



Uncertainty-Aware Mean Opinion Score Prediction

Hui Wang¹, Shiwang Zhao¹, Jiaming Zhou¹, Xiguang Zheng², Haoqin Sun¹, Xuechen Wang¹,
Yong Qin^{1,*}

¹Nankai University, China

²University of Wollongong, Australia

qinyong@nankai.edu.cn

Abstract

Mean Opinion Score (MOS) prediction has made significant progress in specific domains. However, the unstable performance of MOS prediction models across diverse samples presents ongoing challenges in the practical application of these systems. In this paper, we point out that the absence of uncertainty modeling is a significant limitation hindering MOS prediction systems from applying to the real and open world. We analyze the sources of uncertainty in the MOS prediction task and propose to establish an uncertainty-aware MOS prediction system that models aleatory uncertainty and epistemic uncertainty by heteroscedastic regression and Monte Carlo dropout separately. The experimental results show that the system captures uncertainty well and is capable of performing selective prediction and out-of-domain detection. Such capabilities significantly enhance the practical utility of MOS systems in diverse real and open-world environments.

Index Terms: MOS prediction, speech quality assessment, uncertainty estimation

1. Introduction

With the advancement of deep learning, particularly self-supervised learning (SSL), MOS prediction systems have made notable progress in specific domains [1, 2, 3, 4, 5]. For instance, Saeki et al. [6] propose an ensemble learning method combining strong SSL learners with weak learners, achieving the highest scores across several metrics in the VoiceMOS Challenge 2022 [7]. Wang et al. [8] introduce a retrieval method to enhance the MOS prediction system and design a fusion network to optimize the retrieval scope and fusion weights.

While MOS prediction systems have demonstrated commendable overall performance, their deployment in real-world, open environments poses significant challenges. The inconsistency in performance across various samples within the dataset, coupled with a pronounced decline in effectiveness when encountering zero-shot out-of-domain (OOD) samples [9], underscores the limitations of current systems. Active research is underway to address these issues. Qi et al. [10] innovatively integrate supervised LE-SSL-MOS with unsupervised methods, marking a significant advancement in evaluating the OOD French synthesized speech, and the noisy and enhanced speech. Shen et al. [11] develop a speech quality assessment transformer with listener-dependent modeling, showcasing state-of-the-art performance in zero-shot evaluation of singing voice conversion samples. Nevertheless, as highlighted in the VoiceMOS Challenge 2023 [12], the objective to create a model capable of reliably predicting MOS in diverse real-world scenarios

remains unfulfilled, making general MOS prediction in practical applications an ongoing area of research.

From our perspective, a significant limitation hindering the MOS prediction systems from applying to the real and open world is they almost *exclusively focus on point estimates—neglect to provide uncertainty information*. Uncertainty estimation is an important building block of reliable, interpretable, and practical MOS prediction systems. On the one hand, uncertainty can help users to better understand the reliability of the forecast and make more rational decisions based on their needs. On the other hand, through the filtration of uncertainties, the model can reject unreliable predictions, thereby exhibiting stable performance. It is a crucial advantage given the current challenge of inconsistent MOS prediction accuracy and the need to address the diversity of samples in open-world scenarios.

The existing literature scarcely mentions uncertainty estimation in MOS prediction tasks, with only a few studies of weak relevance. Liang et al. [13] propose a deep neural network-based approach to estimate the posterior distribution of the MOS, replacing the traditional point estimate method. But the mean of the posterior distribution could be considered a new form of point estimation and additional information within the distribution is overlooked in testing. Ravuri et al. [14] demonstrate that the uncertainty obtained from SSL models exhibits a correlation with the MOS, facilitating the assessment of speech quality in low-resource settings. This uncertainty serves as a strategy to derive quality estimates, while the level of confidence associated with these estimates has yet to be determined.

Given the significant yet underexplored impact of uncertainty on MOS prediction, we undertake a systematic discussion and investigation into the uncertainties inherent in MOS prediction tasks. Specifically, we analyze the sources of uncertainty in the MOS prediction, categorizing them into aleatoric and epistemic uncertainties. Building on this foundation, we propose an uncertainty-aware MOS prediction framework. This framework not only assesses speech quality but also quantifies the associated uncertainty. Heteroscedastic regression and Monte Carlo (MC) dropout techniques are employed for aleatoric and epistemic uncertainty, respectively. To the best of our knowledge, this is the first MOS prediction system that can make uncertainty estimation. We validate the effectiveness of our uncertainty modeling methods through experiments and explore the benefits of incorporating uncertainty into the application of MOS prediction models in the real and open world.

2. Uncertainty in MOS Prediction

2.1. Category of Uncertainty

Uncertainty is commonly classified into two types: aleatoric uncertainty and epistemic uncertainty [15, 16].

* Corresponding author.

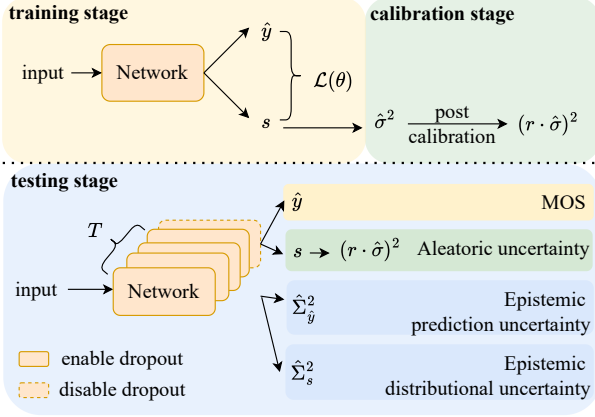


Figure 1: The pipeline of our uncertainty-aware MOS prediction system, including training, calibration, and testing stages.

- Aleatoric uncertainty, or data uncertainty, refers to the inherent randomness or noise present within the data itself. It is irreducible and intrinsic to the data being analyzed.
- Epistemic uncertainty, on the other hand, accounts for the uncertainty in the model parameters. Generally, it is caused by a lack of knowledge of the neural network [17]. Unlike aleatoric uncertainty, epistemic uncertainty can be reduced when more information or data becomes available.

2.2. Sources of Uncertainty

The main sources of uncertainty in MOS prediction include:

1. **The diversity of synthetic speech.** Driven by the influence of synthesized content, the target speaker, and the synthesis algorithm, the produced speech exhibits a high degree of complexity and variability.
2. **The subjectivity of annotation.** MOS reflects people’s subjective opinions on speech quality. Influenced by cultural background, perceptual capabilities, and personal preference, the opinion annotations are highly uncertain.
3. **Small-scale training data.** Gathering manually scored data is challenging due to limited datasets. This scarcity restricts models’ ability to learn diverse speech variations, leading to reduced effectiveness and increased uncertainty.
4. **OOD test samples.** Due to the lack of external knowledge, models cannot interpret OOD samples and often produce unreliable and uncertain predictions.

Of the four aspects discussed, the initial two are categorized under aleatoric uncertainty – rooted in randomness, and the subsequent two under epistemic uncertainty – stemming from knowledge limitations.

3. Methodology

As shown in Figure 1, the model simultaneously learns to predict speech quality MOS and capture aleatoric uncertainty in the training stage by heteroscedastic regression. The aleatoric uncertainty is further calibrated by scalar r in the calibration stage. In the test stage, two kinds of epistemic uncertainty are modeled by the MC dropout, performing T times forward propagation on the network with dropout enabled and collecting the output results to estimate uncertainty.

3.1. Aleatoric Uncertainty Modeling

Aleatoric uncertainty can be further divided into homoscedastic uncertainty and heteroscedastic uncertainty [18]. Homoscedastic uncertainty implies a consistent level of uncertainty across different inputs, heteroscedastic uncertainty refers to the uncertainty changes depending on different inputs. Data of MOS prediction tasks exhibits heteroscedasticity because the level of uncertainty differs across inputs which is influenced by the TTS system and the synthesized text. Therefore we model the aleatoric uncertainty as a data-dependent parameter σ^2 .

Pre-estimation. Following Kendall et al. [18] and Laves et al. [19], we also include σ^2 as one of the training targets of the model. Formally, the aleatoric uncertainty-aware model f_θ takes the data x as input and outputs the target that follows the Gaussian distribution with mean equal to $\hat{y}(x)$ and variance $\hat{\sigma}^2(x)$, which is

$$f_\theta(x) \sim \mathcal{N}(\hat{y}(x), \hat{\sigma}^2(x)). \quad (1)$$

We use the Negative Gaussian Log-Likelihood loss (NLL) as the optimization objective of the model. Given a training set \mathcal{D} , it can be expressed as:

$$\mathcal{L}(\theta) = - \sum_{i=1}^D \log p(y_i | \hat{y}_i, \hat{\sigma}_i^2) \quad (2)$$

$$= \sum_{i=1}^D \left(\frac{\log(\hat{\sigma}_i^2)}{2} + \frac{(y_i - \hat{y}_i)^2}{2\hat{\sigma}_i^2} + \text{const} \right), \quad (3)$$

where $D = |\mathcal{D}|$ and y_i denotes the label. In practice, we directly estimate $s_i := \log(\hat{\sigma}_i^2)$ to avoid numerical instability.

Post-calibration. Based on obtaining a preliminary estimate of aleatoric uncertainty through heteroscedastic regression, we further calibrate the uncertainty by introducing a simple scalar parameter r . More specifically, after training the model f_θ , we adjust its uncertainty on the calibration set \mathcal{C} . The adjusted uncertainty is denoted as $(r \cdot \hat{\sigma})^2$, which is derived by applying a scaling factor r to the original uncertainty $\hat{\sigma}^2$. Thus the probability density function of the target value is $\mathcal{N}(\hat{y}(x), (r \cdot \hat{\sigma}(x))^2)$.

Substituting the new standard deviation $r \cdot \hat{\sigma}(x)$ into the NLL loss function, we obtain the optimization objective for finding the best scaling factor r :

$$\mathcal{L}(r) = - \sum_{i=1}^C \log p(y_i | \hat{y}_i, (r \cdot \hat{\sigma}_i)^2) \quad (4)$$

$$= C \log(r) + \frac{1}{2} r^{-2} \sum_{i=1}^C \frac{(y_i - \hat{y}_i)^2}{(\hat{\sigma}_i)^2} + \text{const}, \quad (5)$$

where $C = |\mathcal{C}|$. Since the model f_θ is fixed during the calibration stage, \hat{y} and $\hat{\sigma}$ are treated as known constants. By setting the derivative of $\mathcal{L}(r)$ with respect to s to zero, we find the optimal value of r is:

$$r = \pm \sqrt{\frac{1}{C} \sum_{i=1}^C \frac{(y_i - \hat{y}_i)^2}{(\hat{\sigma}_i)^2}}, \quad (6)$$

which reflects how the scaling factor minimizes the entire loss function under the given dataset and model prediction uncertainty. Therefore, by Equation 6, we can simply obtain the calibration coefficients without any iterative optimization process.

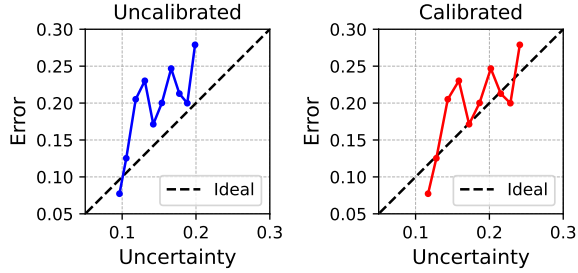


Figure 2: The error-uncertainty curve before and after calibration in BVCC test set.

3.2. Epistemic Uncertainty Modeling

To capture the epistemic uncertainty Σ^2 , we use MC dropout [20]. Dropout is a popular regularization technique that prevents over-fitting by randomly disabling a portion of neurons during the training time. Gal and Hahramani [20] interpret dropout training in networks as approximate Bayesian inference in deep Gaussian processes and propose to enable dropout at test time to quantify the uncertainty.

In our system, the model is trained with dropout. At the test time, we run T stochastic forward passes with dropout enabled at a given probability p . Test time outputs can be seen as samples from the approximate Bayesian. Thus we use the variance of the results as the predictive epistemic uncertainty. Corresponding to the two outputs of the model: the speech quality y and the log of variance s , the epistemic uncertainty is modeled from two perspectives: the epistemic prediction uncertainty $\hat{\Sigma}_y^2$ and the epistemic distributional uncertainty $\hat{\Sigma}_s^2$.

Epistemic prediction uncertainty. The epistemic uncertainty associated with the prediction, \hat{y} , quantifies the variability in the model’s quality prediction outcomes. It measures the reliability of the model in quality prediction for a given input. This concept is formalized as:

$$\hat{\Sigma}_y^2 = \frac{1}{T} \sum_{t=1}^T \left(\hat{y}_t - \frac{1}{T} \sum_{t=1}^T \hat{y}_t \right)^2. \quad (7)$$

Epistemic distributional uncertainty. Since s is actually a feature related to data, this type of uncertainty quantifies the model’s confidence in capturing the underlying distribution of the input data. A high level of epistemic distributional uncertainty signifies the model’s inadequate comprehension of the data. The quantification of epistemic uncertainty for s is:

$$\hat{\Sigma}_s^2 = \frac{1}{T} \sum_{t=1}^T \left(s_t - \frac{1}{T} \sum_{t=1}^T s_t \right)^2. \quad (8)$$

4. Experiments

4.1. Experiments Setup

Datasets. In our experiments, we utilize the BVCC [21] for training, which comprises 7,106 English audio samples from 187 TTS and Voice Conversion (VC) systems included in previous challenges, as well as samples from the ESPnet-TTS [22]. For OOD testing, we employ VCC2018 [23], which consists of English VC system outputs from Voice Conversion Challenge 2018 and has minor overlap with BVCC. Additionally, we use BC2019 [24], which contains Mandarin TTS samples from the

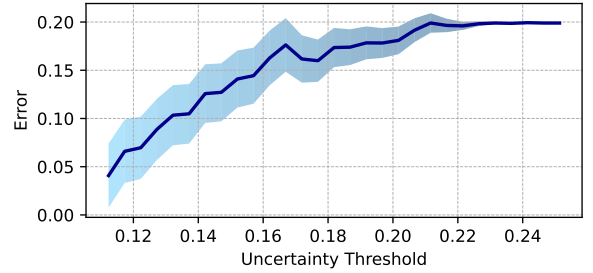


Figure 3: The relationship between uncertainty threshold and MSE of the reliable subset in the BVCC test set.

2019 Blizzard Challenge, and TMHINT-QI-II [25], released for the VoiceMOS challenge 2023, comprising noisy and enhanced speech from speech enhancement systems.

Architecture. We use wav2vec 2.0 base model [26] as the backbone. After feature extraction, we compress the temporal dimension by averaging and then apply a linear layer to reduce it from 768 to 256. Subsequently, we attach two task-specific heads: one for quality scores \hat{y} and another for s . Each head includes a dropout layer followed by two linear layers.

Implementation details. The models are trained with a batch size of 8 and a learning rate of 0.0003. The calibration phase is conducted on the validation set. In estimating epistemic uncertainty using MC dropout during testing, a dropout rate p of 0.5 and $T = 25$ executions are employed.

4.2. Evaluation Metrics

We evaluate the performance from two aspects: quality prediction and uncertainty estimation. The quality prediction aspect utilizes Mean Squared Error (MSE) and system-level Spearman’s Rank Correlation Coefficient (SRCC). For uncertainty estimation, we employ the following metrics:

Negative Log-likelihood (NLL). The NLL measures the fit of the model to the data. Its calculation follows the Equation 3.

Uncertainty calibration error (UCE). The metric assesses the data by dividing it into M equal-width bins B_1, B_2, \dots, B_m based on uncertainty levels, then calculates the average of the differences between each bin’s error and its uncertainty [19]:

$$\text{UCE} := \sum_{m=1}^M \frac{|B_m|}{n} |\text{err}(B_m) - \text{uncert}(B_m)|, \quad (9)$$

where n is the number of inputs and M is 10 in experiments.

Sharpness. It measures a model’s overall uncertainty level on a dataset through the average of uncertainties of all samples in dataset [27].

Area Under Curve (AUC). It’s used to assess the performance of in-domain and OOD binary classification.

4.3. Experimenting with Aleatoric Uncertainty Modeling

We first verify the effectiveness of heteroscedastic regression, denoted as HR, by making a comparison with MSE loss and another aleatoric uncertainty modeling method, denoted as KL [28]. MSE is commonly used in MOS prediction, yet lacks the capability for uncertainty estimation. The KL method models annotator scores as a target Gaussian distribution, minimizing KL divergence to reduce the distance between target and predicted distributions. Table 1 displays the average results of each

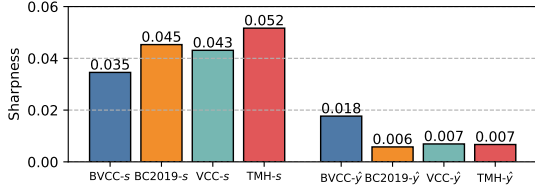


Figure 4: The sharpness for epistemic distributional uncertainty and epistemic prediction uncertainty across for test sets: BVCC, BC2019 (OOD), VCC2018 (OOD), and TMHINT-QI-II (OOD).

method across eight seeds. Our method not only surpasses the KL model across all metrics but also achieves competitive results in quality prediction when compared with the MSE model. This demonstrates the HR model’s strong quality prediction ability and capability to model aleatoric uncertainty effectively.

Table 1: Results of aleatoric uncertainty modeling methods.

Model	UCE ↓	NLL ↓	MSE ↓	SRCC ↑
MSE	-	-	0.223	0.930
KL	0.7523	1.049	0.254	0.927
HR	0.0456	0.657	0.203	0.932
calibrated HR	0.0338	0.632	0.203	0.932

We further validate the effectiveness of the calibration stage. The last two rows in Table 1 show that simple scaling calibration significantly enhances the UCE and NLL metrics, with $p = 0.0168$ and $p = 0.0372$, respectively. The fact that the MSE and SRCC metrics remain unchanged is attributed to the calibration solely addressing uncertainty and not modifying the quality prediction. In Figure 2, we show the error-uncertainty curves before and after calibration. The “ideal” curve represents perfect uncertainty estimation, that is, the uncertainty accurately reflects the model’s prediction error. The “calibrated” curve is closer to the ideal one than the “uncalibrated” curve, demonstrating the effectiveness of our calibration method.

Building on our advancements in accurately modeling aleatoric uncertainties, we apply the forecasted aleatoric uncertainty for selective prediction. Specifically, we determine an uncertainty threshold, discarding predictions beyond this value as unreliable, and subsequently assess the model’s performance on the deemed reliable subset, as illustrated in Figure 3. The shaded area in the figure shows the proportion of unreliable samples, and the curve points indicate the MSE of the reliable subset at various thresholds. As the threshold increases, so does the MSE, demonstrating that aleatoric uncertainty effectively captures accuracy and supports the viability of using uncertainty to manage model performance in practice.

4.4. Experimenting with Epistemic Uncertainty Modeling

We set two OOD scenarios to evaluate the effectiveness of the predictive epistemic uncertainty. We first test the BVCC with three OOD datasets and compare two kinds of epistemic uncertainty captured in Figure 4. These two types of epistemic uncertainty exhibit inconsistent trends and magnitudes: the sharpness of epistemic distributional uncertainty is larger overall and more sensitive to OOD samples, while the epistemic prediction uncertainty is the opposite. We analyze that this is because of the difference caused by the difficulty and target value range of the two output heads to learn. Consequently, epistemic distri-

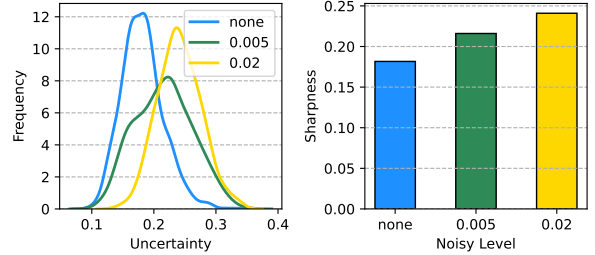


Figure 5: The distribution and sharpness of epistemic distributional uncertainty on data with different noise levels.

butional uncertainty proves more effective for OOD detection. We further investigate using this uncertainty for OOD classification, with results in Table 2. It shows strong performance on all three OOD datasets, particularly TMHINT-QI-II, likely due to its significant difference from the training domain.

Table 2: AUC of OOD detection, which refers to using uncertainty as a basis for classifying the mix BVCC and OOD data.

	BC2019	VCC2018	TMHINT-QI-II
AUC ↑	0.641	0.646	0.768
	noise level of 0.005		noise level of 0.02
AUC ↑	0.723		0.890

The second OOD evaluation scenario examines BVCC test set with different noise levels: none, 0.002, and 0.01, shown in Figure 5. The results indicate the sharpness of the model’s epistemic distributional uncertainty grows with noise increase, highlighting its sensitivity to OOD features. The lower part in Table 2 shows that using this uncertainty for OOD classification tasks offers promising performance, significant for applying the MOS prediction system in real and open-world settings.

4.5. Discussion

In these experiments, we conduct a preliminary analysis that showcases the potential applications of aleatoric and epistemic uncertainties. Accurately capturing aleatoric uncertainty in MOS prediction enables model performance understandable, and controllable by setting of appropriate thresholds to do selective prediction. Meanwhile, epistemic uncertainty displays a heightened sensitivity to OOD data, providing a viable approach for OOD detection. Equipped with selective prediction and OOD detection capabilities, the MOS prediction system can be reliably applied in real and open-world environments.

5. Conclusion

In this paper, we emphasize the importance of uncertainty modeling for a practical MOS prediction system and analyze the uncertainty sources of MOS prediction tasks. We then model aleatoric uncertainty and epistemic uncertainty using calibrated heteroscedastic regression and MC dropout, respectively. The experimental results show that our uncertainty-aware MOS system not only performs well but also holds promise for applications in selective prediction and OOD detection.

6. Acknowledgements

This work has been supported by the National Key R&D Program of China through grant 2022ZD0116307 and NSF China (Grant No.62271270).

7. References

- [1] 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*, 2019, pp. 1541–1545.
- [2] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnnet: Mos prediction for synthesized speech with mean-bias network," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 391–395.
- [3] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 896–900.
- [4] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8442–8446.
- [5] H. Wang, X. Zheng, and Y. Qin, "Intermediate-task learning with pretrained model for synthesized speech mos prediction," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 378–383.
- [6] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab System for Voice-MOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.
- [7] W. C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, 2022, pp. 4536–4540.
- [8] H. Wang, S. Zhao, X. Zheng, and Y. Qin, "RAMP: Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting," in *Proc. INTERSPEECH 2023*, 2023, pp. 1095–1099.
- [9] W.-C. Tseng, W.-T. Kao, and H. yi Lee, "DDOS: A MOS Prediction Framework utilizing Domain Adaptive Pre-training and Distribution of Opinion Scores," in *Proc. Interspeech 2022*, 2022, pp. 4541–4545.
- [10] Z. Qi, X. Hu, W. Zhou, S. Li, H. Wu, J. Lu, and X. Xu, "L-ssl-mos: Self-supervised learning mos prediction with listener enhancement," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–6.
- [11] K. Shen, D. Yan, L. Dong, Y. Ren, X. Wu, and J. Hu, "Squat-Id: Speech quality assessment transformer utilizing listener dependent modeling for zero-shot out-of-domain mos prediction," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–6.
- [12] E. Cooper, W.-C. Huang, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The voicemos challenge 2023: Zero-shot subjective speech quality prediction for multiple domains," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–7.
- [13] X. Liang, F. Cumlin, C. Schüldt, and S. Chatterjee, "DeePMOS: Deep Posterior Mean-Opinion-Score of Speech," in *Proc. INTERSPEECH 2023*, 2023, pp. 526–530.
- [14] A. Ravuri, E. Cooper, and J. Yamagishi, "Uncertainty as a predictor: Leveraging self-supervised learning for zero-shot mos prediction," in *Self-supervision in Audio, Speech and Beyond (SASB) workshop*, 2024.
- [15] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2019.
- [16] M. Hu, Z. Zhang, S. Zhao, M. Huang, and B. Wu, "Uncertainty in natural language processing: Sources, quantification, and applications," *arXiv preprint arXiv:2306.04459*, 2023.
- [17] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. M. Kruspe, R. Triebel, P. Jung, R. Roscher, M. Shahzad, W. Yang, R. Bamler, and X. Zhu, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, pp. 1513–1589, 2021.
- [18] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.
- [19] M.-H. Laves, S. Ihler, J. F. Fast, L. A. Kahrs, and T. Ortmaier, "Well-calibrated regression uncertainty in medical imaging with deep learning," in *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, ser. Proceedings of Machine Learning Research, T. Arbel, I. Ben Ayed, M. de Bruijne, M. Descoteaux, H. Lombaert, and C. Pal, Eds., vol. 121. PMLR, 06–08 Jul 2020, pp. 393–412.
- [20] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. F. Balcan and K. Q. Weinberger, Eds., vol. 48. New York, New York, USA: PMLR, 20–22 Jun 2016, pp. 1050–1059.
- [21] E. Cooper and J. Yamagishi, "How do Voices from Past Speech Synthesis Challenges Compare Today?" in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 183–188.
- [22] T. Hayashi, R. Yamamoto, K. Inoue, T. Yoshimura, S. Watanabe, T. Toda, K. Takeda, Y. Zhang, and X. Tan, "Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7654–7658, 2020.
- [23] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," in *Speaker Odyssey 2018*. ISCA, Jun. 2018, pp. 195–202, the Speaker and Language Recognition Workshop, Speaker Odyssey 2018 ; Conference date: 26-06-2018 Through 29-06-2018.
- [24] Z. Wu, Z. Xie, and S. King, "The blizzard challenge 2019," in *Proc. Blizzard Challenge Workshop*, 2019.
- [25] R. E. Zezario, Y.-W. Chen, S.-W. Fu, Y. Tsao, H.-M. Wang, and C.-S. Fuh, "A study on incorporating whisper for robust speech assessment," 2023.
- [26] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.
- [27] V. Kuleshov, N. Fenner, and S. Ermon, "Accurate uncertainties for deep learning using calibrated regression," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 2796–2804.
- [28] C. Zerva, T. Glushkova, R. Rei, and A. F. T. Martins, "Disentangling uncertainty in machine translation evaluation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 8622–8641.