COMFORMALMOS: UNCERTAINTY-AWARE MOS PREDICTION WITH CONFORMAL INTERVALS AND ORDINAL MODELING

Anonymous authors
Paper under double-blind review

ABSTRACT

Accurately predicting human Mean Opinion Scores (MOS) is essential for evaluating synthetic speech quality in text-to-speech (TTS) and voice conversion (VC) systems. Existing MOS prediction models focus on point estimates and often overlook uncertainty, reducing model selection and deployment reliability. Recent work has sought to address uncertainty estimation using probabilistic losses but lacks formal coverage guarantees. Addressing this limitation, we introduce ComformalMOS, a framework that augments MOS prediction with conformal prediction-based interval estimation to provide statistically valid prediction intervals with guaranteed coverage under exchangeability assumptions, alongside conventional point estimates. During training, ordinal-aware modeling of the MOS score converts one-hot labels into a soft distribution using a Gaussian kernel. By explicitly modeling the ordinal structure of MOS labels, our approach produces reliable uncertainty estimates when softmax-based confidence scores become overconfident on out-of-distribution speech, ensuring that the resulting intervals respect the ordering of MOS scores. We evaluate our method on both point-prediction quality and uncertainty quality. Experiments on BVCC datasets demonstrate that ComformalMOS maintains competitive point prediction performance (MSE = 0.08) while providing prediction intervals with empirically validated coverage rates. This dual capability enhances model reliability for deployment in production TTS and VC systems where uncertainty quantification is critical.

1 INTRODUCTION

Accurately predicting human Mean Opinion Scores (MOS) is essential for evaluating the quality of synthetic speech and music, yet the task remains challenging due to the inherent subjectivity of human ratings and the variability introduced by factors such as noise conditions, accents, and listener bias (Rosenberg & Ramabhadran, 2017; Wells et al., 2024). Large-scale benchmarking efforts report that even state-of-the-art systems achieve only moderate utterance-level correlation with human scores, underscoring the difficulty of reliable automatic MOS estimation (Liu et al., 2025; Huang et al., 2025). Recent models such as UTMOSv2 leverage transfer learning from spectrogram-based image classifiers fused with self-supervised speech embeddings, achieving strong predictive accuracy for speech naturalness (Baba et al., 2024). Other methods improve performance further through waveform spectrogram fusion or cross-modal encoders Hoq et al. (2025).

Despite these advances, existing systems share a key limitation since they output only point estimates of MOS without quantifying uncertainty. This makes them vulnerable to inaccurate predictions in open-world settings, where unseen conditions can degrade prediction reliability. As highlighted by Wang et al. (2024), the lack of uncertainty modeling hinders the safe deployment of MOS systems in real-world applications, including game development, movie production, and virtual assistants. While some probabilistic approaches attempt to capture uncertainty Wang et al. (2024); Hoq et al. (2025), they cannot guarantee that the predicted intervals reliably contain the true MOS, limiting their trustworthiness in deployment.

To address this challenge, we adopt conformal prediction, a distribution-free framework that produces prediction intervals while incorporating ordinal-aware modeling to capture the discrete and ranked nature of MOS ratings. Unlike heuristic uncertainty measures, these intervals provide formal coverage guarantees under the assumption that calibration and test data are exchangeable.

1.1 CONTRIBUTION

In this paper, we introduce ComformalMOS, a novel uncertainty-aware MOS prediction framework that extends our previous MOS prediction model (Elelu et al., a;b). The key contributions of our work are:

- **Conformal Uncertainty Intervals:** We integrate conformal prediction to output statistically valid prediction intervals for MOS. These intervals come with formal coverage guarantees (Shafer & Vovk, 2008), which greatly improves the reliability of uncertainty estimates compared to heuristic approaches.
- **Ordinal-Aware Modeling:** We explicitly respect the ordinal structure of MOS labels by converting one-hot score labels into Gaussian-smoothed probability distributions over discrete score levels (Ritter-Gutierrez et al., 2025), allowing the model to capture the partial correctness of neighboring scores. This ordinal regression approach aligns the training objective with the nature of human ratings and improves point estimation accuracy.
- **Improved Accuracy and Robustness:** On the BVCC dataset, ComformalMOS achieves competitive point-prediction accuracy compared to strong baselines (Baba et al., 2024; Hoq et al., 2025). Our m2d backbone model with $\alpha = 0.05$ attains a system-level MSE of 0.08, representing a 7.0% reduction from FUSE-MOS. Conformal intervals further provide calibrated uncertainty, effectively flagging low-confidence or out-of-distribution inputs and enhancing robustness in real-world scenarios.

2 RELATED WORK

Early MOS prediction methods relied on hand-crafted features or shallow neural networks to predict point estimates of human ratings. In the speech domain, non-intrusive metrics such as PESQ and Quality-Net were extended to MOS prediction, but they did not provide uncertainty estimates (Fu et al., 2018). With the rise of deep learning, end-to-end models improved system-level correlation with human MOS ratings (Lo et al., 2019; Baba et al., 2024), yet these approaches still exhibit high variance at the utterance level due to subjective rating noise and continue to produce only point estimates.

More recently, research driven by competition has advanced MOS prediction. The VoiceMOS Challenge 2024 and AudioMOS Challenge 2025 introduced tasks for speech and music quality prediction, attracting numerous submissions. Hoq et al. (2025) proposed FUSE-MOS, a network that fuses raw waveform and log-Mel features to predict both a MOS point estimate and its posterior distribution (Hoq et al., 2025). In the VoiceMOS 2024 challenge, Baba et al. showed that their top-ranked T05 system leveraged a pretrained image encoder on spectrograms alongside a self-supervised speech encoder, fusing these features to improve MOS naturalness prediction (Baba et al., 2024).

In the music domain, the ASTAR-NTU team's winning entry for AudioMOS 2025 employed pretrained MuQ audio and RoBERTa text encoders with cross-attention, and crucially handled the ordinal nature of MOS by converting one-hot labels into soft distributions via a Gaussian kernel (Ritter-Gutierrez et al., 2025). Respecting label ordinality has been shown to improve predictive accuracy. Building on this line of inquiry, Elelu et al. introduced ComformalMOS, a self-supervised alternative to CLAP for music evaluation, demonstrating that the choice of backbone strongly impacts MOS prediction performance (**?**).

While these methods enhance point-prediction accuracy, most overlook prediction reliability and uncertainty estimation. This omission is critical because MOS labels are inherently subjective and noisy, leading to high variance. A few recent works have begun addressing this gap, such as Wang et al. (2024), who explicitly modeled aleatoric and epistemic uncertainties, enabling selective prediction and out-of-domain detection (Wang et al., 2024). However, such probabilistic approaches lack

formal coverage guarantees compared to conformal prediction (Shafer & Vovk, 2008). Although conformal methods have been applied in other domains to produce calibrated uncertainty intervals, to our knowledge, they have not yet been applied to MOS prediction. Our work integrates conformal prediction with ordinal modeling for MOS prediction.

3 METHODOLOGY

The proposed framework estimates speech quality Mean Opinion Scores (MOS) by combining ordinal regression with uncertainty quantification through split conformal calibration. First, pretrained audio representations are passed to an ordinal-aware prediction head, which outputs a probability distribution over evenly spaced MOS bins. These probabilities are then combined to produce a point estimate of perceptual quality. During training, the prediction head is optimized using Gaussian-smoothed ordinal targets, ensuring that the model respects the ordinal structure of MOS labels. In a separate calibration stage, split conformal prediction computes a single scalar half-width from calibration residuals, yielding finite-sample valid prediction intervals around each point estimate after clipping to the MOS range. Figure 1 illustrates this framework and the flow from input representations to final prediction intervals.

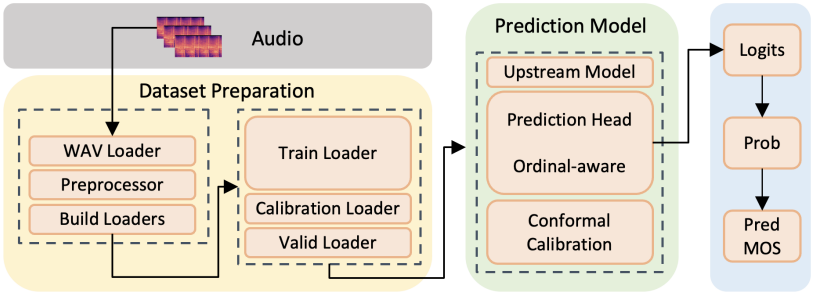


Figure 1: Overview of the proposed ComformalMOS framework. Input audio representations are first processed by a frozen upstream feature extractor. These features are fed into an ordinal prediction head, which uses Gaussian label smoothing to produce point MOS estimates. Finally, post-hoc conformal calibration generates statistically valid prediction intervals, providing reliability-aware predictions.

3.1 UPSTREAM MODEL EXTRACTION

The upstream model serves as a feature extractor that transforms an input waveform x into a high-level representation suitable for MOS prediction Niizumi et al. (2025). We denote this component as ϕ. Its purpose is to capture rich acoustic, prosodic, and phonetic information from the raw waveform, enabling the downstream prediction head to work with compact, fixed-dimensional embeddings instead of raw time-domain signals. This reduces data requirements and simplifies learning.

Formally, given an input waveform x, the upstream encoder produces

$$z = \phi(x) \in \mathbb{R}^D, \tag{1}$$

where z is a D-dimensional embedding vector.

Many pretrained speech models like M2D or Wav2Vec2, output a sequence of frame-level embeddings

$$Z = (h_1, \ldots, h_T) \in \mathbb{R}^{T \times D}. \tag{2}$$

In such cases, we apply temporal mean pooling to reduce the sequence to a single fixed-length representation:

$$\bar{h} = \frac{1}{T} \sum_{t=1}^{T} h_t \ \in \ \mathbb{R}^D. \tag{3}$$

To preserve the general-purpose speech representations learned during pretraining, the upstream encoder is kept frozen during training. Only the downstream prediction head is trained on the MOS-labeled data.

3.2 PREDICTION HEAD

The prediction head g_θ maps the upstream feature vector $\bar{h}$ into three outputs: (i) logits over K ordinal bins, (ii) the corresponding class probabilities, and (iii) a scalar MOS estimate Cao et al. (2020). We implement g_θ as a lightweight two-layer MLP with LayerNorm and dropout regularization.

MOS values lie on the continuous interval $[1, 5]$. A naïve approach would train directly on scalar MOS targets using an ℓ_1 loss. However, minimizing mean absolute error does not directly optimize rank-based metrics such as Spearman's rank correlation coefficient (SRCC), which measure the monotonic ordering of predictions rather than their absolute accuracy Wang et al. (2025). The model can achieve a high SRCC even if its predictions are uniformly shifted, as long as it correctly ranks clips from better to worse. To better align training with ranking behavior, we discretize the MOS space into K equal-width bins spanning [1,5], following the choice of K recommended by Ritter-Gutierrez et al. (2025):

$$c_k = 1 + \frac{4(k-1)}{K-1}, \qquad k = 1, \ldots, K. \tag{4}$$

Training the model to classify a sample into the correct bin encourages it to place samples in the right relative order, which supports stronger SRCC performance.

The standard cross-entropy loss with one-hot bin labels ignores the ordinal structure of MOS, therefore misclassifying a sample by one bin ($N \to N+1$) is penalized the same as a much larger error ($N \to N+10$). To address this, we replace hard one-hot targets with Gaussian-smoothed soft labels that distribute probability mass to neighboring bins according to their proximity to the true score. Formally, for a target MOS $y \in [1, 5]$,

$$\tilde{q}_k(y) = \exp\Big(- \frac{(y - c_k)^2}{2\sigma^2} \Big), \tag{5}$$

Following the approach of (Ritter-Gutierrez et al., 2025), we set σ slightly larger than the bin width to smooth the target distribution, enforcing local continuity while preserving ordinal relationships. This Gaussian label softening explicitly encodes the ordinal relationships among the 20 bins, ensuring that predictions close to the true MOS are penalized less than larger deviations, while preserving the rank structure of the scores.

The model outputs a scalar MOS prediction as the expected value of the bin centers:

$$\hat{y}(x) = \sum_{k=1}^{K} c_k \, p_k(x), \tag{6}$$

which preserves ordinal geometry and ensures $\hat{y}(x) \in [1, 5]$. We train the model using a combination of KL divergence between the predicted probabilities and the Gaussian-smoothed ordinal targets, along with an auxiliary ℓ_1 loss on the scalar MOS prediction to stabilize training.

3.3 CONFORMAL CALIBRATION

Although the prediction head produces a single MOS estimate $\hat{y}(x)$, point predictions alone do not indicate the model's level of confidence. In practical evaluation settings, reliable uncertainty intervals are crucial for assessing whether a predicted score can be trusted. To address this, we adopt *split conformal prediction*, a simple yet powerful post-hoc calibration method that provides finite-sample coverage guarantees without making distributional assumptions Lei et al. (2013); Dhillon et al. (2024).

We partition the labeled data into a training set, used to learn the prediction head parameters θ, and a separate calibration set, which comprises approximately 10% of the labeled samples. After training, we compute the absolute prediction errors (residuals) on the calibration set and use their empirical

distribution to determine a quantile corresponding to the desired coverage level (90%). This quantile becomes the half-width $\hat{q}$ of our prediction intervals.

For a new input x, we output an uncertainty interval

$$I(x) = [\,\hat{y}(x) - \hat{q},\ \hat{y}(x) + \hat{q}\,] \cap [1, 5], \tag{7}$$

clipped to the valid MOS range. By construction, these intervals satisfy $\mathbb{P}\{Y \in I(X)\} \geq 1 - \alpha$, meaning that the true MOS lies inside the interval with at least $(1 - \alpha)$ probability under the assumption that calibration and test data are drawn from the same underlying distribution.

The resulting intervals adapt to model accuracy by widening when calibration residuals are large (indicating high uncertainty or potential distribution shift) and narrowing when the model is confident and accurate. On average, the interval width is approximately $2\hat{q}$, offering an intuitive measure of prediction reliability.

Since conformal calibration is applied post-hoc, it does not require retraining the model and incurs minimal computational overhead. Computing the calibration residuals and determining the quantile scale linearly with the number of calibration samples makes this approach efficient even for large datasets. At inference time, generating prediction intervals only requires a simple addition and subtraction of the precomputed half-width q, which adds minimal cost compared to the forward pass of the prediction head.

4 Experimental Setup

4.1 Dataset Preparation

In our experiments, we utilize the BVCC Huang et al. (2022) for training, which comprises 7,106 English audio samples collected from 187 TTS and Voice Conversion (VC) systems included in previous challenges. Although the dataset contains a modest number of samples, it is widely used in speech MOS prediction research and provides diverse samples from multiple TTS and VC systems. Each audio clip is annotated by multiple human raters, ensuring high-quality labels, and such dataset sizes are standard in the field due to the cost and effort of collecting human MOS annotations. For model training and evaluation, we split the BVCC dataset into a training set of 2,029 samples, a validation set of 387 samples, and a calibration subset of 225 samples reserved from the training data for conformal interval estimation. This separation ensures that model fitting, calibration, and validation are performed on distinct subsets, supporting robust training and reliable uncertainty quantification.

4.2 Training Details

We trained our MOS prediction models on the BVCC dataset using NVIDIA RTX A5000 GPUs (24 GB memory). Each run was trained for up to 1000 epochs with early stopping monitored on validation loss, using a patience of 20 epochs. In practice, most runs stopped well before the maximum, typically after 300–400 epochs, once the validation loss plateaued, preventing overfitting and saving computation. This strategy ensures that the model achieves optimal generalization without unnecessary training. Training a full run generally requires about 2 hours. The batch size was set to 32 for training to balance stable gradient updates with GPU memory constraints. For validation and testing, we used a smaller batch size of 8 to reduce memory usage during inference Masters & Luschi (2018). We used stochastic gradient descent (SGD) with momentum because its inherent noise can prevent memorization and improve generalization on small to medium-sized datasets, while providing stable convergence for our ordinal-smoothed MOS prediction task. The optimizer was configured with a learning rate of 1×10^{-4}, momentum of 0.9, and weight decay of 1×10^{-4} to further regularize the model and prevent overfitting. To preserve the quality of pretrained representations, the upstream encoder was kept frozen while only the downstream prediction head was updated. Model checkpoints were saved based on the lowest validation loss achieved.

4.3 Evaluation Metrics

To assess point-prediction quality, we report the following standard evaluation metrics for MOS at both the utterance and system levels:

- **Mean Squared Error (MSE):** Measures the average squared difference between predicted and ground-truth MOS scores.
- **Linear Correlation Coefficient (LCC):** Quantifies the linear relationship between predictions and true scores.
- **Spearman's Rank Correlation Coefficient (SRCC):** Evaluates the monotonic ordering of predictions with respect to ground-truth MOS.
- **Kendall Tau Rank Correlation Coefficient (KTAU):** Measures the ordinal association between predicted and true rankings.

To assess the quality of our uncertainty estimates, we evaluate the conformal prediction intervals using the following metrics:

- **Coverage:** Measures the proportion of test samples for which the ground-truth MOS falls within the predicted interval $I(x)$, indicating how well the intervals capture the true values:

$$\text{Coverage} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{y_i \in I(x_i)\}, \tag{8}$$

  where N is the number of test samples and $\mathbf{1}\{\cdot\}$ is the indicator function. Coverage is compared to the nominal level $1-\alpha$.
- **Calibration Error.** Measures the absolute deviation between empirical coverage and the nominal coverage, reflecting how accurately the intervals are calibrated:

$$\text{Calibration Error} = \big|\text{Coverage} - (1-\alpha)\big|. \tag{9}$$

- **Average Interval Width.** Represents the mean width of the conformal intervals across all samples, capturing the overall sharpness of the uncertainty estimates:

$$\text{Average Width} = \frac{1}{N}\sum_{i=1}^{N} \big(\overline{y}_i - \underline{y}_i\big), \tag{10}$$

  where $I(x_i) = [\underline{y}_i, \overline{y}_i]$.
- **Sharpness (RMS Radius).** Measures a model's overall uncertainty level on a dataset through the root-mean-square of the half-widths of the intervals, providing an alternative concentration measure:

$$\text{Sharpness (RMS)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\overline{y}_i - \underline{y}_i}{2}\right)^2}. \tag{11}$$

5 Experimental Results

5.1 Evaluating ComformalMOS on BVCC

We assess the performance of our proposed ComformalMOS framework on the BVCC dataset at both the utterance and system levels, and compare it with prior state-of-the-art models (UTMOS and FUSE-MOS). Table 1 summarizes the results. We first observe that UTMOS and FUSE-MOS provide strong baselines, achieving low MSE and high correlation metrics. While UTMOS attains slightly better utterance-level accuracy, FUSE-MOS performs best in system-level ranking. By incorporating the upstream model M2D2, conformal prediction, and ordinal-aware modeling into the MOS estimation pipeline, our ComformalMOS models achieve competitive or superior system-level performance. In particular, the ComformalMOS m2d models consistently reach the highest system-level LCC while maintaining the lowest MSE.

We evaluated multiple α values 0.01, 0.05, 0.1, and 0.2 to examine their impact on conformal prediction intervals for MOS estimation. The choice of α directly affects the width of the intervals and the associated coverage probability. Smaller values produce narrower intervals but risk undercoverage, while larger α values yield wider intervals with higher coverage. Selecting multiple levels

allows us to study the trade-off between interval tightness and reliability, aiming for intervals that are both informative and statistically valid. Empirical studies have shown that appropriate selection of α is essential for controlling error rates and ensuring valid coverage in conformal prediction (Chernozhukov et al., 2018). Based on these considerations, we selected an α of 0.05 as our primary setting, corresponding to 95% nominal coverage, which balances interval width with reliability while maintaining the best overall performance.

This shows that the novel pipeline improves the reliability of predictions without degrading point-estimation accuracy. At the utterance level, ComformalMOS m2d maintains strong correlations close to UTMOS and FUSE-MOS, with only a slight increase in MSE.

Table 1: Comparison of FUSE-MOS, prior methods, and Conformal-MOS (Ours) on BVCC datasets. Best values in each column are in **boldface**.

Model	Utterance				System			
	MSE↓	LCC↑	SRCC↑	KTAU↑	MSE↓	LCC↑	SRCC↑	KTAU↑
UTMOS	**0.165**	**0.899**	**0.896**	-	0.090	0.936	0.936	-
FUSE-MOS	0.191	0.887	0.887	-	0.086	0.946	**0.947**	-
Our Models:								
m2d ($\alpha = 0.01$)	0.229	0.877	0.875	0.697	0.082	0.952	0.941	0.802
m2d ($\alpha = 0.05$)	0.228	0.878	0.876	**0.698**	**0.080**	**0.953**	0.943	**0.805**
m2d ($\alpha = 0.1$)	0.228	0.877	0.875	0.697	0.081	0.952	0.941	0.802
m2d ($\alpha = 0.2$)	0.229	0.877	0.876	0.697	0.083	0.951	0.941	0.801
wav2vec ($\alpha = 0.01$)	0.329	0.815	0.808	0.621	0.153	0.909	0.905	0.737
wav2vec ($\alpha = 0.05$)	0.334	0.814	0.808	0.620	0.162	0.908	0.906	0.739
wav2vec ($\alpha = 0.1$)	0.335	0.813	0.807	0.619	0.164	0.906	0.905	0.735
wav2vec ($\alpha = 0.2$)	0.348	0.808	0.800	0.612	0.181	0.898	0.904	0.736

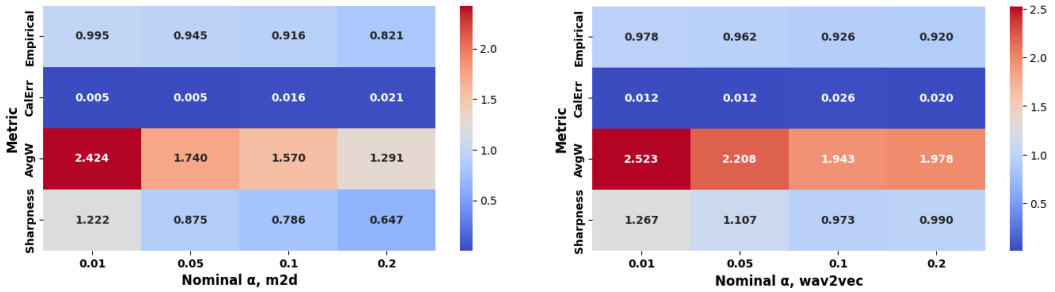
In contrast, the ComformalMOS wav2vec models underperform notably across all metrics, indicating that the choice of backbone strongly affects performance under the conformal framework. Overall, these results highlight that conformal and ordinal prediction enhance system-level robustness, especially when combined with the m2d backbone. Our best-performing ComformalMOS m2d $\alpha = 0.05$ configuration surpasses all baselines at the system level, while retaining competitive utterance-level accuracy. This confirms that conformal prediction can be effectively integrated into MOS estimation pipelines to improve prediction reliability without compromising prediction quality.

5.2 Backbone Effects on Conformal-MOS Calibration

Figures 2a and 2b present the calibration performance of our Conformal-MOS models on the BVCC dataset. Ideally, empirical coverage should closely match the nominal coverage, calibration error should be small, and intervals should be as narrow (sharp) as possible while remaining well-calibrated.

The m2d models exhibit consistently accurate coverage. Empirical coverage aligns closely with nominal targets at most α levels, with low calibration error, especially for the m2d backbone. As expected, increasing α leads to narrower prediction intervals and improved sharpness. This demonstrates the expected coverage-sharpness trade-off and confirms that the conformal calibration procedure is functioning properly on this backbone.

In contrast, the wav2vec models show higher calibration error and less precise coverage, particularly at lower α values. During the α of 0.01, the empirical coverage falls short of the target 0.99, and the intervals are much wider with higher sharpness values. This indicates that the weaker wav2vec backbone produces less reliable uncertainty estimates under the conformal framework.



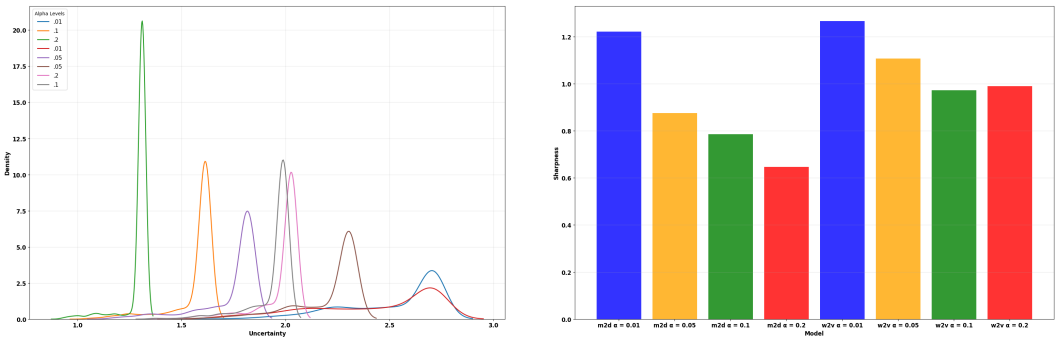
(a) M2D2 Result (b) Wav2Vec Result

Figure 2: Calibration performance of Conformal-MOS on the BVCC dataset using two different backbones. (a) M2D2 achieves empirical coverage closely matching nominal targets across multiple α levels, with low calibration error and relatively narrow (sharp) intervals, indicating reliable uncertainty estimates. (b) Wav2Vec shows larger calibration error, empirical coverage that falls short at low α levels, and wider intervals, highlighting less reliable uncertainty estimation.

Overall, these results confirm that conformal prediction produces well-calibrated uncertainty intervals when combined with a strong backbone like m2d, while performance degrades when the underlying point predictions are less accurate. This highlights the importance of backbone choice for reliable uncertainty estimation.

5.3 Qualitative Analysis

Figure 3a illustrates the density distribution of uncertainty values across multiple prediction models. The density peaks highlight the concentration of uncertainty values, showing variations in prediction interval widths across different alpha levels and backbones. Notably, some models exhibit narrower uncertainty distributions, indicating higher confidence in predictions, while others display broader distributions, suggesting lower certainty.



(a) Seaborn Plot (b) Bar Chart

Figure 3: Visualization of uncertainty and sharpness across different Conformal-MOS models and backbones. (a) The Seaborn density plot shows the distribution of prediction interval widths (uncertainty) for each model and alpha level. Peaks indicate where most intervals lie, with narrower distributions reflecting higher confidence and broader distributions indicating lower certainty. (b) The bar chart compares sharpness values across models and alpha levels. Lower sharpness corresponds to narrower, more precise intervals, while higher sharpness indicates wider, less confident predictions.

Figure 3b compares the sharpness values across different models and configurations. Sharpness, measured as the root-mean-square of prediction interval widths, reflects the precision of the models. Models with lower sharpness values demonstrate narrower intervals, indicating higher confidence. Conversely, models with higher sharpness values suggest wider intervals and lower precision. We

notice a decrease in sharpness as the alpha value is increased, with a 0.2 alpha level providing the best sharpness for m2d.

5.4 Discussion

Our proposed ComformalMOS framework provides more than a single-point estimate of perceived speech quality by producing prediction intervals with formal statistical guarantees. This shift is particularly important because MOS prediction is inherently subjective and therefore, different listeners may assign different scores to the same utterance, and even the same listener may vary depending on context, fatigue, or cultural background. As a result, human ratings exhibit natural variability that cannot be captured by point estimates alone. Traditional regression-based approaches ignore this subjectivity, offering a single predicted MOS without reflecting how confident the model is in that value. By contrast, ComformalMOS transforms MOS evaluation into a reliability-aware process, where both the estimated quality and its associated uncertainty are explicitly quantified. This not only aligns better with the noisy and subjective nature of MOS but also provides practitioners with actionable confidence measures for decision-making.

The interval provides a range of plausible MOS values for a given utterance with a formal coverage guarantee. For example, a 90% interval [2.8, 3.0] means that across the test distribution, at least 90% of true MOS values will fall within predicted ranges. On the 1–5 MOS scale, widths below 0.5 indicate reliable predictions, widths of 0.5–1.0 reflect moderate confidence, and widths above 1.0 highlight substantial uncertainty. Out-of-interval samples often correspond to hard-to-predict cases, such as atypical noise, accented speech, or distributional shifts relative to the calibration set. In deployment, narrow intervals around high MOS indicate reliable quality, while wide intervals signal uncertainty and may warrant further testing or human evaluation. In risk-sensitive settings, such as call center monitoring or hearing aid assessment, intervals enable threshold-based decisions: if the lower bound exceeds an acceptable MOS, the system is trusted; otherwise, deployment can be deferred. Conformal intervals, therefore, support reliability-aware decision-making and risk management.

We further note that treating MOS as a continuous variable in regression may produce unrealistic predictions and poorly calibrated intervals. In contrast, ordinal methods align with the inherent ranked nature of MOS ratings, producing probability distributions over discrete scores. When combined with conformal prediction, ordinal models typically yield narrower, better-calibrated intervals that are more consistent across the rating scale, especially near the extremes. This alignment with human rating behavior improves both interpretability and coverage stability.

6 Conclusion

Our experimental findings provide several key insights into the effectiveness of integrating M2D2, conformal prediction and ordinal-aware modeling into MOS estimation pipelines. First, the results on the BVCC dataset demonstrate that the ComformalMOS framework can enhance system-level performance without sacrificing utterance-level accuracy. This confirms that the framework can improve prediction reliability while preserving point-estimation quality.

Second, the backbone analysis highlights the critical role of base model quality in enabling effective calibration. While the m2d backbone produced well-calibrated uncertainty estimates with empirical coverage closely matching nominal targets and low calibration error, the wav2vec backbone struggled, showing higher calibration errors and wider intervals. This suggests that calibration amplifies the strengths and weaknesses of the underlying backbone and is most beneficial when combined with strong base predictors.

Third, our qualitative analyses of uncertainty distributions and sharpness further illustrate the expected coverage–sharpness trade-off. As α increases, prediction intervals become narrower while coverage gradually decreases. Models with lower sharpness values produce tighter and more confident predictions, while broader distributions reflect greater uncertainty.

Overall, these results show that ComformalMOS can be a powerful addition to MOS prediction systems, improving reliability and interpretability of predictions, particularly when applied to high-performing backbones.

REFERENCES

Kosuke Baba, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech. In *Proc. IEEE Spoken Language Technology Workshop (SLT)*, 2024.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Victor Chernozhukov, Kaspar Wüthrich, and Zhu Yinchu. Exact and robust conformal inference methods for predictive machine learning with dependent data. In *Conference On learning theory*, pp. 732–749. PMLR, 2018.

Guneet S Dhillon, George Deligiannidis, and Tom Rainforth. On the expected size of conformal prediction sets. In *International Conference on Artificial Intelligence and Statistics*, pp. 1549–1557. PMLR, 2024.

Kehinde Elelu, Josh Siegel, Ali Saffary, Hung Luong, Simeon Babatunde, and Ebuka Okpala. Convm2d2: Improving generative music evaluation using self-supervised alternative to clap. *Authorea Preprints*, a.

Kehinde Abdulsalam Elelu, Joshua E Siegel, Saffary Ali, Babatunde Simeon, Ebuka Okpala, et al. Convm2d2: Improving generative music evaluation using self-supervised alternative to clap. In *TTIC Summer Workshop on Foundations of Speech and Audio Foundation Models 2025*, b.

Szu-Wei Fu, Yu Tsao, Hsin-Te Hwang, and Hsin-Min Wang. Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm. In *Proc. Interspeech*, pp. 1873–1877, 2018.

Emon Hoq, Neha Gupta, Dennis Omondi, and Ifeoma Nwogu. Fuse-mos: Fusion of speech embeddings for mos prediction with uncertainty quantification. In *Proc. Interspeech*, pp. 2350–2354, 2025.

Wen-Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi. The voicemos challenge 2022. *arXiv preprint arXiv:2203.11389*, 2022.

Wen-Chin Huang, Hui Wang, Cheng Liu, Yi-Chiao Wu, Andros Tjandra, Wei-Ning Hsu, Erica Cooper, Yong Qin, and Tomoki Toda. The audiomos challenge 2025. *arXiv preprint arXiv:2509.01336*, 2025.

Jing Lei, James Robins, and Larry Wasserman. Distribution-free prediction sets. *Journal of the American Statistical Association*, 108(501):278–287, 2013.

Cheng Liu, Hui Wang, Jinghua Zhao, Shiwan Zhao, Hui Bu, Xin Xu, Jiaming Zhou, Haoqin Sun, and Yong Qin. Musiceval: A generative music corpus with expert ratings for automatic text-to-music evaluation. *arXiv preprint arXiv:2501.10811*, 2025. URL https://arxiv.org/abs/2501.10811.

Chi-Chun Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. Mosnet: Deep learning based objective assessment for voice conversion. In *Proc. Interspeech*, pp. 1541–1545, 2019.

Dominic Masters and Carlo Luschi. Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*, 2018.

Daisuke Niizumi, Daiki Takeuchi, Masahiro Yasuda, Binh Thien Nguyen, Yasunori Ohishi, and Noboru Harada. M2d2: Exploring general-purpose audio-language representations beyond clap. *arXiv preprint arXiv:2503.22104*, 2025. URL https://arxiv.org/abs/2503.22104.

Fernando Ritter-Gutierrez, Yu-Cheng Lin, Jian-Cheng Wei, Jia Hao Marcus Wong, Nancy F. Chen, and Hung-yi Lee. Astar-ntu solution to audiomos challenge 2025 track1. *arXiv preprint arXiv:2507.09904*, 2025. submitted to ASRU 2025.

Andrew Rosenberg and Bhuvana Ramabhadran. Bias and statistical significance in evaluating speech synthesis with mean opinion scores. In *Interspeech*, pp. 3976–3980, 2017.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Chien-Chun Wang, Kuan-Tang Huang, Cheng-Yeh Yang, Hung-Shin Lee, Hsin-Min Wang, and Berlin Chen. Qamro: Quality-aware adaptive margin ranking optimization for human-aligned assessment of audio generation systems. *arXiv preprint arXiv:2508.08957*, 2025.

Hao Wang, Shuyuan Zhao, Jian Zhou, Xin Zheng, Hao Sun, Xuan Wang, and Yong Qin. Uncertainty-aware mean opinion score prediction. In *Proc. Interspeech*, pp. 1215–1219, 2024.

Dan Wells, Andrea Lorena Aldana Blanco, Cassia Valentini-Botinhao, Erica Cooper, Aidan Pine, Junichi Yamagishi, and Korin Richmond. Experimental evaluation of mos, ab and bws listening test designs. *INTERSPEECH 2024: Speech and Beyond*, 2024.