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

Mean Opinion Score (MOS): It is a Human-Centric Evaluation metric which relies on subjective human judgments, offering a nuanced comparison of different TTS systems based on perceived quality.

Motivation to automate:

- **Efficiency**: Faster evaluation time.
- **Scalability**: Large number of TTS systems.
- **Consistency**:Objective measure.
- **Cost reduction**: Expensive to hire humans.

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

Proposed approach

- **Ranking and PR Matrix**: Evaluate system positions using a list, like **L** = (1, 3, 2), with each value representing an absolute MOS.
- **PR(L) Matrix Insights**: Matrix PR(L) offers crucial details - sign for directionality (higher or lower) and magnitude for rank order differences in the relative position.

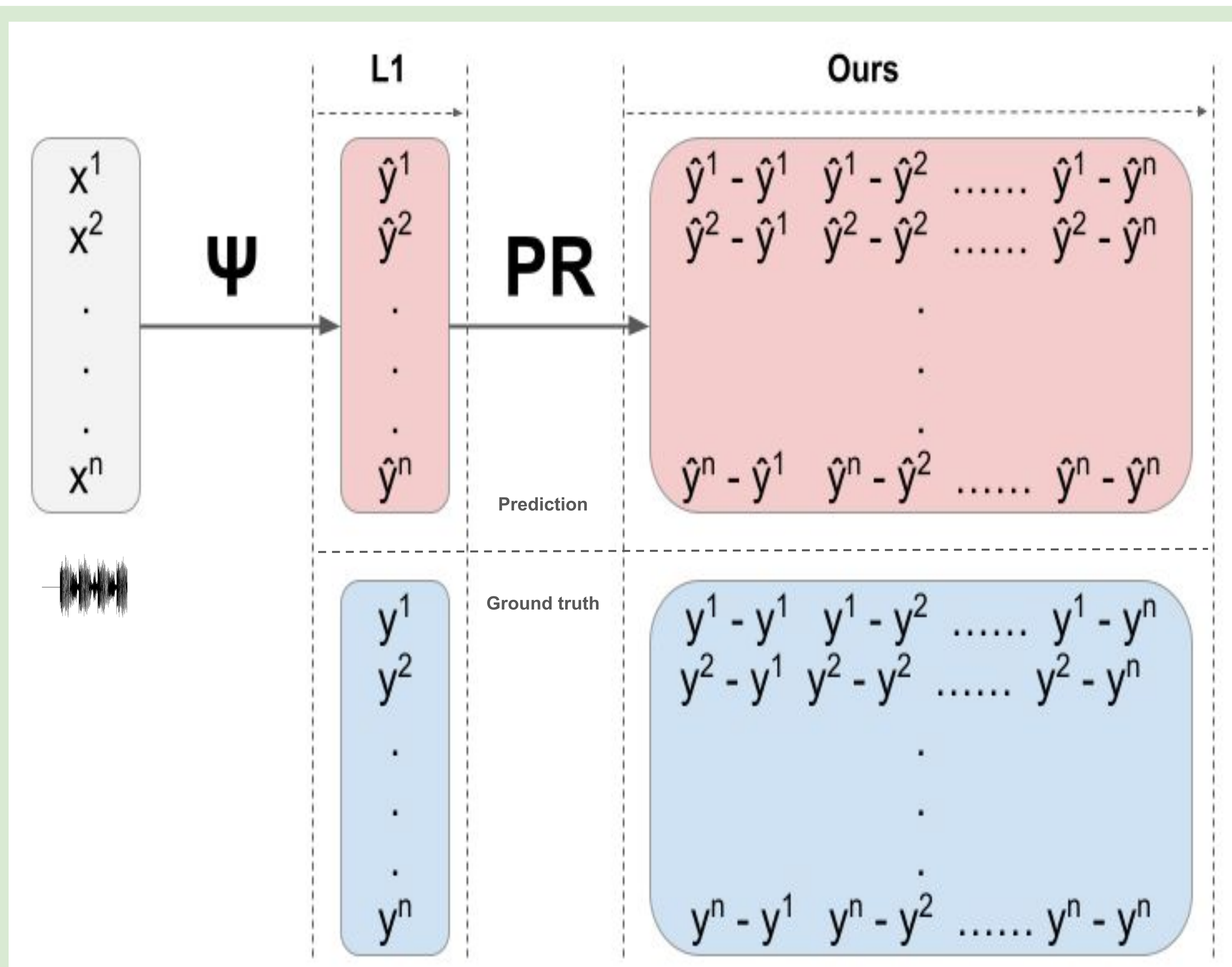

Table 4. Testing the $\mathcal{PRS}$ method in zero-shot, few-shot and, semi-supervised settings on a dataset [8]. E-$\mathcal{PRS}$ with $\lambda_c = 0.1$ configuration is used for Stage 1 and Stage 2 finetuning. The results are averaged over three runs with random seeds. The row marked with * model 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				$\mathcal{PRS}$ / L1			
		MSE ↓	LCC ↑	SRCC ↑	KTAU ↑	MSE	LCC	SRCC	KTAU	MSE	LCC	SRCC	KTAU
Zero-shot setting													
0	0	16.350	0.617	**0.651**	0.457	3.150	0.532	0.538	0.387	16.350	0.617	**0.651**	0.457
Few-shot setting													
10	0	13.160	0.657	0.690	0.486	0.980	0.715	0.708	0.509	0.640	0.701	**0.744**	0.542
136	0	6.960	0.873	**0.842**	0.652	0.660	0.845	0.825	0.632	0.750	0.865	**0.843**	0.652
Semi-supervised setting													
0*	136*	12.414	0.651	0.686	0.484	-	-	-	-	-	-	-	-
0	136	4.000	0.807	**0.778**	0.580	13.050	0.721	0.744	0.550	9.910	0.720	0.773	0.572
0	676	1.980	0.768	**0.778**	0.582	11.190	0.701	0.747	0.551	23.920	0.623	0.751	0.553
10	126	0.750	0.783	**0.786**	0.582	2.750	0.703	0.686	0.493	2.900	0.675	0.705	0.509
10	666	1.160	0.770	**0.782**	0.583	8.790	0.663	0.696	0.503	11.910	0.606	0.672	0.483
136	540	0.650	0.858	**0.839**	0.646	0.660	0.845	0.825	0.632	1.330	0.860	**0.840**	0.650

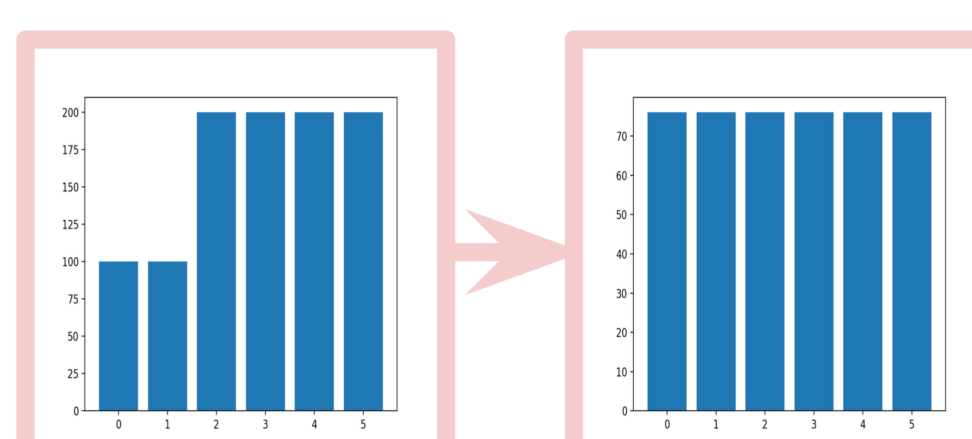

Table 6. Testing the BApMOS selection algorithm for $\mathcal{PRS}/\mathcal{PRS}$ configurations, similar to Table 4. Here the SRCC metric was used to compare the performance.

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

BApMOS

Ensures a balanced distribution of selected pseudo MOS values or uniform prior probability of the histogram

Conclusion

- Novel MOS Prediction Method: Introduces the PRS method, a unique approach for capturing ranking information.
- MSE and LCC Evaluation Challenge: Questions the reliability of MSE and LCC as metrics for comparing MOS prediction systems.
- Semi-Supervised Fine-Tuning Enhancement: Highlights potential performance improvement through better selection methods in the semi-supervised fine-tuning

Hemant Yadav: hemantya@iiitd.ac.in

Erica Cooper: ecooper@nii.ac.jp

Junichi Yamagishi: jyamagis@nii.ac.jp