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

# When The MOS Predictor Asks For Training Annotation In Cross Lingual/Domain Adaptation

*Natacha Miniconi*[1], *Meysam Shamsi*[1], *Anthony Larcher*[1]

[1]LIUM, University of Le Mans, France
{natacha.miniconi}, {meysam.shamsi}, {anthony.larcher}@univ-lemans.fr

## Abstract

The Mean Opinion Score (MOS) is widely used to assess speech synthesis quality, but requires costly human evaluation. Automatic MOS predictors have been developed to estimate MOS. Training and generalizing these predictors across different languages and domains remain difficult due to the high cost of labeled data. To optimize MOS prediction while minimizing human annotation efforts, we explore active learning. As far as we know, this is the first study to investigate the use of active learning as training sample selection strategies for enhancing MOS prediction. We investigate its effectiveness on two tasks: cross-domain and cross-lingual adaptation, comparing multiple selection strategies that rely on uncertainty or diversity measures. Our sample selection strategies have been compared with random selection. Among these strategies, Monte Carlo (MC) Dropout proved effective for cross-lingual adaptation, while perturbation noise performed well for cross-domain adaptation.

**Index Terms**: Synthesis speech quality assessment, MOS prediction, Active learning

## 1. Introduction

Speech synthesis has made significant progress in recent years, but its quality is still commonly evaluated using the Mean Opinion Score (MOS), a metric that relies on a costly subjective human evaluation. Although MOS is the most widely accepted metric, it can be controversial due to its dependence on individual perception [1, 2, 3].

The VoiceMOS Challenges [4, 5] was launched to develop automated quality predictors capable of handling multiple languages and domains [6, 7] including under-resourced languages where annotated data is especially limited. These predictors aim to reduce the reliance on human evaluation, which is both costly and time-consuming. The VoiceMOS Challenge demonstrated that these automated predictors can achieve results close to human perception, underlining their potential to provide scalable and efficient evaluation methods [8, 9, 10]. The scarcity of MOS-annotated data is an issue for effective model training and for ensuring that MOS predictors generalize to other languages or domains, where annotated data may be even more limited.

Active learning approaches can help overcome the scarcity of MOS-annotated data by selectively identifying the most informative samples for labelling, with different selection strategies, thereby minimizing human intervention and reducing overall costs. Given that speech synthesis itself generates synthetic data, data collection costs can be considered low, as no direct human intervention is needed to produce new samples. This makes active learning particularly well-suited to optimizing MOS prediction models, even on different datasets or in different languages, leveraging potential cross-domain/lingual transfer learning. Moreover, this approach can provide insights into the linguistic and acoustic characteristics of the training data, helping to refine dataset design and ensuring an efficient distribution of speech samples across different conditions and languages. To our knowledge, this is the first study to explore the use of active learning for MOS prediction in synthetic speech.

Our hypothesis is the following: active learning can optimize MOS prediction across different datasets or languages. This approach may also reduce the need for human annotations while enabling cross-domain and cross-lingual transfer learning. In order to validate this hypothesis, we define two tasks:

- **Task 1**: investigate whether a model trained on English data can be adapted to a different dataset in the same language.
- **Task 2**: evaluate whether a model trained on English data can be adapted to predict MOS for a different language.

## 2. Related Work

### 2.1. Active Learning Algorithm

Active learning is designed to optimize model performance by selectively querying the most informative samples from an unlabeled dataset for training set. This approach allows efficient use of annotation efforts by focusing on the most valuable data points, leading to improved model performance with minimal labeled data and a minimal budget. Active learning is widely applied in various domains, including physics [11], natural language processing, and speech processing [12, 13]. The general architecture of active learning consists of the following key components [14, 13]:

**1. Initial Model Training:** the process begins with an initial model, which is typically trained on a small labeled dataset. This model serves as a starting point for the active learning cycle.

**2. Unlabeled Data Pool:** a large pool of unlabeled data is maintained, from which the model selects samples for annotation that are expected to provide the most benefit.

**3. Sample Selection Strategy:** at each iteration, an acquisition function is applied to identify the most informative samples from the unlabeled pool, assigning a selection score for each sample. Various selection strategies can be used, such as:

- *Random sampling:* selecting samples randomly from the unlabeled pool, used for comparison in our study.
- *Uncertainty-based sampling:* selecting samples for which the model exhibits the highest uncertainty with the help of various methods like the Monte Carlo (MC) dropout [15], Bald [16] or Adversarial perturbations [16].
- *Diversity-based sampling:* ensuring selected samples are diverse and representative of the entire data distribution with


10.21437/Interspeech.2025-496

the help of various methods, clustering and distance between the unlabeled sample and the already labeled training samples [16].

- *Query-by-Committee:* using an ensemble of models to identify samples with the highest disagreement among predictions [13].

**4. Data Annotation & Model Updating:** the selected samples are sent for annotation, where ground truth labels are obtained. This step can be performed by a human or a machine with a pseudo-annotation. Once the newly labeled data is available, it is added to the existing training set, and the model training continues to incorporate the new information [14].

**5. Evaluation:** after this additional training, the model is evaluated on a test set with different metrics [13].

# 3. Methodology

Our protocol simulates an active learning framework, adapting a pretrained baseline model [7] by iteratively incorporating informative samples from an unlabeled data pool. The process is structured into the following key steps:

## 3.1. Initial Model Training

We begin with a baseline model designed to predict the MOS by regression. This model captures generalizable features from an available dataset [7].

## 3.2. Active Learning Iterations

With the baseline model established, an iterative active learning process is employed to enrich a task-specific datasets. This process involves the following steps repeated over multiple iterations:

- **Unlabeled Data Selection:** in each iteration, a subset of samples is selected from the unlabeled data pool using various sampling strategies that prioritize the most informative samples for annotation and model refinement, thereby maximizing model improvement.
- **Model Updating:** the selected samples are added incrementally to the training data, and the model is finetuned to incorporate the newly acquired information, progressively enhancing its performance on task-specific data.
- **Evaluation:** after each iteration, the updated model is evaluated using a separate test set to measure the overall performance.

## 3.3. Sample Selection

We use two strategies for sample selection with several selection criteria described below.

- **Uncertainty**

Uncertainty is modeled using two distinct approaches. The first approach is based on **MC Dropout** [15], which involves performing several forward passes (25 passes, following [17]) with dropout in wav2vec and linear layers, allowing us to estimate the model's uncertainty. The second, called perturbation noise, introduces **white noise** to the input data with a 5 dB level and compares the predictions MOS obtained from the noisy input to those obtained from the original data. If the addition of noise leads to significant changes in the predictions, it indicates that the model is fragile and lacks robustness on the chosen sample.

- **Diversity**

We employ cosine similarity metric as defined in the equation 1, to measure the dissimilarity between the unlabeled samples $d_j$ and the training samples $d_i$.

$$S_{div}(d_j) = \frac{1}{N} \sum_{i=0}^{N} \left(1 - \frac{\mathbf{X}_{d_j} \cdot \mathbf{X}_{d_i}}{\|\mathbf{X}_{d_j}\| \, \|\mathbf{X}_{d_i}\|}\right) \quad (1)$$

For a given unlabeled sample $d_j$ represented by $\mathbf{X}_{d_j}$ in a embedding space, we compute the diversity score $S_{div}$ as an indicator of how distinct the data sample is from the existing training data.

# 4. Experiments

## 4.1. Experimental Setup

### *4.1.1. Datasets*

For our experiments, we use the BVCC dataset [18] to train the initial system. It contains 7,106 English audio samples from 187 Text-to-Speech (TTS) and Voice Conversion (VC) systems used in previous challenges, along with additional samples generated by ESPnet-TTS. We also rely on the SOMOS dataset, recently evaluated [19], for Task 1. SOMOS contains 20,100 English audio samples from 200 different TTS systems. The 882 French audio samples from 21 TTS systems included in the Blizzard Challenge 2023 (BC 2023) dataset are used for Task 2 [20]. All audio segments are resampled to 16 kHz. All these datasets include MOS annotation.

### *4.1.2. MOS predictor model*

Our initial model, based on the SSL-MOS framework [7] from previous Voice MOS challenges, is based on the wav2vec 2.0 base model [21] with frozen parameters. Embeddings from each layer of wav2vec 2.0 are concatenated using Multi-level Feature Aggregation (MFA) and compressed using a mean pooling. Two additional linear layers are used to reduce the dimension and output a single prediction of the MOS. A dropout layer with a rate of 0.2 is applied between the two linear layers. In total, the model has 84M trainable parameters.

### *4.1.3. Data Architecture for Active Learning*

In our approach, we proceed to 4 iterations, representing an incremental annotation budget, of active learning to progressively refine the model and adapt it to the targeted task. Throughout these iterations, the training dataset evolves, allowing the model to continuously improve its performance based on newly acquired information. Our baseline is established with the initial model before further refinements. The simulation does not include real-time annotation; instead, we have direct access to the ground-truth annotations requested by the model. The model selects a sample, and its corresponding annotation is immediately available for training. See Table 1 for details.

Table 1: *Repartition of dataset (train/validation/test/unlabeled data) for each iteration, where $i$ represents the iteration index*

| DATA | Train | Validation | Test | Unlabeled data |
|---|---|---|---|---|
| **Initial Model Training** | | | | |
| BVCC | 4974 | 1066 | 0 | - |
| SOMOS | 0 | 0 | 3000 | - |
| BC 2023 | 0 | 0 | 294 | - |
| **Iteration 1-4** | | | | |
| BVCC | 4974 | 0 | 0 | 0 |
| SOMOS | $20 \times i$ | 3000 | 3000 | $14100 - 20 \times (i-1)$ |
| BC 2023 | $20 \times i$ | 147 | 294 | $441 - 20 \times (i-1)$ |

**Initial Model Training:** we train the model using the BVCC dataset, which is split into two parts: a *training set*, used to

learn the initial model, and a *validation set* (Table1). This stage aims to establish a baseline model before adaptation to specific tasks.

**Active Learning Iterations 1-4 (Task-Specific Adaptation):** once the initial training is completed, the model is adapted to each tasks with four iterations. The SOMOS dataset (Task 1) is divided into three subsets: an *unlabeled dataset*, a *validation set* which is used to save the best model at each iteration, and a *test set* (Table1). In each active learning iteration, 20 candidates are selected from the *unlabeled dataset* based on different selection strategies and incrementally added to the training data after simulating the annotation process. The model is then evaluated on the *test set* of SOMOS. The same approach is applied with BC 2023 (Task 2).

**Weighted Batch Sampling**: we employ a weighted sampling strategy that prioritizes the inclusion of samples from underrepresented data. We assign a weight $w$ to each sample based on the inverse of the number of available training samples; $w = \frac{1}{N}$ Where $N$ represents the total number of BC 2023 or SOMOS samples in the training set. The BVCC samples receive a corresponding weight based on their dataset size, ensuring a balanced representation in each training batch. This weighting helps to control the importance of new data for adaptation.

#### *4.1.4. Hyperparameter*

The models are trained with a batch size of 4 a learning rate of 0.0005, an SGD optimizer, 100 epochs and an early stopping of 10 epochs on the *validation set*. The experiments were conducted on a NVIDIA RTX 8000 GPU, with computation times ranging from 1 to 3 hours depending on the selected strategy and dataset used.

### 4.2. Evaluation Metrics

For the MOS prediction task, the commonly used evaluation metrics, including system-level and utterance-level, are Mean Squared Error (MSE), Linear Correlation Coefficient (LCC), Spearman's Rank Correlation Coefficient (SRCC), and Kendall's Tau Rank Correlation (KTAU)[4, 5]. Following [4], we evaluate the performance of our strategies using SRCC at utterance-level. We focus on SRCC because our goal is to capture the relative ranking of utterance quality rather than the absolute MOS values, which are often biased. To ensure the reliability and robustness of our approach, we conducted the active learning process over 5 runs with different seeds.

### 4.3. Loss Function

The loss function used is the Concordance Correlation Coefficient (CCC) loss [22], chosen to align the model's optimization with the evaluation metric SRCC, unlike the baseline approach which used L1 Loss [7]. This choice was motivated by previous research [23] on optimizing MOS predictors, which demonstrated that replacing L1 loss with a correlation-based loss led to improved performance [7].

## 5. Results

In order to evaluate the performance of our strategies, we compare each strategy with a random selection (Figure. 1).

Additionally, to analyze the characteristics of the selected samples, we compare their distribution based on duration and MOS against the entire *unlabeled dataset*. This allows us to detect potential over-representation of certain types of data and determinate whether certain types of data are preferred over others. We use Probability Density Functions (PDFs) to visualize and compare the distributions (Figure 2,3). Moreover, we apply a Kolmogorov-Smirnov (KS) test to statistically assess whether

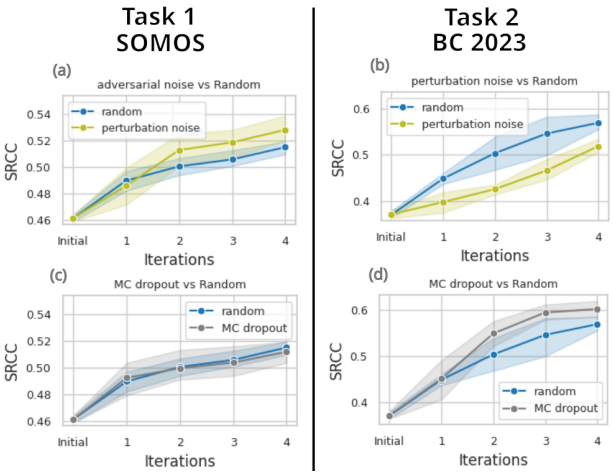



Figure 1: *Evaluation of uncertainty strategies in active learning with SRCC metric in utterance-level (y-axis) for each iteration (x-axis) with a 95 % confidence interval for Tasks 1 and 2. (a) and (b) for perturbation noise strategy. (c) and (d) represent the performance of MC dropout.*

there are significant differences between the selected samples and the *unlabeled dataset* with a ($p$ -$value < 0.01$). This helps determine whether the selection process introduces substantial deviations from the target distribution.

### 5.1. Experiments with Uncertainty Strategy

#### *5.1.1. Perturbation Noise*

The uncertainty of each sample, as selection score, is measured by calculating the absolute difference between the predicted MOS of the clean and noisy versions. Samples with the largest differences are considered the most uncertain, and the top 20 are selected for inclusion in the training process.

For Task 1, selection based on perturbation noise outperforms random selection (Figure 1a), suggesting that this level of perturbation effectively captures uncertainty. When analyzing the number of iterations, we observe that at iteration 1, which corresponds to a small annotation budget, performance remains similar to random selection. It is only at iteration 2 that the approach demonstrates improvements beyond random selection.

In this task, the model selects samples annotated with higher MOS, more frequently than their proportion in the *unlabeled dataset*. The distribution of MOS in this dataset is predominantly between 3 and 4, with very few samples between 1 and 2. Having more MOS samples within a range of 3 to 4 appears to be beneficial for this strategy, as adding noise to low-quality audio could amplify distortions and make evaluation more challenging. Additionally, we observe that this approach tends to select longer audio samples compared to the general distribution of the unlabeled data. Our KS test confirms a significant difference in both MOS and duration distributions between the selected samples and the unlabeled dataset. These differences could be explained by the greater linguistic variability present in longer audio files, which leads to higher uncertainty and makes them more informative for the model. This behaviour could indicate that the model initially trained on a different dataset, learns to adapt its selection criteria to the new domain by prioritising samples with higher variability.

In Task 2, adding white noise resulted in a performance drop (Figure 1b), failing to provide meaningful improvements.

This suggests that the perturbation noise did not effectively capture uncertainty in this case.

The distribution of both duration and MOS in the selected samples differs from that of the unlabeled data. Our KS test confirmed this significant difference. Interestingly, the algorithm tends to select longer audio files, as for Task 1, but unlike SOMOS, the BC 2023 *unlabeled dataset* contains fewer long-duration samples. This difference in dataset characteristics could contribute to the model's difficulties, as it may limit the availability of sufficiently informative long samples for uncertainty assessment. Thus, the selection process leads to an overrepresentation of MOS between 4 and 5. The observed biases in selection suggest that the model relies too heavily on patterns learned from the source language, leading to inefficiencies when applied to a dataset with different statistical properties.

#### *5.1.2. MC dropout*

To establish a strategy incorporating uncertainty for a regression task, we use MC dropout during inference on the *unlabeled dataset*. The uncertainty of each sample, as selection score, is measured by computing the variance of the predicted MOS across multiple forward passes. In each iteration, we then select the 20 most uncertain samples, corresponding to those with the highest variance among their predictions.

For Task 2 (see Figure 1d), the MC dropout strategy outperforms random selection in terms of SRCC performance, but the results remain relatively close. Analysing the number of iterations, we observe that, at iteration 1, corresponding to a small annotation budget, the performance of MC dropout is similar to that of random selection. The performance difference between the two strategies becomes more pronounced starting from iteration 2. The selected samples follow a distribution almost identical to that of the *unlabeled dataset* in terms of both duration and MOS (Figure 2). This indicates that the selection process does not introduce any overrepresentation and ensures for that duration and MOS, based on these characteristics, are well represented. However, for Task 1, the results show that

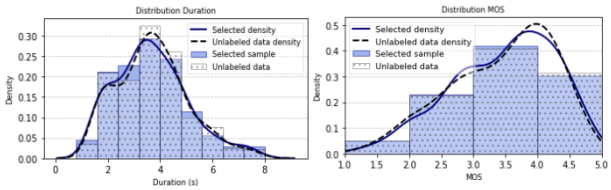



Figure 2: *The distribution of the duration at the left and at the right MOS for the unlabeled dataset and the samples selected from BC 2023 by the MC dropout strategy, according to KS test, we don't have any evidence they are different.*

predictions are very close to random (Figure 1c), indicating that this approach is less effective on this dataset compared to the Task 2. The selected samples follow a distribution almost identical to that of the *unlabeled dataset* in terms of both duration and MOS. This suggests that the MC dropout strategy does not introduce any selection bias in terms of these attributes. These results align with previous observations from [24] and [25], which demonstrated that uncertainty-based active learning methods can fail to generalize effectively across different tasks.

### 5.2. Experiments with Diversity

To incorporate diversity into the sample selection process, we aim to select the most diverse samples from the *unlabeled dataset* in relation to the current training dataset. Diversity as the selection score is measured with equation 1. Feature embeddings are extracted using wav2vec 2.0, with temporal averaging applied to obtain a fixed-length vector for each sample. We then select the top 20 samples with the highest selection score.

We observe that our approach is not effective for Task 1 and 2. This technique primarily selects audio samples with MOS around 2 and shorter durations (Figure 3), leading to an overrepresentation of such samples in the selected subset for both SOMOS and BC 2023. This is the opposite of the result reported in [26] for TTS corpus design. For SOMOS, the overrepresentation of short-duration audio samples is particularly pronounced. This extreme imbalance could be the main reason in the failure of the diversity strategy on SOMOS, as the selected subset may not provide sufficient diversity in terms of content. KS test confirms that the distribution of selected samples significantly differs from that of the *unlabeled dataset*.

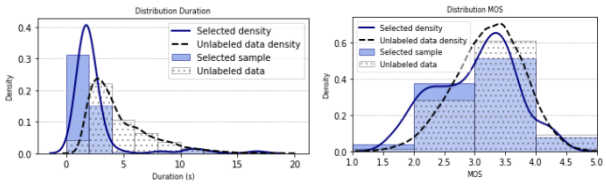



Figure 3: *The distribution of the duration at the left and MOS at the right for the unlabeled data and the samples selected from SOMOS by the diversity strategy, according to KS test we have different distribution between selected and unlabeled data.*

## 6. Conclusion

This work explored, for the first time, an active learning strategy to improve MOS prediction of synthetic speech across domains and languages. The main contribution of this paper is to propose an efficient method to reduce the annotation costs. In cross-domain adaptation (Task 1), our results show that selection with perturbation noise yields higher results in terms of SRCC than random selection. On the other hand, MC dropout was less effective compared to random selection. For cross-lingual adaptation (Task 2), MC dropout show potential and delivered higher performance than random selection, while perturbation noise was not effective, suggesting that further investigation is needed to understand its impact across different languages. In both tasks, selection based on diversity yields lower SRCC results than random selection, likely due to the over-representation of certain data types. This limitation could present an opportunity for improvement: by incorporating mechanisms that ensure a more balanced distribution, these approaches could be more effective. Future active learning strategies could be designed to align with a target distribution, preventing sampling biases. This issue might also be linked to the fact that Wav2Vec 2.0 does not capture task-relevant characteristics. By an exploration on the embedding space we haven't detect an observable relation between MOS/TTS system and sample distance. A future investigation would be helpful. Also, we explore hybrid approaches that combine different selection methods which show promising preliminary results. The perspective of designing efficient datasets for model training, strategies such as perturbation noise, which primarily target long-duration samples, could help guide the creation of corpora better suited to model.

## 7. Acknowledgements

We deeply thank Simon Petitrenaud for his advices. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101007666. This work was granted access to the HPC resources of IDRIS under the allocation 2024-A0161014876 made by GENCI.

## 8. References

[1] A. Kirkland, S. Mehta, H. Lameris, G. E. Henter, E. Szekely, and J. Gustafson, "Stuck in the mos pit: A critical analysis of mos test methodology in tts evaluation," in *12th ISCA Speech Synthesis Workshop 2023 (SSW2023)*, pp. 41–47.

[2] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech Language*, vol. 84, p. 101577, 2024.

[3] J. Camp, T. Kenter, L. Finkelstein, and R. Clark, "MOS vs. AB: Evaluating Text-to-Speech Systems Reliably Using Clustered Standard Errors," in *Interspeech*, 2023, pp. 1090–1094.

[4] 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.

[5] W.-C. Huang, E. Cooper, Y. Tsao, H.-M. Wang, T. Toda, and J. Yamagishi, "The VoiceMOS Challenge 2022," in *Interspeech*, 2022, pp. 4536–4540.

[6] T. Sellam, A. Bapna, J. Camp, D. Mackinnon, A. P. Parikh, and J. Riesa, "Squid: Measuring speech naturalness in many languages," *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

[7] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization Ability of MOS Prediction Networks," *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 8442–8446, 2021.

[8] Z. Qi, X. Hu, W. Zhou, S. Li, H. Wu, J. Lu, and X. Xu, "LE-SSL-MOS: Self-Supervised Learning MOS Prediction with Listener Enhancement," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023, pp. 1–6.

[9] H. Wang, S. Zhao, X. Zheng, and Y. Qin, "RAMP: Retrieval-Augmented MOS Prediction via Confidence-based Dynamic Weighting," in *Interspeech*, 2023, p. 1095–1099.

[10] W.-C. Huang, E. Cooper, and T. Toda, "MOS-Bench: Benchmarking Generalization Abilities of Subjective Speech Quality Assessment Models," *arXiv preprint arXiv:2411.03715*, 2024.

[11] W. Han, H. Li, H. Ruan, L. Ma, J. Sun, and B. Schuller, "Active learning for dimensional speech emotion recognition," in *Interspeech*, 2013, pp. 2841–2845.

[12] T. V. Ho, K. Dohi, and Y. Kawaguchi, "Stream-based Active Learning for Anomalous Sound Detection in Machine Condition Monitoring," in *Interspeech*, 2024, pp. 102–106.

[13] E. Surzhikova and J. Proppe, "regAL: Python Package for Active Learning of Regression Problems," 2024.

[14] D. Li, Z. Wang, Y. Chen, R. Jiang, W. Ding, and M. Okumura, "A Survey on Deep Active Learning: Recent Advances and New Frontiers," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–21, 2024.

[15] 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, vol. 48. PMLR, 2016, pp. 1050–1059.

[16] K.-H. Huang, "Deepal: Deep active learning in python," 2021.

[17] H. Wang, S. Zhao, J. Zhou, X. Zheng, H. Sun, X. Wang, and Y. Qin, "Uncertainty-Aware Mean Opinion Score Prediction," in *Interspeech*, 2024, pp. 1215–1219.

[18] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" *arXiv preprint arXiv:2105.02373*, 2021.

[19] G. Maniati, A. Vioni, N. Ellinas, K. Nikitaras, K. Klapsas, J. S. Sung, G. Jho, A. Chalamandaris, and P. Tsiakoulis, "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis," in *Interspeech*, 2022, pp. 2388–239.

[20] O. Perrotin, B. Stephenson, S. Gerber, and G. Bailly, "The Blizzard Challenge 2023," in *18th Blizzard Challenge Workshop*, 2023, pp. 1–27.

[21] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 12 449–12 460.

[22] L. Goncalves, A. N. Salman, A. R. Naini, L. Moro-Velázquez, T. Thebaud, P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024 - Speech Emotion Recognition Challenge: Dataset, Baseline Framework, and Results," in *The Speaker and Language Recognition Workshop*, 2024, pp. 247–254.

[23] B. Hu and Q. Li, "Correlation Loss for MOS Prediction of Synthetic Speech," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2022, pp. 1910–1914.

[24] D. Lowell, Z. C. Lipton, and B. C. Wallace, "Practical obstacles to deploying active learning," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 21–30.

[25] A. Snijders, D. Kiela, and K. Margatina, "Investigating multi-source active learning for natural language inference," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 2187–2209.

[26] M. Shamsi, J. Chevelu, N. Barbot, and D. Lolive, "Corpus design for expressive speech: impact of the utterance length," in *10th International Conference on Speech Prosody 2020*. Tokyo, Japan: ISCA, May 2020, pp. 955–959.