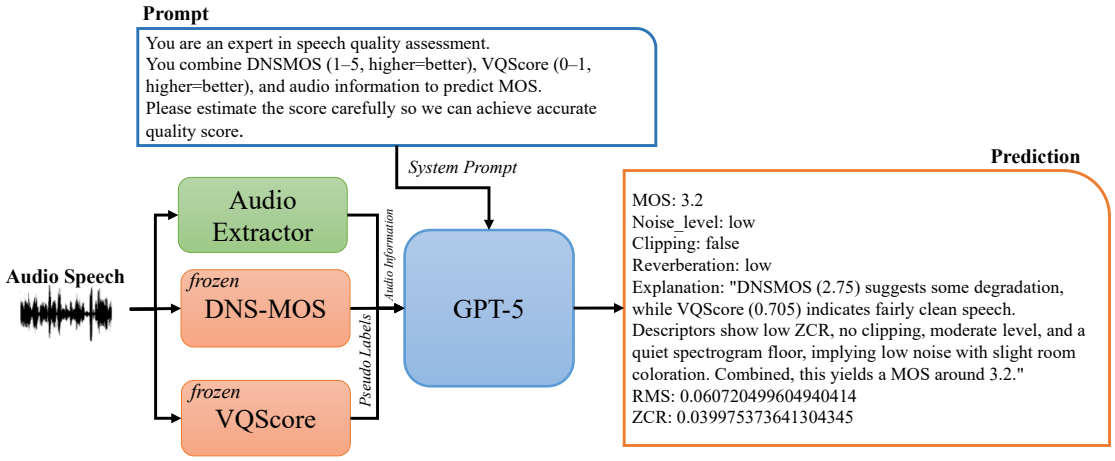


Fig. 1. Detailed architecture of GatherMOS

tion accuracy, whereas richer acoustic features consistently improve results. These findings suggest that GatherMOS has potential as an LLM-based meta-evaluator for non-intrusive speech quality, complementing existing methods and pointing toward more scalable evaluation with reduced dependence on annotated data.

The remainder of this paper is organized as follows. Section II presents the proposed GatherMOS method. Section III describes the experimental setup and results. Finally, Section IV provides conclusions and future work.

2. GATHERMOS

2.1. Zero-shot GatherMOS

Given an input waveform $x \in \mathbb{R}^T$, we extract several acoustic descriptors that summarize temporal, spectral, and perceptual information. The RMS reflects the overall signal energy, while the ZCR indicates the level of noisiness or voicing. Duration and clipping detection are included to capture temporal length and signal distortion when the maximum amplitude approaches unity. In addition, 13 MFCC are computed and averaged across frames to represent spectral envelope characteristics. Finally, mel-spectrogram statistics, including per-bin mean and variance as well as global maximum and minimum values, are extracted to capture spectral distribution patterns. In addition to handcrafted descriptors, we incorporate two lightweight predictors DNSMOS ($s_{\text{DNS}} \in [1, 5]$) and VQScore($s_{\text{VQ}} \in [0, 1]$). These pseudo-labels act as weak supervision signals that correlate with perceptual quality while remaining computationally efficient.

We aggregate descriptors and pseudo-labels using an LLM, specifically GPT-5. The inputs are serialized into a structured textual prompt, and the model is instructed to output a continuous MOS prediction $\hat{s}$ along with auxiliary explanatory attributes as shown in Fig. 1. Formally, GatherMOS is expressed as:

$$\hat{s}, a = f_{\text{LLM}}\big(\{d_i\}_{i=1}^{N},\ s_{\text{DNS}},\ s_{\text{VQ}}\big)\,, \tag{1}$$

where $\{d_i\}_{i=1}^{N}$ are extracted descriptors, s_{DNS} and s_{VQ} denote pseudo-labels, and a represents additional attributes (e.g., noise level, clipping, reverberation).

2.2. Few-Shot GatherMOS

When representative few-shot examples (x_j, s_j) are available, they are incorporated into the prompt to provide in-context guidance during inference. Specifically, these few-shot samples are provided as input to the LLM along with the other descriptors:

$$\hat{s}, a = f_{\text{LLM}}\big(\{d_i\}_{i=1}^{N},\ s_{\text{DNS}},\ s_{\text{VQ}},\ \{(x_j, s_j)\}_{j=1}^{K}\big)\,, \tag{2}$$

where $\{(x_j, s_j)\}_{j=1}^{K}$ denotes the few-shot support set consisting of K labeled examples. In this formulation, the support set is provided solely as part of the model input context and does not involve any update to model parameters. The LLM is therefore guided by both pseudo-labels and few-shot examples at inference time.

3. EXPERIMENTS

3.1. Experimental setup

The proposed approaches are evaluated on the VoiceBank-DEMAND dataset [15], which is also included in the test evaluation of the VoiceMOS Challenge 2024 [16]. We select a test set of 200 utterances, consisting of clean speech, noisy speech corrupted by four noise types (car, white, music, and cafeteria) at an SNR of 0 dB, and enhanced speech produced by five representative enhancement systems: BSSE [17], MPSENet [18], CMGAN [19], DEMUCS [20], and Wiener

Table 1. Evaluation scores of GatherMOS with a few sample test set.

Systems	LCC	SRCC
DNSMOS	0.5538	0.5231
VQScore	0.4631	0.6359
NaiveEnsemble	0.6255	0.5490
GatherMOS-ZS	0.6310	0.6420
GatherMOS-FS	**0.6653**	**0.8473**

filtering [1]. Human evaluation is conducted with ten listeners (six male and four female), where each utterance is rated by five listeners. For objective assessment, we report two correlation-based metrics: the linear correlation coefficient (LCC) and Spearman's rank correlation coefficient (SRCC) [21], where higher values indicate stronger alignment between predicted scores and ground-truth human ratings.

3.2. Few samples evaluation

In this experiment, we first use a small set of ten speech samples covering different acoustic environments and enhancement systems. Starting with a limited number of samples enables a controlled comparison with the baselines, facilitates detailed inspection of individual predictions, and allows us to verify the behavior of GatherMOS before proceeding to full-scale evaluation. For DNSMOS [14], we adopt the publicly available checkpoint provided on GitHub to obtain evaluation scores. Similarly, for VQScore [6], we use the official GitHub checkpoint with its recommended configuration. For NaiveEnsemble, we compute the average of the scores predicted by DNSMOS and VQScore. To ensure fairness, we normalize the outputs of each model so that both scores are on the same scale prior to averaging. For the proposed approach, we investigate two variants of GatherMOS. GatherMOS-ZS refers to the zero-shot setting that leverages basic acoustic features, including RMS, ZCR, clipping ratio, and duration. GatherMOS-FS corresponds to the few-shot setting, which incorporates the same acoustic features as GatherMOS-ZS together with a small support set of labeled examples. Specifically, three representative few-shot examples corresponding to low, medium, and high MOS levels are used. As illustrated in Fig. 2, these examples are selected to span different quality ranges. The few-shot examples are unseen from the test set to mimic a more realistic scenario, and they are selected from the UDASE task of the 7th CHiME Challenge [22, 23].

Table 1 summarizes the evaluation results in terms of SRCC and LCC. The baselines, DNSMOS and VQScore, show moderate correlation performance. NaiveEnsemble, which simply averages the outputs of the two baselines, provides a marginal improvement. In addition, GatherMOS-ZS outperforms all baselines, showing the benefit of incorporating acoustic features together with pseudo-labels into the LLM framework. Furthermore, using a small set of few-shot examples further improves prediction accuracy. Specifically, GatherMOS-FS increases the SRCC substantially, reaching 0.8473, which highlights the effectiveness of few-shot examples in guiding the LLM toward more accurate speech quality prediction.

Table 2. Evaluation scores of GatherMOS with all sample test set.

Systems	LCC	SRCC
DNSMOS	0.6021	0.5314
VQScore	0.5753	0.4476
NaiveEnsemble	0.6106	0.5177
CNN-BLSTM	0.3192	0.2971
MOS-SSL	0.4888	0.4732
GatherMOS-ZS	0.6439	0.6014
GatherMOS-ZS*	**0.6495**	**0.6069**
GatherMOS-FS	0.5653	0.4770

Here is the few shot information for your knowledge to adjust the score appropriately:

Example 1: Input: RMS=0.00691, ZCR=0.19525, Duration=4.43, Clipping=False, DNSMOS=1.20984, VQScore=0.6356 Output: MOS=1.125

Example 2:Input: RMS=0.01495, ZCR=0.1051, Duration=4.99, Clipping=False, DNSMOS=2.56121, VQScore=0.68099 Output: MOS=3.5

Example 3:Input: RMS=0.01715, ZCR=0.15325, Duration=4.6, Clipping=False, DNSMOS=2.53536, VQScore=0.68106 Output: MOS=4.375

Fig. 2. Few-shot prompt examples used for in-context guidance.

3.3. All samples evaluation

In this experiment, we evaluate GatherMOS and baseline methods on the full set of speech samples. We also introduce GatherMOS-ZS*, which incorporates additional MFCC and spectrogram features into the model. To ensure prediction stability, test samples are evaluated in minibatches of ten, with the LLM session reset between minibatches, which helps mitigate potential cross-sample conditioning effects in LLM inference. In addition, we build two trainable neural speech assessment baselines: a CNN-BLSTM-ATT model following the architecture in [4], and a MOS-SSL model following [7], which integrates a wav2vec module. For these two trainable models, we investigate whether it is feasible

to train them from scratch under very limited labeled-data. Specifically, we train CNN-BLSTM and MOS-SSL using the same three labeled samples as a stress-test to evaluate their feasibility under extreme low-resource conditions, rather than as a competitive fully supervised setting.

Table 2 reports the performance in terms of SRCC and LCC. DNSMOS and VQScore achieve moderate correlation, and NaiveEnsemble provides marginal improvements. CNN-BLSTM and MOS-SSL, trained with limited data, show weaker correlation compared to other methods. GatherMOS-ZS, which uses basic acoustic features (RMS, ZCR, clipping ratio, and duration), consistently outperforms all baselines. GatherMOS-ZS*, which further includes MFCC and spectrogram features, provides a slight improvement over GatherMOS-ZS, showing the contribution of richer acoustic representations. Next, GatherMOS-FS, which leverages a small few-shot support set, performs worse than the zero-shot variants on the full test set. As shown in Table 1, GatherMOS-FS achieves the best performance on the few-shot subset, indicating that the model can exploit few-shot examples effectively when the support set closely matches the test samples. The drop in performance on the full set suggests that the few-shot examples do not sufficiently represent the broader distribution, leading to domain bias and overfitting to the support examples. This highlights a tradeoff between in-domain few-shot adaptation and cross-condition generalization. In contrast, the zero-shot variants, especially GatherMOS-ZS*, generalize more robustly across diverse environments and enhancement systems.

3.4. Scatter plot analysis

In this experiment, we visualize scatter plots of predicted versus ground-truth MOS scores to analyze how each method performs on the full test set. For the first case, NaiveEnsemble produces predictions that are compressed into a narrow mid-range of the MOS scale, resulting in moderate correlation and difficulty in capturing the full variability of human quality scores. MOS-SSL also suffers from a limited prediction range, with outputs clustered between 2.5 and 3.5, which weakens its ability to represent the full MOS spectrum and leads to lower correlation values. In contrast, GatherMOS-FS spreads predictions more broadly across the MOS scale, but the alignment with the ground-truth scores is weaker, as reflected by increased scatter around the diagonal trend. This suggests that although GatherMOS-FS improves score coverage, it produces less consistent alignment with ground-truth scores across the full range. Interestingly, GatherMOS-ZS, which integrates a zero-shot strategy and combines several acoustic features (RMS, ZCR, clipping ratio, and duration) with pseudo-labels, achieves better correlation and predictions that align more closely with the diagonal trend. Finally, GatherMOS-ZS*, which additionally incorporates MFCC and spectrogram features, produces the best alignment with ground-truth scores. This highlights that enriching the feature set with complementary acoustic information enables the model to generalize more reliably and allows the LLM to better calibrate predictions to reflect subjective human judgments across the full MOS range.

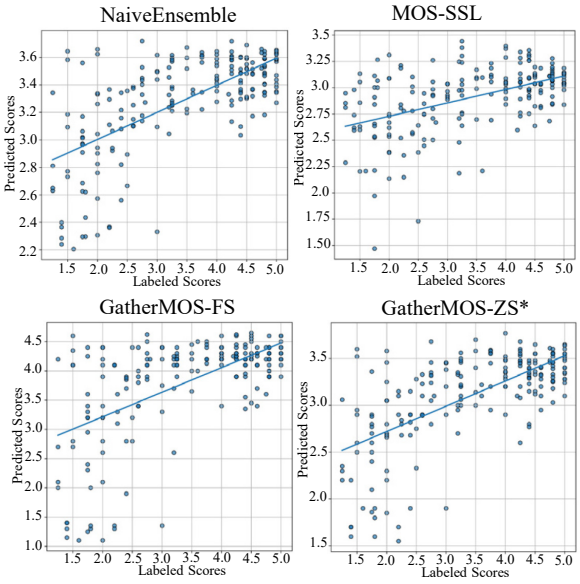


Fig. 3. Scatter Plot Analysis Between GatherMOS and Other Models.

4. CONCLUSION

In this work, we explored the use of generative AI for non-intrusive speech quality estimation and introduced GatherMOS, a framework that acts as a meta-evaluator by combining acoustic feature representations with pseudo-labels from DNSMOS and VQScore. By leveraging the reasoning capabilities of large language models, GatherMOS integrates these diverse signals to produce more reliable MOS predictions. Our experiments show that incorporating richer acoustic features, such as MFCC and spectrogram statistics, consistently enhances performance. We also observed that few-shot in-context examples can provide substantial gains when the support set aligns with the test domain, but may reduce accuracy when mismatched. Across evaluations on the VoiceBank-DEMAND dataset, GatherMOS outperforms DNSMOS, VQScore, naive ensembles, and learning-based baselines trained on limited data. These findings suggest that LLM-based aggregation offers a promising direction for scalable and practical speech quality evaluation, reducing reliance on large annotated datasets while complementing existing approaches. Future work will extend the study to larger and more diverse datasets, and further explore strategies to improve the stability of few-shot prompting and the meta-evaluator.

5. REFERENCES

[1] P. C. Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.

[2] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Quality Index (HASQI) Version 2," *Journal of the Audio Engineering Society*, vol. 62, no. 3, pp. 99–117, 2014.

[3] J. G. Beerends, C. Schmidmer, J. Berger, M. Obermann, R. Ullmann, J. Pomy, and M. Keyhl, "Perceptual objective listening quality assessment (POLQA), the third generation ITU-T standard for end-to-end speech quality measurement part i—temporal alignment," *Journal of The Audio Engineering Society*, vol. 61, no. 6, pp. 366–384, 2013.

[4] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[5] R. E. Zezario, S.-W Fu, F. Chen, C.-S Fuh, H.-M. Wan, and Yu Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2023.

[6] S.-W. Fu, K.-H. Hung, Y. Tsao, and Y.-C. F. Wang, "Self-supervised speech quality estimation and enhancement using only clean speech," in *Proc. ICLR*, 2024, pp. 1–22.

[7] E. Cooper, W.-H. Huang, T. Toda, and J. Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. ICASSP*, 2022, pp. 8442–8446.

[8] S. Wang, W. Yu, Y. Yang, C. Tang, Y. Li, J. Zhuang, X. Chen, X. Tian, J. Zhang, G. Sun, L. Lu, and C. Zhang, "Enabling auditory large language models for automatic speech quality evaluation," in *Proc. ICASSP*, 2025, pp. 1–5.

[9] C. Chen, Y. Hu, S. Wang, H. Wang, Z. Chen, C. Zhang, C.-H. H. Yang, and E. Chng, "Audio large language models can be descriptive speech quality evaluators," in *Proc. ICLR*, 2025, pp. 1–15.

[10] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NIPS*, 2020, pp. 12449 – 12460.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28492–28518.

[12] M. Amiri, H.O. Shahreza, and I. Kodrasi, "Exploring in-context learning capabilities of ChatGPT for pathological speech detection," *arXiv 2503.23873*, 2025.

[13] R. E. Zezario, S. M. Siniscalchi, H.-M. Wang, and Y. Tsao, "A study on zero-shot non-intrusive speech assessment using large language models," in *Proc. ICASSP*, 2025, pp. 1–5.

[14] C. K. A Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *Proc. ICASSP*, 2021, pp. 6493–6497.

[15] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust text-to-speech.," in *Proc. SSW*, 2016, pp. 146–152.

[16] W.-C. Huang, S.-W. Fu, E. Cooper, R. E. Zezario, T. Toda, H.-M. Wang, J. Yamagishi, and Y. Tsao, "The Voicemos challenge 2024: Beyond speech quality prediction," in *Proc. SLT*, 2024, pp. 803–810.

[17] K.-H. Hung, S.-W. Fu, H.-H. Tseng, H.-T. Chiang, Y. Tsao, and C.-W. Lin, "Boosting Self-Supervised Embeddings for Speech Enhancement," in *Proc. Interspeech*, 2022, pp. 186–190.

[18] Y.-X. Lu, Y. Ai, and Z.-H. Ling, "MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra," in *Proc. Interspeech*, 2023, pp. 3834–3838.

[19] R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. Interspeech*, 2022, pp. 936–940.

[20] A. Défossez, G. Synnaeve, and Y. Adi, "Real Time Speech Enhancement in the Waveform Domain," in *Proc. Interspeech*, 2020, pp. 3291–3295.

[21] C. Spearman, "The proof and measurement of association between two things," *The American Journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[22] S. Leglaive, L. Borne, E. Tzinis, M. Sadeghi, M. Fraticelli, S. Wisdom, M. Pariente, D. Pressnitzer, and J. R. Hershey, "The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement," in *Proc. CHIME*, 2023, pp. 1–7.

[23] S. Leglaive, M. Fraticelli, H. ElGhazaly, L. Borne, M. Sadeghi, S. Wisdom, M. Pariente, J. R Hershey, D. Pressnitzer, and J. P Barker, "Objective and subjective evaluation of speech enhancement methods in the UDASE task of the 7th CHiME challenge," *Computer Speech and Language*, vol. 89, pp. 101685, 2025.