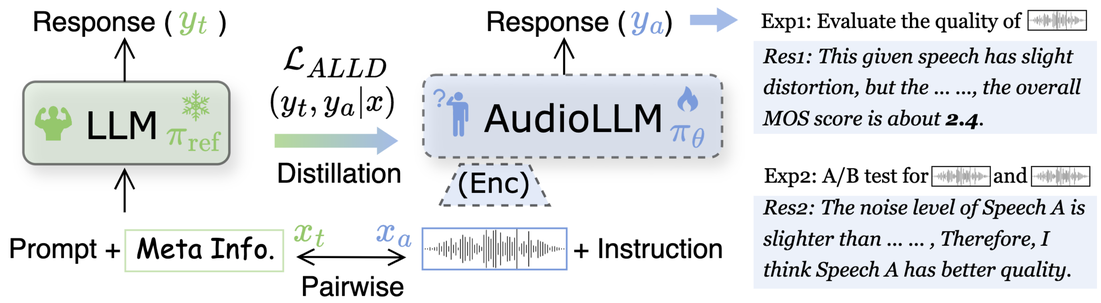
Audio Large Language Models Can Be Descriptive Speech Quality Evaluators

Chen Chen[†,1,2] Yuchen Hu[1] Siyin Wang[3] Helin Wang[4] Zhehuai Chen[2] Chao Zhang[†,3] Chao-Han Huck Yang[†,2] Eng Siong Chng[†,1]
[1]NVIDIA [2]Nanyang Technological University [2]Tsinghua University [1]Johns Hopkins University

Overview

- Audio LLMs to perceive the quality of input speech for multimodal agents
- The first descriptive speech quality evaluation dataset is introduced to bridge the speech evaluation gap for audio LLMs
- By employing token-level distillation, proposed post-training alignment approach effectively mitigates the language capability degradation of audio LLM.

DPO based Audio LLM Distillation



Speech-Text Alignment with LLM distillation (ALLD): aims to align the audio LLM response y_a to y_t via token-level distillation, where π_{ref} is exceptionally set as an expert LLM.

Mean Opinion Score (MOS) Prediction: alignment objective of π_θ with a learned reward function $r_\phi(x, y)$ as:

$$\max_{\pi_\theta} \mathbb{E}_{(x_a,x_t)\sim\mathcal{D}, y\sim\pi_\theta(y|x_a)}[r_\phi(x_a, y)] - \beta D_{\text{KL}}(\pi_\theta(y|x_a) \parallel \pi_{\text{ref}}(y|x_t)) \tag{1}$$

where M is the set of indices of the masked audio frame, $y_i^{(k)}$ is the cluster assignment derived the k-th layer of teacher model through online k-means.

DPO token-level distillation: After sampling enough y_a, a dataset of comparisons $\mathcal{D} = \{x_a^{(i)}, x_t^{(i)}, y_a^{(i)}, y_t^{(i)}\}_{i=1}^N$ is composed for preference optimization, and the training objective can be re-written as:

$$\mathcal{L}_{\text{ALLD}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x,y_a,y_t)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_t|x)}{\pi_{\text{ref}}(y_t|x)} - \beta\log\frac{\pi_\theta(y_a|x)}{\pi_{\text{ref}}(y_a|x)}\right)\right] \tag{2}$$

where y_t and y_a are formulated as preferred-dispreferred completions for Bradley-Terry model.

Reference

[1] Rafailov et al. DPO: Your LM is secretly a reward model. *NeurIPS*, 2024.
[2] Wang et al. Enabling auditory LLMs for automatic speech quality evaluation. *ICASSP*, 2025.

Dataset page →

Results on LLM based Speech Quality Prediction

- MOS prediction results with LCC, SRCC, MSE, and BLEU for audio LLMs.

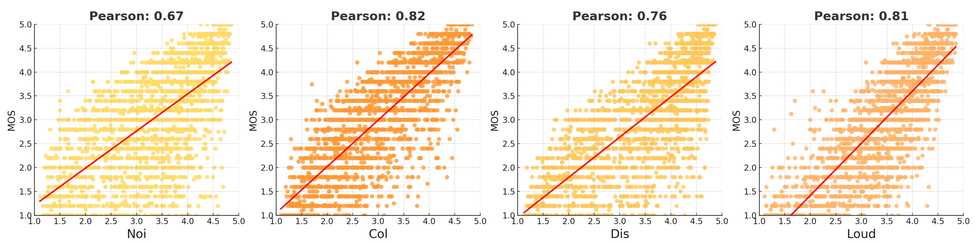
ID	Model	Tuning Manner	LCC ↑	SRCC ↑	MSE ↓	BLEU ↑
		Regression Model w/o Description				
1	CNN-SA-AP	Full-ft	0.90	0.89	0.23	*N.A.*
2	WavLM		0.90	0.90	0.24	
3	Wav2vec2		**0.93**	0.92	0.27	
		Audio LLMs w/ Description				
4	SALMONN-7B	Q-former + LoRA	0.87	0.87	0.34	25.49
5	SALMONN-13B		0.87	0.87	0.33	25.07
6	Qwen-Audio	Enc + Proj.	0.88	0.87	0.26	23.52
7	Qwen2-Audio	IA3	0.25	0.24	1.45	16.79
8		LoRA (Enc & Dec)	0.75	0.74	0.52	18.80
9		Enc-only	0.89	0.89	0.24	23.41
10		Dec-only	0.76	0.75	0.55	19.62
11		Full-ft	0.91	0.90	0.21	23.84
12		**ALLD**	0.92	0.92	0.20	25.22
13		**ALLD** (2×)	**0.93**	**0.93**	**0.17**	**25.84**

- Unseen speech domains: sub-numbers between brackets represent +/- *in-domain* performance.

Unseen Speech Domains	Model	LCC↑	SRCC↑	MSE↓	BLEU↑
LIVE: *Phone; Skype recording*	Wav2vec2	$0.86_{(-0.07)}$	$0.86_{(-0.06)}$	$0.14_{(-0.13)}$	-
	ALLD	$0.86_{(-0.06)}$	$0.86_{(-0.06)}$	$0.14_{(-0.06)}$	$26.62_{(+1.40)}$
FOR: *forensic speech dataset*	Wav2vec2	$0.93_{(-0.00)}$	$0.92_{(-0.00)}$	$0.13_{(-0.14)}$	-
	ALLD	$0.94_{(+0.02)}$	$0.93_{(+0.01)}$	$0.10_{(-0.10)}$	$25.98_{(+0.76)}$
P501: *Annex C files from P.501*	Wav2vec2	$0.94_{(+0.01)}$	$0.94_{(+0.02)}$	$0.43_{(+0.16)}$	-
	ALLD	$0.92_{(-0.00)}$	$0.92_{(-0.00)}$	$0.19_{(-0.01)}$	$27.23_{(+2.01)}$

Results on Statistical Refutation of ALLD

- The relationship between MOS and four sub-dimensions: ***Noi**siness*, ***Col**oration*, ***Dis**continuity*, and ***Loud**ness*.
- The red line represents the linear regression line fitted to the data points, showing the linear trend between each metric and MOS.



- Generation Prompt template for LLaMA-3.1 70B is shown as follows: *I will give you a tuple of meta information for speech quality evaluation, it contains 5 factors are rating from 1 to 5. For all these factors, higher is better. input is {demonstration data point}, then you should output: {customized response}*