

# PARTIAL RANK SIMILARITY MINIMIZATION METHOD FOR QUALITY MOS PREDICTION OF UNSEEN SPEECH SYNTHESIS SYSTEMS IN ZERO-SHOT AND SEMI-SUPERVISED SETTING

Hemant Yadav\*, Erica Cooper, Junichi Yamagishi, Sunayana Sitaram, Rajiv Ratn Shah

{hemantya,rajivrtn}@iiitd.ac.in, {ecooper,jyamagis}@nii.ac.jp, sunayana.sitaram@microsoft.com

## ABSTRACT

This paper introduces a novel objective function for quality mean opinion score (MOS) prediction of unseen speech synthesis systems. The proposed function measures the similarity of relative positions of predicted MOS values, in a mini-batch, rather than the actual MOS values. That is the partial rank similarity is measured ( $PRS$ ) rather than the individual MOS values as with the L1 loss. Our experiments on out-of-domain speech synthesis systems demonstrate that the  $PRS$  outperforms L1 loss in zero-shot and semi-supervised settings, exhibiting stronger correlation with ground truth. These findings highlight the importance of considering rank order, as done by  $PRS$ , when training MOS prediction models. We also argue that mean squared error and linear correlation coefficient metrics may be unreliable for evaluating MOS prediction models. In conclusion,  $PRS$ -trained models provide a robust framework for evaluating speech quality and offer insights for developing high-quality speech synthesis systems. Code and models are available at [github.com/nii-yamagishilab/partial\\_rank\\_similarity/](https://github.com/nii-yamagishilab/partial_rank_similarity/)

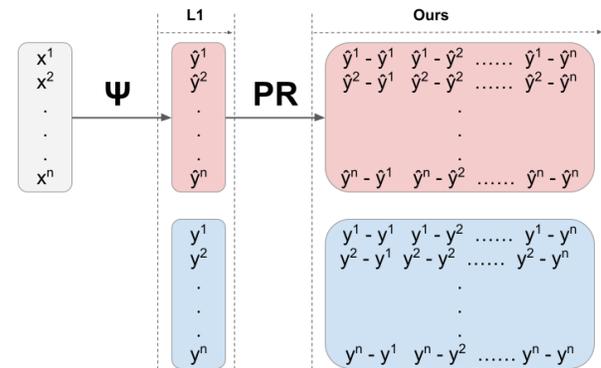
**Index Terms**— MOS, automatic MOS prediction, Rank order, Naturalness, Quality, L1, Text-to-speech, Voice conversion

## 1. INTRODUCTION

Recent advances in machine learning have significantly improved synthesized speech, which consequently has become more integrated into our daily lives. Unlike machine translation, which uses BLEU score [1] for algorithmic evaluation, text-to-speech (TTS) synthesis and voice conversion (VC) heavily rely on human ratings from listening tests. Crowdsourcing [2] and web-based tests have expanded participant pools and accelerated experimentation; however, these are still more costly and time-consuming than automated evaluation metrics. Thus, there is increasing interest in developing reliable objective quality measures for synthesized speech.

Mean opinion score (MOS) serves as an attractive evaluation methodology for researchers due to its ability to provide a single, easily comparable numerical result. In a MOS test, listeners evaluate synthesized samples one by one and assign them an integer rating on a scale (e.g., 1-5) on the basis of some criteria such as naturalness. All ratings per system are averaged together to obtain a final mean score. With the recent advances in machine learning, attention has turned to data-driven synthesized speech quality prediction – in particular, automatic MOS prediction. Early works on neural network-based data-driven MOS prediction [3, 4, 5] found that although MOS ratings from the same listening test as the training data could be well-predicted, these models do not generalize well to data from other listening tests due to differences in the listener pool, testing interface,

\*The work was performed while at National Institute of Informatics (NII), Tokyo, Japan.



**Fig. 1.** A typical MOS prediction pipeline. It consists of a function approximator  $\psi$  to predict the MOS scores, given an audio file. Standard practice is to calculate the L1 loss using the predictions and ground-truth MOS values. In this work, we apply a partial ranking function  $PRS$  and then apply the p-norm loss over the output matrices of prediction and ground-truth MOS values.

systems under consideration, and many other factors outlined by [6]. The authors of [7, 8] showed that finetuning self-supervised learning (SSL) based models for speech, such as Wav2Vec2 [9], could increase the generalization ability of automatic MOS predictors on out-of-domain (OOD) datasets. To mitigate the domain mismatch between pretrained SSL models, which have only seen examples of natural speech, and the MOS prediction task for synthesized speech, [10] conducted domain-adaptive pretraining [11]. They show improvements on an OOD dataset, most notably in the zero-shot and few-shot settings. However, predicting unseen systems from OOD listening tests remains challenging. In fact, this is a crucial scenario for researchers and engineers utilizing automatic quality predictors. They often develop and assess new, unseen systems, including those for different languages, including low-resource languages.

It was noted in [7] that in the zero-shot prediction scenario, where the model has not been finetuned on any labeled data from the target listening test, that mean squared error (MSE) can be very high even when the correlations with true MOS values are reasonable. We hypothesize that, indeed, predicting the correct ordering of synthesis systems with respect to their naturalness is more meaningful than predicting the absolute MOS values. As an example, if we use a rating scale from 1-5 and keep the rank order of MOS ratings for the audio samples the same but shift and skew the overall distribution of their scores towards either end of the scale to simulate listener and other contextual biases, then MSE will increase substantially, although ranking-based correlations will remain high. Using metrics such as MSE and linear correlation coefficient (LCC),

which are dependent on the absolute MOS values, can be misleading in evaluating different MOS predictors, especially in the zero-shot OOD setting.

In the same spirit, the authors of UTMOS [12] proposed a loss that enforces correct rank order, obtaining conclusive improvement on an OOD dataset and supporting our hypothesis. However, the authors of UTMOS in their paper did not discuss why the rank order is important nor did they investigate the performance of their loss function in zero-shot or semi-supervised settings. In contrast, we justify our loss function using the partial rank order within a mini-batch and show that MSE and LCC are unreliable metrics for evaluating MOS prediction systems. The core idea of our method is most similar to UTMOS [12]. The most notable difference is in the loss formulation. Their loss contains a margin term to avoid penalizing small errors, but which has the consequence that the loss could be zero even if the rank order is incorrect. This is an undesirable behavior when predicting MOS values. Lastly, different from prior work, we also study the effect of extending the total number of comparisons beyond the current batch size. The differences between UTMOS and the proposed method will be described in more detail in Section 2.

In this paper, we propose a method that addresses the challenging case of *zero-shot and few-shot quality MOS prediction for unseen, OOD speech synthesis systems*. Rather than focusing on absolute measures, we aim to measure *similarity of partial rank order matrices obtained from MOS values for multiple (but not necessarily all) samples and systems, particularly in terms of naturalness*. Our contributions are as follows:

1. We explain why relative position in the rank order is important to consider when solving the MOS prediction task.
2. We formulate a loss function on the basis of the relative position in the rank order that covers parts of systems to be evaluated and call it Partial Rank Similarity (*PRS*) loss.
3. We introduce a BALanced pseudo MOS (BAPMOS) selection approach for choosing unlabeled audio samples for use in semi-supervised training.
4. We empirically demonstrate the effectiveness of the proposed loss function to make quality predictions on unseen OOD speech synthesis systems in zero-shot, few-shot, and semi-supervised settings.

## 2. METHODOLOGY

In this section, we present the proposed *PRS* criterion. The method is motivated by the idea that the relative position of an audio sample in the ranking based on a partial list of training samples, which are ordered by their relative quality, is an important aspect of solving the MOS prediction task as opposed to only considering the absolute MOS value as in [7]. Therefore, before delving into the specifics of the *PRS* criterion, the concept of relative position in the ranking and partial rank matrix needs to be explored in greater detail.

### 2.1. Relative position in the ranking and partial rank matrix

Let us consider a list  $l = (l_1, l_2, l_3) = (1, 3, 2)$  where each element represents an absolute MOS value assigned to a different system. The list may not contain samples from all speech synthesis systems but a subset of them. Although the original ratings are ordinal values,

we treat the MOS values as continuous for simplicity. To represent the relative position of each value with respect to all other values in the list, we define a matrix called the partial rank matrix. This matrix stores the position of each value in the list relative to every other value and also to itself. For example, the elements of the first row of the matrix are  $l_1 - l_1 = 0$ ,  $l_1 - l_2 = -2$ , and,  $l_1 - l_3 = -1$ , respectively. By extending this idea to all rows, we can construct the partial rank matrix for all values in  $l$ , as shown in Equation 1.

$$PR(l) = \begin{bmatrix} 0 & l_1 - l_2 & l_1 - l_3 \\ l_2 - l_1 & 0 & l_2 - l_3 \\ l_3 - l_1 & l_3 - l_2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -2 & -1 \\ 2 & 0 & 1 \\ 1 & -1 & 0 \end{bmatrix} \quad (1)$$

A visual representation of the partial rank operation to the predicted and ground-truth MOS values is shown in Figure 1.

Matrix  $PR(l)$  captures two fundamental pieces of information: directionality and magnitude. The sign in the matrix indicates the directionality, allowing us to determine whether the reference value is ranked higher or lower than all other indices. The magnitude simply represents the rank order difference, indicating *how much* higher or lower each value is than the reference value. Having established a solid foundation in understanding relative position in the rank, now we discuss the key aspects of the proposed *PRS* loss and its variants.

### 2.2. The *PRS* loss function

During the training process, let us consider a batch of size  $n$  containing  $n$  input audio samples, denoted as  $\mathbf{X} = (x_1, x_2, \dots, x_n)$ , and their corresponding MOS values. Our goal is to learn a prediction function that can closely estimate the MOS scores given the input audio signals. To achieve this, we assume the existence of a non-linear function  $\Psi$  that approximates the MOS value on the basis of the provided audio, such that  $\hat{y}_i = \Psi(x_i)$ . We propose an objective function that minimizes the total losses with respect to the training data. The objective function is defined in Equation 2:

$$\mathcal{L}_{PRS} = \left( \sum_{i=1}^n \sum_{j=1}^n \lambda * |PR_{ij}(\hat{\mathbf{Y}}) - PR_{ij}(\mathbf{Y})|^p \right)^{1/p} \quad (2)$$

Where:

$$\hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n), \quad (3)$$

$$\hat{y}_i = \Psi(x_i), \quad (4)$$

$$\lambda = \begin{cases} 1 & \text{if } \{PR_{ij}(\hat{\mathbf{Y}}) \cdot PR_{ij}(\mathbf{Y})\} \leq 0, \\ \lambda_c \leq 1 & \text{otherwise,} \end{cases} \quad (5)$$

and  $\lambda_c$  is a hyper-parameter. In Equation 2, the loss represents a measure of the difference between  $ij$ -th elements of the predicted  $PR_{ij}(\hat{\mathbf{Y}})$  and the ground-truth matrices  $PR_{ij}(\mathbf{Y})$ , by utilizing a  $p$ -norm. Additionally, the weight factor  $\lambda$  allows us to control the contribution of each index pair  $(i, j)$  in the total loss calculation. In one possible use case, if the two values ( $PR_{ij}(\hat{\mathbf{Y}})$  and  $PR_{ij}(\mathbf{Y})$ ) have the same sign (either both positive or both negative) in Eq. (2), they are penalized less ( $\lambda = \lambda_c < 1$ ) than in cases where they have opposite signs. In other words, if the MOS prediction model misclassifies the relative order of the  $i$ -th and  $j$ -th samples, we penalize more.

One benefit of using the loss in Eq. (2) over the loss used by [7] (which we call L1) is that the minimization takes into consideration

other values and not just an individual value by incorporating the notion of relative positions in the ranking into the learning process. The model is explicitly encouraged to learn the correct rank order of the samples, whereas L1 regression does not consider the interaction between the samples. Furthermore, the number of comparisons (column) for each audio sample (row), in the  $\mathcal{PRS}$  matrix is not restricted by the current batch size and can be easily extended by maintaining a cache of previous MOS values. We save the output of previous batches in a dictionary to be used for comparisons with the audio samples in the current batch. This is done because of the GPU memory limit. We call this variant “Extended  $\mathcal{PRS}$ ,” (E- $\mathcal{PRS}$  for short). The proposed  $\mathcal{PRS}$  loss is a new approach for predicting the MOS of audio signals, is easy to implement, and can be used with any neural network architecture. Lastly, we also investigate the combined E- $\mathcal{PRS}$  and L1 loss as shown in Eq. (6).

$$\mathcal{L} = \alpha * \mathcal{L}_{E-\mathcal{PRS}} + \beta * \left( \sum_{i=1}^n |\hat{y}_i - y_i|^p \right)^{1/p} \quad (6)$$

Similar to  $\mathcal{PRS}$ ,  $\max(0, |PR_{ij}(\hat{\mathbf{Y}}) - PR_{ij}(\mathbf{Y})| - \gamma)$  is the loss used by the authors of UTMOS [12]. One major drawback is that their loss function does not always enforce correct rank order; i.e., even if the rank order is incorrect, the loss may be zero – that is, all values less than  $\gamma$  will be neglected. In contrast, the  $\mathcal{PRS}$  loss uses  $\lambda_c$  to penalize less if the MOS prediction model orders the ranks of the  $i$ -th and  $j$ -th samples correctly. Lastly, if the MOS values of very similar systems are not reliable (assumption), then having a margin to ignore very small values is a good choice. This can be used as a regularizer in Equation 2.

### 2.3. Pseudo MOS values selection algorithm for semi-supervised training

Assume we possess  $n$  audio samples, each associated with their respective MOS values (labeled), and  $m$  audio samples without the MOS values (unlabeled). In the semi-supervised setting, we initially train the model using the labeled samples through supervised learning. Subsequently, using the trained model, we estimate the MOS values for the unlabeled samples, which are referred to as pseudo MOS values. In the following phase, we merge the labeled and (selected) unlabeled samples and repeat the supervised learning as in the initial step. We iterate this procedure until a predefined stopping criterion is met. A straightforward selection algorithm would be to choose all unlabeled samples. One drawback is that not all the pseudo MOS values are accurate, which could destabilize the subsequent training phase. Therefore, a need arises for a better selection algorithm to pick pseudo MOS values that are likely to be correct.

In this work, we propose a simple yet effective selection algorithm. Since it is challenging to define what is correct, we propose to simply balance the pseudo MOS values and call our method BALanced pseudo MOS (BAPMOS) selection. Given  $m$  unlabeled audio samples and their corresponding pseudo MOS values  $\hat{\mathbf{Y}} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m)$ , our method operates as follows: (i) We construct a histogram with  $b$  bins (hyperparameter), each containing a count specified by  $\mathbf{C} = (c_1, c_2, \dots, c_b)$ . If the resulting distribution is imbalanced, the method is prone to over-classify the majority group due to its higher prior probability. To address this issue, (ii) we randomly sample the minimum count,  $\min(C)$ , pseudo MOS values from each bin and discard the remaining values. The total number of

**Table 1.** Summary of the different datasets used in this work.

Dataset	Lang	# Samples			# ratings per sample
		Train	Dev	Test	
<b>Stage 1</b>					
BVCC [7]	en	4,974	1,066	1,066	8
ASV2019 [7]	en	-	-	6,026	1-26
BC2019 [7]	ch	-	-	450	10-17
COM2018 [7]	ja	-	-	1,586	1-9
<b>Stage 2</b>					
BC2019 [8]	ch	Labeled: 136 Unlabeled: 540	136	540	10-17
[13]	Gitksan	-	-	25	12

selected pseudo MOS values for the iterative training is  $b * \min(C)$ . This simply ensures a balanced distribution of selected pseudo MOS values or uniform prior probability of the histogram.

## 3. EXPERIMENTS

### 3.1. Experimental Design

Two types of experiments are conducted in this paper: in Stage 1, an SSL model is first finetuned with the proposed criterion on the basis of the labeled training data. The evaluation is then performed on held-out data in the same domain as the training data. Zero-shot evaluations are also performed on three OOD sets that are not included in the training data.

In the Stage 2 experiments, we show and discuss the results of training a MOS prediction model with the proposed criterion on an OOD set, either by zero-shot, few-shot, or semi-supervised learning, and we also investigate the use of the BAPMOS selection approach in the semi-supervised setting.

### 3.2. Experimental Conditions

**Pretrained Model:** Our approach utilizes the pretrained w2v\_small model [9], which has 95 million parameters and generates 768-dimensional output embeddings from an input audio sample. This model was trained on the standard Librispeech dataset [14], which comprises 960 hours of speech data.

**Loss Function:** We use our proposed  $\mathcal{PRS}$  loss as described in Eq. (6). We perform all the experiments with  $p = 1$  and squared  $p = 2$  norm. Similar to [7], we have found that the  $p = 1$  almost always gives slightly better results. Therefore, we only report results with  $p = 1$ . Furthermore, the values of  $\lambda_c$ ,  $\alpha$  and,  $\beta$  are set to 1.0, 1.0 and, 0.0 in Eqs. (5) and (6) respectively, unless mentioned otherwise. Lastly, in the case of E- $\mathcal{PRS}$ , the contribution of the extended columns to the loss is scaled by  $1/10$ .

**Finetuning** For Stage 1, we finetune the Wav2Vec2.0 model on the BVCC [15] training set using the  $\mathcal{PRS}$  loss unless mentioned otherwise. Similarly to [7], we average the frame-level features of the last Wav2Vec2.0 layer and apply a linear regressor on top of it. The entire resulting model is then finetuned to solve the MOS prediction task using the BVCC training dataset.

We also further finetune the Stage 1 model, best weights, on different OOD datasets for Stage 2 experiments. Three different sets of finetuning loss function configurations are used:  $\mathcal{PRS} / \mathcal{PRS}$ , L1 / L1 and,  $\mathcal{PRS} / \text{L1}$ . Stage 2 finetuning consists of one of three setups: zero-shot, few-shot, or a semi-supervised scenario. In the zero-shot scenario, no finetuning is done i.e., 0 labeled and 0 unlabeled samples. In the few-shot scenario, small numbers of labeled

**Table 2.** Comparison of Stage 1 finetuned models, including prior work on the in-domain dataset.

Methods	Utterance				System			
	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑
L1 [7]	<b>0.227</b>	0.868	0.866	0.690	<b>0.121</b>	0.938	0.942	0.790
UTMOS [12]	5.870	0.869	0.866	0.687	4.810	0.948	0.951	0.806
$\mathcal{L}_{PRS}, \lambda_c = 1.0$	10.670	0.879	0.878	0.704	8.800	0.951	0.951	0.811
$\mathcal{L}_{E-PRS}, \lambda_c = 1.0$	12.320	0.881	<b>0.881</b>	0.707	10.120	0.947	0.949	0.805
$\mathcal{L}_{E-PRS}, \lambda_c = 0.1$	7.240	0.872	0.869	0.692	6.260	0.944	0.941	0.800
$\mathcal{L}_{E-PRS}, \lambda_c = 0.0$	3.490	0.602	0.862	0.684	2.320	0.643	0.920	0.760
$\mathcal{L}, \lambda_c = 1.0, \beta = 0.01$	0.307	<b>0.883</b>	<b>0.881</b>	<b>0.710</b>	0.229	<b>0.953</b>	<b>0.952</b>	<b>0.813</b>
$\mathcal{L}, \lambda_c = 0.1, \beta = 0.01$	0.490	0.874	0.871	0.700	0.490	0.937	0.938	0.790
$\mathcal{L}, \lambda_c = 0.0, \beta = 0.01$	0.300	0.820	0.880	0.700	0.200	0.874	0.940	0.792

**Table 3.** Comparison of zero-shot capabilities of  $PRS$  Stage 1 finetuned Wav2Vec2.0 model, its variants, and results from prior work on three out-of-domain datasets.

Methods	Utterance											
	ASV2019				BC2019				COM2018			
	MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE	LCC	SRCC	Ktau	MSE	LCC	SRCC	KTAU
L1 [7]	<b>1.498</b>	<b>0.470</b>	0.491	0.352	3.672	0.553	0.559	0.409	1.200	0.476	0.423	0.297
UTMOS [12]	4.610	0.462	0.479	0.342	26.990	0.658	0.684	0.489	14.750	0.463	0.431	0.307
$\mathcal{L}_{PRS}, \lambda_c = 1.0$	8.430	0.464	0.475	0.339	45.250	0.649	0.681	0.493	25.520	0.466	0.436	0.309
$\mathcal{L}_{E-PRS}, \lambda_c = 1.0$	9.010	0.464	0.479	0.342	51.800	0.635	0.654	0.464	29.090	0.502	0.463	0.331
$\mathcal{L}_{E-PRS}, \lambda_c = 0.1$	2.750	<b>0.470</b>	<b>0.499</b>	<b>0.357</b>	19.510	0.637	<b>0.686</b>	<b>0.500</b>	6.34	<b>0.515</b>	<b>0.490</b>	<b>0.350</b>
$\mathcal{L}_{E-PRS}, \lambda_c = 0.0$	4.500	0.253	0.480	0.342	38.100	0.604	0.651	0.467	2.64	0.401	0.443	0.315
$\mathcal{L}, \lambda_c = 1.0, \beta = 0.01$	1.800	0.471	0.486	0.347	<b>2.820</b>	0.646	0.663	0.472	<b>0.81</b>	0.467	0.431	0.306
$\mathcal{L}, \lambda_c = 0.1, \beta = 0.01$	2.280	0.448	0.463	0.329	3.28	0.643	0.673	0.484	0.810	0.437	0.421	0.297
$\mathcal{L}, \lambda_c = 0.0, \beta = 0.01$	1.660	0.413	0.467	0.333	2.650	<b>0.669</b>	0.664	0.480	0.740	0.442	0.416	0.295

samples are used for finetuning. In the semi-supervised setting, we generate predicted pseudo MOS values on the available unlabeled samples either using the Stage 1 or Stage 2 finetuned models. Then, we use these pseudo MOS values combined with the real scores to finetune the model further. During Stage 2 finetuning, we evaluate the model after each epoch, and if and only if the Spearman rank correlation coefficient (SRCC) metric improves on the development set, we regenerate the pseudo MOS values and continue finetuning.

**Dataset for Stage 1:** We evaluate the performance of our approach trained using the BVCC dataset, which was derived from a comprehensive listening test conducted by [15]. The dataset consists of 7,106 audio samples from 187 systems, including text-to-speech synthesis, voice conversion, and natural speech. Each sample has eight ratings, which are averaged to obtain a MOS label for that sample. Listeners rated samples on a discrete scale from 1 (very bad) to 5 (very good) in terms of naturalness. We use the same training, development, and test sets as [7], preserving a distribution of 70%/15%/15%(4,974/1,066/1,066). To assess the generalization ability of our approach, similar to [7], we also tested the BVCC-trained models on three OOD listening test datasets: ASV2019 [16] (English), BC2019 [17] (Mandarin Chinese), and COM2018 [18] (Japanese). Testing was conducted in a zero-shot manner; i.e., without any further finetuning on these three OOD datasets. This evaluation protocol allows us to examine how well the model performs on unseen OOD data that is different from the training domain.

**Dataset for Stage 2:** For the Stage 2 finetuning experiments, we adopt the OOD track dataset from the Interspeech 2022 VoiceMOS challenge [8], which is the same original data as BC2019 except with different splits: there are 136 labeled training samples and 540 audio-only unlabeled training samples for use in semi-supervised training, including an “unlabeled training” set. We also use a dataset

from [13], consisting of five samples from each of four TTS systems and natural reference speech in the Gitksan language, an Indigenous language of Canada, for testing our approach on a real low-resource language. Table 1 shows the statistics of all the datasets used in this work.

**Metrics:** Similar to [7, 8], to evaluate MOS prediction models, we employ four widely used metrics: mean squared error (MSE), linear correlation coefficient (LCC), Spearman rank correlation coefficient (SRCC), and Kendall’s Tau rank correlation (KTAU). The LCC, SRCC, and KTAU values range from -1 to 1, with values closer to 1 indicating a better correlation between predicted and ground-truth values. Among them, SRCC and KTAU are more useful metrics for our proposed loss function since MSE and LCC are dependent on absolute MOS values.

### 3.3. Stage 1 experiment: in-domain vs. out-of-domain

Table 2 shows comparison results of the Stage 1 models on the in-domain BVCC test dataset. First, we can confirm that since the predictive models using the proposed  $PRS$  loss and its variant ( $\mathcal{L}_{PRS}$  and  $\mathcal{L}_{E-PRS}$ ) do not take into account the absolute MOS values during the learning process, they naturally result in larger MSEs, but this outcome is expected. Next, comparing the values of LCC, KTAU, and SRCC, we can confirm that the correlation coefficients of the proposed methods ( $\mathcal{L}_{PRS}$ ,  $\mathcal{L}_{E-PRS}$ , and  $\mathcal{L}$ ) are comparable to or even slightly higher than those of L1 and UTMOS when appropriate  $\lambda_c$  values are utilized. Finally, the results of using a loss  $\mathcal{L}$  that also takes L1 into account at the same time naturally confirms that the MSE is also reduced. In summary, if one wants to know only the rank ordering, the proposed loss function is sufficient; if one wants to approximate the MOS values as well, L1 is necessary.

**Table 4.** Testing the  $\mathcal{PRS}$  method in zero-shot, few-shot and, semi-supervised settings on a dataset [8]. E- $\mathcal{PRS}$  with  $\lambda_c = 0.1$  configuration is used for Stage 1 and Stage 2 finetuning. The results are averaged over three runs with random seeds. The row marked with \* model is trained with the pseudo MOS values generated only once at the starting.

Number of labeled samples	Number of unlabeled samples	1st finetuning loss / 2nd finetuning loss											
		$\mathcal{PRS} / \mathcal{PRS}$				L1 / L1				$\mathcal{PRS} / \text{L1}$			
		MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
<b>Zero-shot setting</b>													
0	0	16.350	0.617	<b>0.651</b>	0.457	3.150	0.532	0.538	0.387	16.350	0.617	<b>0.651</b>	0.457
<b>Few-shot setting</b>													
10	0	13.160	0.657	0.690	0.486	0.980	0.715	0.708	0.509	0.640	0.701	<b>0.744</b>	0.542
136	0	6.960	0.873	<b>0.842</b>	0.652	0.660	0.845	0.825	0.632	0.750	0.865	<b>0.843</b>	0.652
<b>Semi-supervised setting</b>													
0*	136*	12.414	0.651	0.686	0.484	-	-	-	-	-	-	-	-
0	136	4.000	0.807	<b>0.778</b>	0.580	13.050	0.721	0.744	0.550	9.910	0.720	0.773	0.572
0	676	1.980	0.768	<b>0.778</b>	0.582	11.190	0.701	0.747	0.551	23.920	0.623	0.751	0.553
10	126	0.750	0.783	<b>0.786</b>	0.582	2.750	0.703	0.686	0.493	2.900	0.675	0.705	0.509
10	666	1.160	0.770	<b>0.782</b>	0.583	8.790	0.663	0.696	0.503	11.910	0.606	0.672	0.483
136	540	0.650	0.858	<b>0.839</b>	0.646	0.660	0.845	0.825	0.632	1.330	0.860	<b>0.840</b>	0.650

**Table 5.** Testing the  $\mathcal{PRS}$  method on Gitksan language [13], similar to Table 4. Readers must keep in mind that because only 25 samples were available, we discard the MOS values and treat them as unlabeled samples in the semi-supervised setting. However, for development and testing purposes, we use the ground-truth MOS values for comparison.

Number of labeled samples	Number of unlabeled samples	1st finetuning loss / 2nd finetuning loss											
		$\mathcal{PRS} / \mathcal{PRS}$				L1 / L1				$\mathcal{PRS} / \text{L1}$			
		MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
<b>Zero-shot setting</b>													
0	0	6.210	0.810	0.790	0.640	0.940	0.760	0.690	0.530	6.210	0.810	0.790	0.640
<b>Semi-supervised setting</b>													
0	25	5.440	0.835	0.851	0.696	4.400	0.717	0.763	0.608	1.200	0.791	0.848	0.680

Table 3 shows zero-shot comparison results of the Stage 1 models on the three OOD test datasets. First, this evaluation is done in a zero-shot manner, so naturally, the overall correlation coefficients are lower, and the MSEs are larger. We then see that the models trained with the proposed loss function have a similar level of correlation coefficients to the case trained with L1 evaluated on the OOD test sets. Some minor but consistent improvement is also observed. For instance, a system using the  $\mathcal{L}_{E-\mathcal{PRS}}$  with  $\lambda_c = 0.1$  has consistently better rank correlations (SRCC and KTAU) than L1 and UTMOS on three out of three OOD datasets. The improvement is more evident in unseen languages, that is, BC2019 and COM2018.

Finally, regarding the combined  $\mathcal{PRS}$  and L1 loss, we see a small amount of degradation concerning the rank correlations. This suggests that the two losses are not working in tandem and that minimizing the absolute values is not a good strategy for solving the MOS prediction task in the OOD setting. To summarize, E- $\mathcal{PRS}$  with  $\lambda_c = 0.1$  has the best generalization ability given its performance gains on unseen languages.

### 3.4. Stage 2 experiment: a comparison of zero-shot, few-shot, and semi-supervised settings

In the Stage 2 experiment, we analyze the performance of MOS predictors in zero-shot, few-shot, and semi-supervised settings. As explained in Section 3.2, we finetune the Stage 1 model using small amounts of labeled samples for the few-shot setting, whereas we generate pseudo MOS labels for unlabeled training audio samples and finetune a model by mixing the labeled samples and pseudo la-

beled ones for the semi-supervised setting.

Table 4 shows results on the BC2019 dataset [8]. First, we see that both few-shot and semi-supervised learning improved correlation coefficients. This is true even for semi-supervised cases where no labeled samples are used. As for the combinations of the losses used for the first and second finetuning, we see that the models using the  $\mathcal{PRS}$  loss for the first finetuning generally resulted in higher rank correlation coefficients after the second finetuning. This trend can be clearly seen from the SRCC values in the table. This demonstrates the generalization ability of the  $\mathcal{PRS}$  loss. Interestingly, the semi-supervised setting with the  $\mathcal{PRS} / \mathcal{PRS}$  condition has smaller MSE values as well. The next observation is that increasing the unlabeled data for semi-supervised learning (136 to 676 samples and 126 to 666 samples) does not result in any performance gains. This could be attributed to using all unlabeled samples with their pseudo MOS values during finetuning. Lastly, the empirical results show that iteratively regenerating the pseudo MOS values is necessary and is more accurate than if the pseudo MOS values are generated only once at the starting as shown in Table 4, in the row marked with \*.

Semi-supervised learning is particularly helpful for low-resource language scenarios since it is not straightforward to find native listeners. We therefore additionally analyzed the performance of MOS predictors in zero-shot and semi-supervised settings on a MOS dataset in the Gitksan language [13]. Table 5 shows results of the zero-shot and semi-supervised inference on the MOS dataset for the Gitksan language. We can see the same trend – the semi-supervised learning without using any labeled samples improved the prediction performance, and the  $\mathcal{PRS} / \mathcal{PRS}$  condition resulted in the highest

**Table 6.** Testing the BApMOS selection algorithm for  $\mathcal{PRS}/\mathcal{PRS}$  configurations, similar to Table 4. Here the SRCC metric was used to compare the performance.

Number of labeled samples	Number of unlabeled samples		Number of bins for a histogram			
			5	10	20	30
<b>Few-shot setting</b>						
136	0	<b>0.842</b>	-	-	-	-
<b>Semi-supervised setting</b>						
0	136	0.778	-	-	-	-
0	676	0.778	-	-	-	-
<b>Semi-supervised setting + BApMOS selection</b>						
0	136		<b>0.804</b>	<b>0.800</b>	0.800	-
0	676		0.780	0.797	<b>0.809</b>	0.799

rank correlation coefficients.

### 3.5. Stage 2 experiment: a comparison of semi-supervised learning using the BApMOS selection strategy

Next, we compare semi-supervised learning with and without the proposed BApMOS selection algorithm on the BC2019 dataset as shown in Table 6. We only report the SRCC values.

First, by comparing the results of semi-supervised learning on 136 samples with and without the BApMOS selection algorithm, we can see that the proposed BApMOS selection algorithm works effectively. It considerably boosts the performance over simply using all of the pseudo labels. As expected, it is not as good as few-shot learning which uses the ground-truth labels. Furthermore, increasing the number of unlabeled samples from 136 to 676 results in a slight performance gain, which was not the case earlier. This again proves the importance of a selection algorithm rather than just using all the unlabeled samples.

Furthermore, we make two observations: (i) diversity of selected pseudo-MOS values is detrimental to the performance of the  $\mathcal{PRS}$  method. When using 676 unlabeled samples, increasing the number of bins boosts the performance significantly. (ii) The total number of selected samples is more important than diversity if there are very few selected samples, as in the case of 30 and 20 bins in 676 and 136 unlabeled samples, respectively. Since this is a promising result, we hope that using better selection methods will result in additional performance gains, as shown by the success of semi-supervised learning methods in the past [19]. We leave this for future work.

## 4. DISCUSSION

The MOS test is affected by not only the quality of the speech, but also by the various contexts during the listening test, which cause MOS values to fluctuate. The need to model the influence of this context is an important decision regarding automatic MOS prediction.

If we believe that the variation in MOS values also needs to be modeled in the current target context, then we will need to use the MOS values as supervised labels for training. However, since this policy learns a context-dependent model, it is not expected to generalize to test sets in different contexts.

On the other hand, we found that the MOS prediction model generalizes better to the OOD test set when the learning criterion is

based on the rank order of the systems, rather than using context-dependent MOS values directly as the learning target. Although the context of that OOD test set cannot be properly considered in a zero-shot manner, we show that some context information can be captured by semi-supervised learning if unlabeled speech data is available.

The semi-supervised learning proposed in this paper has room for improvement. Specifically, the proposed semi-supervised learning used unlabeled speech data for training, but the development set still contains labeled speech. By using unlabeled speech data even in the development set, the semi-supervised learning of the MOS prediction model will be more useful. Lastly, as of now, we have not selected the samples based on any criterion other than simply making the prior of the histogram uniform. We randomly select the samples from the bin, but if instead a heuristic is used, that could lead to further improvements. One possible heuristic is to drop any sample whose relative pseudo MOS value is higher than the natural speech. There could be more ways to do selection but we leave this for future work.

## 5. CONCLUSIONS AND FUTURE WORK

This paper introduced the  $\mathcal{PRS}$  method for predicting Mean Opinion Score (MOS) values given an audio sample. By considering the relative position in the ranking of MOS values within each training batch,  $\mathcal{PRS}$  provides a novel approach to capture ranking information. In this study, we also present  $E\text{-}\mathcal{PRS}$ , an extension of  $\mathcal{PRS}$  to incorporate samples for comparison beyond the current batch size for better generalization. Comparative evaluations with existing methods demonstrated comparable performance on in-domain and superior performance on OOD datasets. The experimental results highlight the generalization ability of  $\mathcal{PRS}$  in MOS prediction tasks. Contrary to popular belief, we posit that MSE and LCC are unreliable evaluation metrics for comparing MOS prediction systems. We also demonstrated that performance can be further improved if a better selection method is used in the semi-supervised finetuning stage, similar to [19].

Our future work includes the following ideas. Instead of averaging the features from the last layer of Wav2Vec2.0, using a recurrent neural network (RNN) as the last layer during finetuning as proposed by [12] may also improve the performance. We will also consider investigating the use of attention to average the frame-level features. Furthermore, similar to [10], additional unsupervised domain-adaptive pre-training of the Wav2Vec2.0 model to learn better features may result in performance improvements in the zero-shot and few-shot settings. Lastly, a better selection algorithm to sample the pseudo MOS values in the semi-supervised setting on the basis of some heuristics could lead to improvements as well. We also intend to explore whether employing an ensemble of MOS models enhances the reliability of predictions, resembling the MOS test conducted with multiple human annotators.

## 6. ACKNOWLEDGMENTS

The authors thank Xin Wang for his valuable feedback, and Aidan Pine for making his listening test data available for this work. This study is supported by JST CREST Grant Number JPMJCR18A6 and by MEXT KAKENHI grant 21K11951. RR Shah is partly supported by the CAI and CDNM at IIIT Delhi, India. Hemant Yadav is supported by Microsoft Research India PhD Fellowship program.

## 7. REFERENCES

- [1] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, July 2002, pp. 311–318, Association for Computational Linguistics.
- [2] Sabine Buchholz, Javier Latorre, and Kayoko Yanagisawa, “Crowdsourced assessment of speech synthesis,” in *Crowdsourcing for Speech Processing: Applications to Data, Collection, Transcription and Assessment*, Maxine Eskénazi, Gina-Anne Levow, Helen Meng, Gabriel Parent, and David Suen-dermann, Eds., chapter 7, pp. 173–214. John Wiley & Sons, Ltd, Chichester, 2013.
- [3] Chen-Chou 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 2019*, 2019, pp. 1541–1545.
- [4] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin, “MBNET: MOS prediction for synthesized speech with mean-bias network,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 391–395.
- [5] Jennifer Williams, Joanna Rownicka, Pilar Oplustil, and Simon King, “Comparison of speech representations for automatic quality estimation in multi-speaker text-to-speech synthesis,” *Speaker Odyssey*, 2020.
- [6] Slawomir Zielinski, Francis Rumsey, and Søren Bech, “On some biases encountered in modern audio quality listening tests—a review,” *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.
- [7] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi 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.
- [8] Wen Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, “The VoiceMOS Challenge 2022,” in *Proc. Interspeech 2022*, 2022, pp. 4536–4540.
- [9] Alexei Baeviski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [10] Wei-Cheng Tseng, Wei-Tsung Kao, and Hung 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.
- [11] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith, “Don’t stop pretraining: Adapt language models to domains and tasks,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020, pp. 8342–8360, Association for Computational Linguistics.
- [12] Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari, “UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022,” in *Proc. Interspeech 2022*, 2022, pp. 4521–4525.
- [13] Aidan Pine, “Low resource speech synthesis,” M.S. thesis, University of Edinburgh.
- [14] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: an asr corpus based on public domain audio books,” in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.
- [15] E. Cooper and J. Yamagishi, “How do voices from past speech synthesis challenges compare today?,” in *Proc. SSW*, 2021, pp. 183–188.
- [16] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, Lauri Juvela, Paavo Alku, Yu-Huai Peng, Hsin-Te Hwang, Yu Tsao, Hsin-Min Wang, Sébastien Le Maguer, Markus Becker, Fergus Henderson, Rob Clark, Yu Zhang, Quan Wang, Ye Jia, Kai Onuma, Koji Mushika, Takashi Kaneda, Yuan Jiang, Li-Juan Liu, Yi-Chiao Wu, Wen-Chin Huang, Tomoki Toda, Kou Tanaka, Hirokazu Kameoka, Ingmar Steiner, Driss Matrouf, Jean-François Bonastre, Avashna Govender, Srikanth Ronanki, Jing-Xuan Zhang, and Zhen-Hua Ling, “ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech,” *Computer Speech & Language*, p. 101114, 2020.
- [17] Zhizheng Wu, Zhihang Xie, and Simon King, “The blizzard challenge 2019,” in *Proc. Blizzard Challenge Workshop*, 2019, vol. 2019.
- [18] Xin Wang, Jaime Lorenzo-Trueba, Shinji Takaki, Lauri Juvela, and Junichi Yamagishi, “A comparison of recent waveform generation and acoustic modeling methods for neural-network-based speech synthesis,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4804–4808.
- [19] Félix de Chaumont Quitry, Asa Oines, Pedro Moreno, and Eugene Weinstein, “High quality agreement-based semi-supervised training data for acoustic modeling,” in *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2016, pp. 592–596.