Towards Post-training Alignment for Large Speech Models

Speaker: Hu Yuchen (NTU 4th-year Ph.D. Student)

Date: Sep 30, 2024

Background: Large Speech Models

Whisper | SeamlessM4T | Canary

VALL-E | VoiceCraft | CosyVoice

SpeechGPT | Qwen-Audio | SALMONN

How to **efficiently** align them to specific scenarios?

Background

Pretraining & SFT:

$$\mathcal{L}(x, y) = \sum_{l=1}^{L} -\log \mathcal{P}_{\theta}(y_l | y_{l-1}, \cdots, y_1, x)$$

Learn a distribution --> uncertain

Post-training:

Sampling → Annotating → Learning

↑

Knowledge

Outline

Speech-to-Text Understanding

- Motivation
- Method
- Results on ASR & AST
- Conclusion & Discussion

Text-to-Speech Synthesis

- Motivation
- Method: UNO
- Experimental Results
- Extension: RIO
- Conclusion & Discussion

Self-Taught Recognizer: Toward Unsupervised Adaptation for Speech Foundation Models

Yuchen Hu[1*], Chen Chen[1*], Chao-Han Huck Yang[2], Chengwei Qin[1],
Pin-Yu Chen[3], Eng Siong Chng[1], Chao Zhang[4]

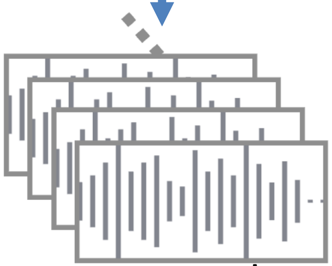
[1] Nanyang Technological University [2] NVIDIA Research
[3] IBM Research [4] Tsinghua University

Motivation

To deploy an ASR system in a practical scenario:

(NTU Canteen)

Collect



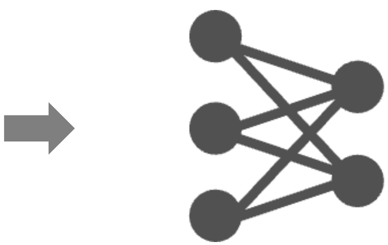
Speech

Manual Labeling

+

Transcription

Supervised Domain Adaptation

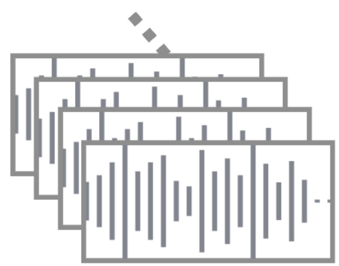


Neural Model

A very convenient approach is:

Whisper

+



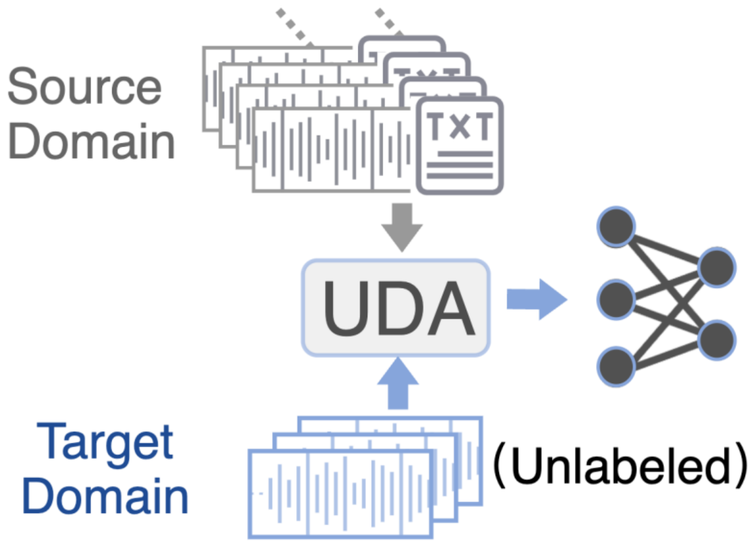
Unlabeled Speech

Unsupervised Domain Adaptation

UDA

Motivation

UDA in ASR:



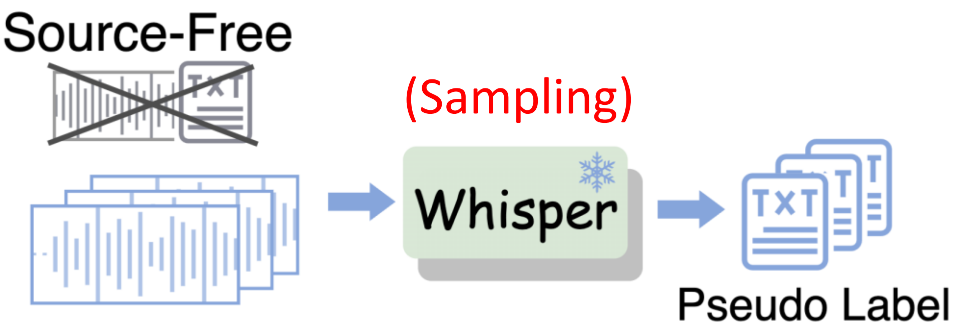
Human's UDA solution:

- "Unsupervised " is for adaptation process, but the learning schedule is semi-supervised.

- Considering the exhibited ability of large speech model:
  Can we skip the source-domain data for target domain adaptation? ➔ Source-free UDA

Method (Self-training --> Post-training)

1) Pseudo Labeling:

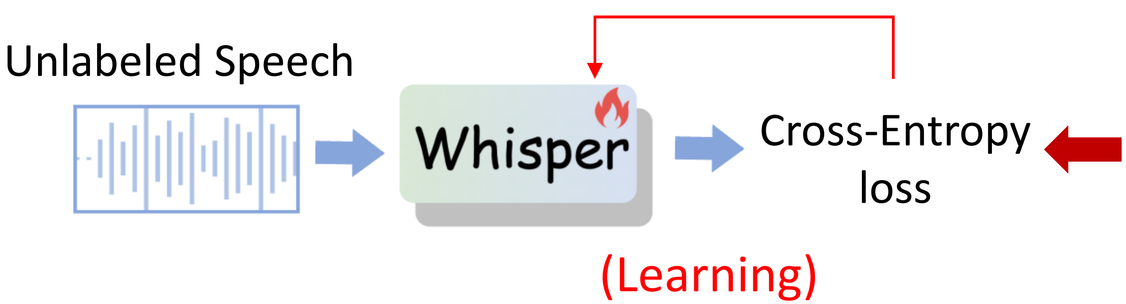


Keep this utterance or **discard**?

--> Monte Carlo sampling

2) Informed Finetuning:

How to assign weight for each token?

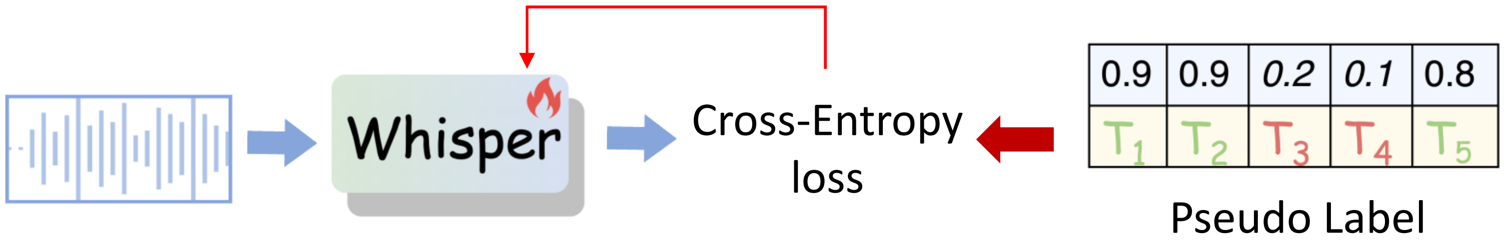


(Annotating)

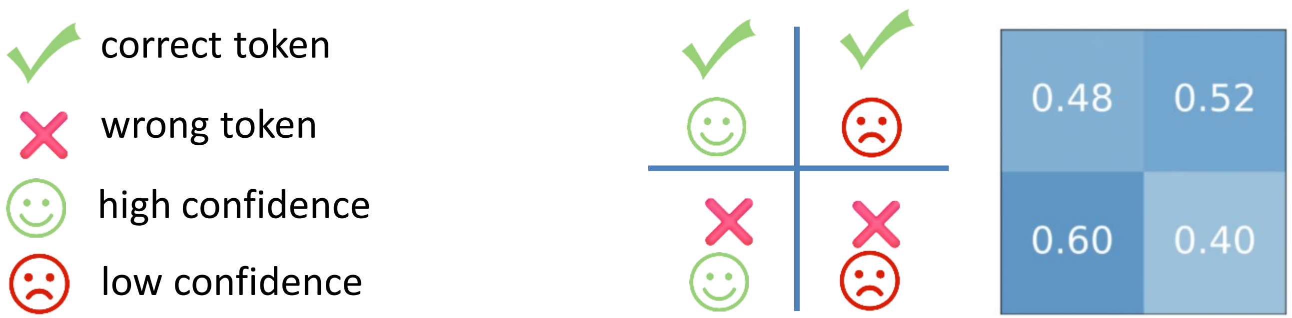
0.9	0.9	*0.2*	*0.1*	0.8
T_1	T_2	T_3	T_4	T_5

Pseudo Label

Candidate 1: Confidence Score



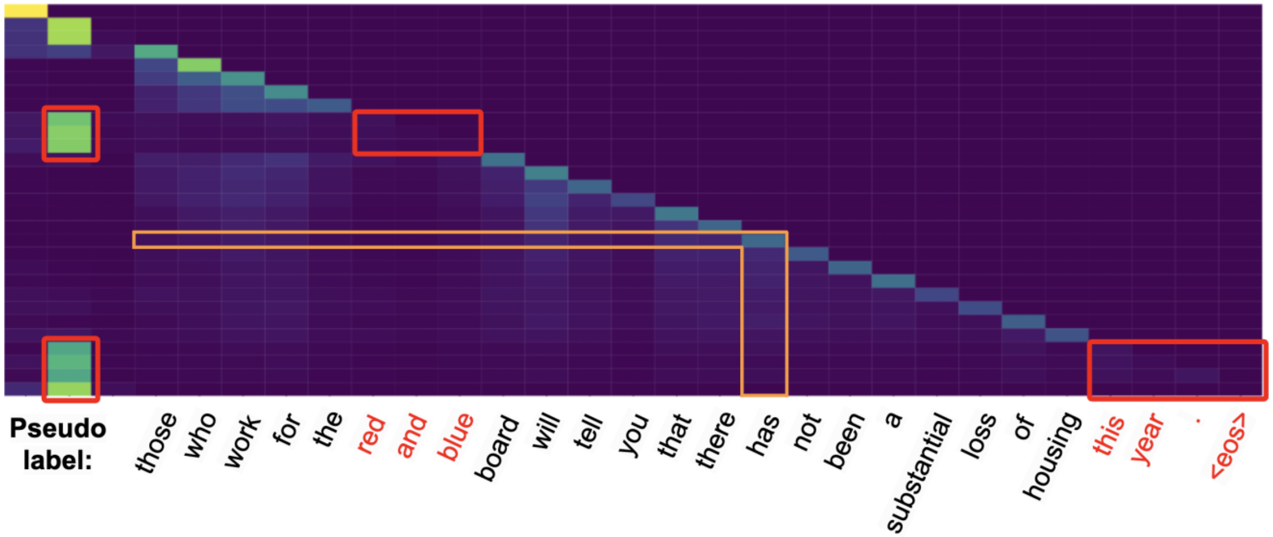
Experimental observation: decoding performance on CHiME-4 test-real



Confidence score is unreliable!

Candidate 2: Self-Attention Matrix

<|transcribe|>



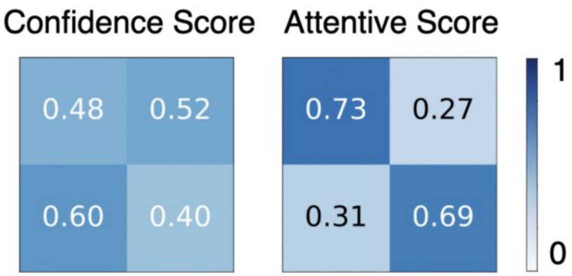
Attentive score:

$$A_l = \sum_{j=4}^{l} W_{l,j} + \sum_{i=l+1}^{L} W_{i,l},$$

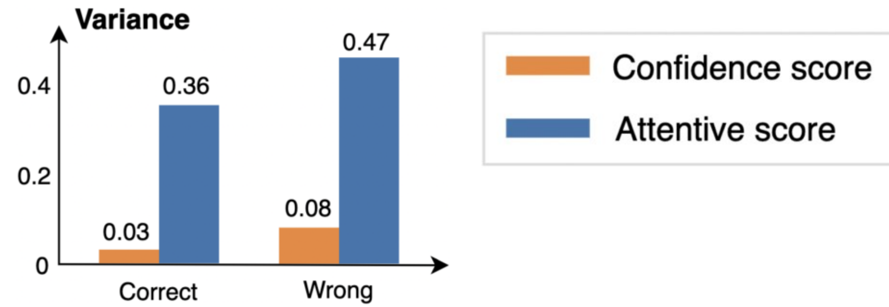
The importance of l-th token in whole utterance[8]

Correct/Wrong

*Is A_l more **reliable** than C_l ?*



Is A_l stable for guide finetuning?



Conclusion: attentive score is **more reliable** but **less stable** than confidence score.

STAR: Integrate *A* and *C* for each token

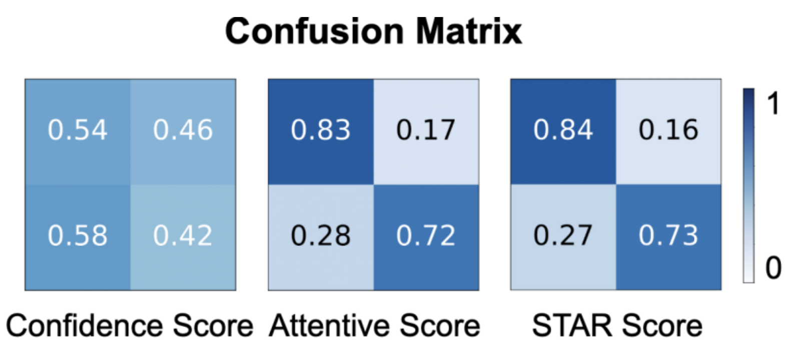
Criteria: - If *A*-*C* conflict, then follow *A:*

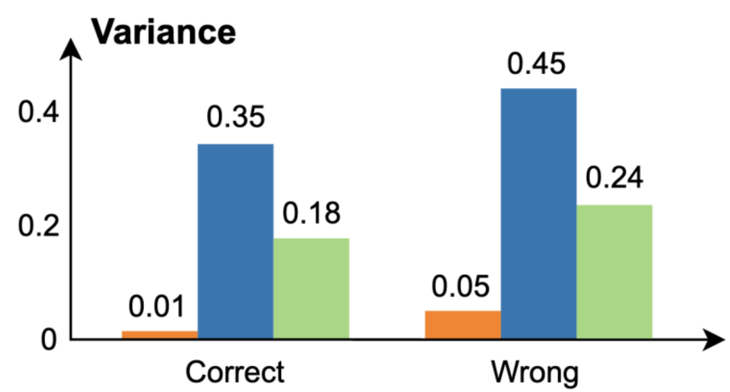
$$\mathcal{S}_l^{\text{conf}} = [\sigma(\mathcal{A}_l^2/\mathcal{C}_l - \lambda) + \sigma(\mathcal{C}_l^2/\mathcal{A}_l - \lambda)] * \mathcal{A}_l$$

- If *A*-*C* consistent, then calibrate *A* using *C:*

$$\mathcal{S}_l^{\text{cons}} = [\sigma(\lambda - \mathcal{A}_l^2/\mathcal{C}_l) * \sigma(\lambda - \mathcal{C}_l^2/\mathcal{A}_l)] * \mathcal{A}_l * e^{(\mathcal{C}_l - \mathcal{A}_l)/\tau}.$$

Quick validation:





Effectiveness on Various Domains

STAR = Self-TAught Recognizer

Testing Scenario		Whisper (frozen)	Whisper (self-train.)	$\text{UTT}_{\text{filter}}$	$\text{TOK}_{\text{reweight}}$ $\mathcal{C}_l$	$\mathcal{A}_l$	**STAR (ours)**	Whisper (real label)
			Background Noise					
CHiME-4	*test-real*	6.8	6.9	6.4	6.5	6.2	$\mathbf{6.0}_{-11.8\%}$	5.2
	test-simu	9.9	10.1	9.7	9.8	9.5	$\mathbf{9.4}_{-5.1\%}$	8.7
	dev-real	4.6	4.5	4.3	4.3	4.1	$\mathbf{3.9}_{-15.2\%}$	3.2
	dev-simu	7.0	7.0	6.6	6.7	6.6	$\mathbf{6.4}_{-8.6\%}$	5.9
LS-FreeSound	*babble*	40.2	37.6	35.0	33.5	31.3	$\mathbf{30.2}_{-24.9\%}$	27.2
	airport	15.6	15.5	15.2	15.3	15.0	$\mathbf{14.8}_{-5.1\%}$	14.5
	car	2.9	3.0	2.8	2.8	2.6	$\mathbf{2.5}_{-13.8\%}$	2.4
RATS	*radio*	46.9	47.2	46.0	45.5	44.9	$\mathbf{44.6}_{-4.9\%}$	38.6
			Speaker Accents					
CommonVoice	*African*	6.0	5.8	5.5	5.4	5.0	$\mathbf{4.8}_{-20.0\%}$	4.6
	Australian	5.8	5.7	5.6	5.5	5.2	$\mathbf{5.1}_{-12.1\%}$	4.3
	Indian	6.6	6.5	6.3	6.4	6.1	$\mathbf{6.0}_{-9.1\%}$	5.7
	Singaporean	6.5	6.2	5.8	5.8	5.4	$\mathbf{5.1}_{-21.5\%}$	4.9
			Specific Scenarios					
TED-LIUM 3	*TED talks*	5.2	4.9	4.7	4.8	4.3	$\mathbf{4.1}_{-21.2\%}$	3.6
SwitchBoard	*telephone*	20.8	20.5	19.8	19.3	18.6	$\mathbf{18.1}_{-13.0\%}$	15.3
LRS2	*BBC talks*	8.5	8.3	7.6	7.9	7.4	$\mathbf{7.0}_{-17.6\%}$	5.6
ATIS	*airline info.*	3.6	3.5	3.3	3.3	3.2	$\mathbf{2.9}_{-19.4\%}$	2.0
CORAAL	*interview*	21.5	21.3	20.8	20.7	20.4	$\mathbf{20.1}_{-6.5\%}$	17.9

Whisper zero-shot

Previous Semi-ASR

Ours

Real-label training

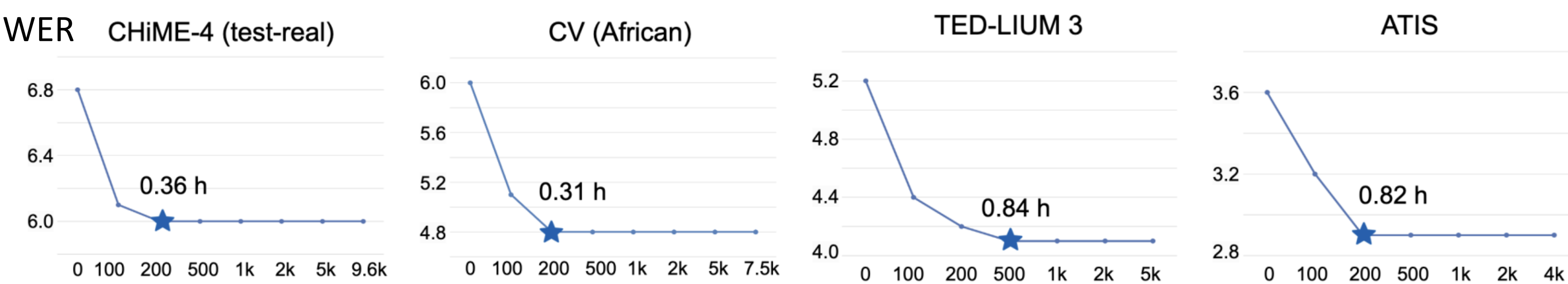
Analysis

STAR can avoid forgetting:

Model	LS-FreeSound			RATS	CommonVoice				TED-3	SWBD	ATIS
	babble	*airport*	*car*		*af*	*au*	*in*	*sg*			
Frozen	40.2	**15.6**	2.9	46.9	**6.0**	**5.8**	**6.6**	6.5	5.2	**13.3**	3.6
Self-train.	38.2	16.6	2.9	47.3	6.4	5.9	6.7	6.3	5.3	13.7	3.4
STAR	**33.3**	15.7	**2.8**	**46.1**	6.1	**5.8**	6.7	**5.6**	**5.0**	13.5	**2.9**

Train on CHiME-4 and test on OOD

STAR enjoys high data efficiency:



train samples

Generalization

- Other models

Model	Baseline	Self-train.	STAR	Real
Whisper-V3-1.5B	6.8	6.9	$6.0_{-11.8\%}$	5.2
Whisper-Med-0.8B	8.9	8.8	$8.0_{-10.1\%}$	7.1
OWSM-V3.1-1.0B	8.4	8.1	$7.5_{-10.7\%}$	6.5
Canary-1.0B	8.2	8.0	$7.2_{-12.2\%}$	6.4
Parakeet-TDT-1.1B	8.0	7.8	$7.0_{-12.5\%}$	6.2

- Other task (Speech Translation on FLEURS)

$X \rightarrow En$	Baseline	Self-train.	STAR	Real
Ar	21.9	22.1	$23.3_{+1.4}$	24.5
De	33.7	34.0	$35.9_{+2.2}$	36.5
Es	23.9	24.1	$24.8_{+0.9}$	26.4
Fa	16.6	16.3	$17.6_{+1.0}$	19.0
Hi	22.4	22.5	$23.4_{+1.0}$	24.4
Zh	16.3	16.3	$17.1_{+0.8}$	17.9

Ablation Study

- Different whisper sizes

Model Size	# Param.	Baseline	STAR	Real
large-v3	1,550 M	6.8	$6.0_{-11.8\%}$	5.2
large-v2		7.7	$6.9_{-10.4\%}$	6.0
large		7.5	$7.0_{-6.7\%}$	6.8
medium.en	769 M	8.9	$8.0_{-10.1\%}$	7.1
small.en	244 M	12.7	$10.6_{-16.5\%}$	9.0
base.en	74 M	32.4	$17.7_{-45.4\%}$	16.1

- Different training methods

Approach	# Param.*	Baseline	STAR	Real
Regular Finetuning				
Full	1550 M	6.8	$6.0_{-11.8\%}$	5.2
Enc-only	635 M		$6.3_{-7.4\%}$	5.0
Dec-only	907 M		$6.1_{-10.3\%}$	4.4
Parameter-Efficient Finetuning				
LoRA	16 M	6.8	$6.0_{-11.8\%}$	5.1
Reprogram.	0.4 M		$6.7_{-1.5\%}$	6.7

Iterative Post-training

Model	Test set	# Iterations 0	1	2	3	4	5	Real label
large-v3	*test-real*	6.8	6.0	5.9	5.7	5.7	5.7	5.2
medium.en	*test-real*	8.9	8.0	7.9	7.9	7.8	7.8	7.1
small.en	*test-real*	12.7	10.6	10.3	10.3	10.3	10.3	9.0
base.en	*test-real*	34.4	17.7	17.2	17.2	17.0	17.0	16.1
large-v3	*test-simu*	9.9	9.4	9.3	9.0	8.9	8.9	8.7
large-v3	*dev-real*	4.6	3.9	3.9	3.8	3.8	3.8	3.2
large-v3	*dev-simu*	7.0	6.4	6.4	6.4	6.3	6.3	5.9
large-v3	*af*	6.0	4.8	4.8	4.7	4.7	4.7	4.6
large-v3	*au*	5.8	5.1	5.0	4.6	4.5	4.5	4.3
large-v3	*in*	6.6	6.0	5.8	5.8	5.8	5.8	5.7
large-v3	*sg*	6.5	5.1	5.1	5.1	5.1	5.1	4.9

- Almost no more improvements after 3 iterations

Conclusion & Discussion

Easy-to-use:

- A pretrained Model + 1-hour ***unlabeled*** speech
- **13.5%** relative WER reduction across **14** target domains (noise, accent, etc.)

Generalization:

- Other models: SeamlessM4T, OWSM, Canary
- Other task: Speech Translation

Anti-forgetting:

- Avoid common catastrophic forgetting in domain adaptation

Discussion

- Large models' attention matrix can present their uncertainty
- Self-improvement is possible in large speech foundation Model

Enhancing Zero-shot Text-to-Speech Synthesis with Human Feedback

Chen Chen[1*], Yuchen Hu[1*], Wen Wu[2], Helin Wang[3], Eng Siong Chng[1], Chao Zhang[4]

[1] Nanyang Technological University [2] University of Cambridge
[3] Johns Hopkins University [4] Tsinghua University

Zero-shot TTS

Zero-shot TTS: speak a sentence with *cloned voice*

Target Text

I have a dream.

Reference Speech

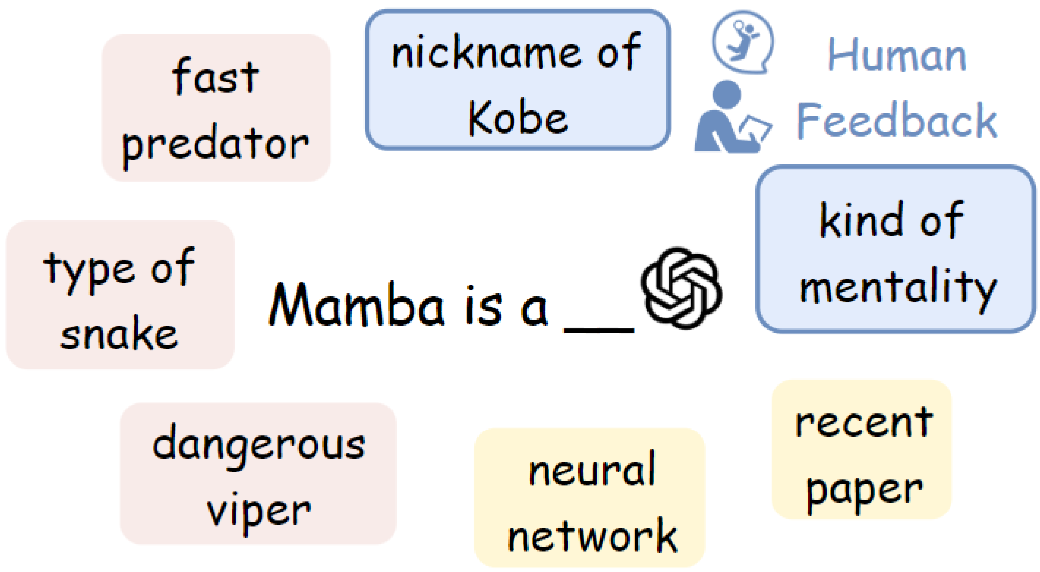
Synthesized Speech

"I have a dream."

Current open-sourced TTS model is not robust enough:

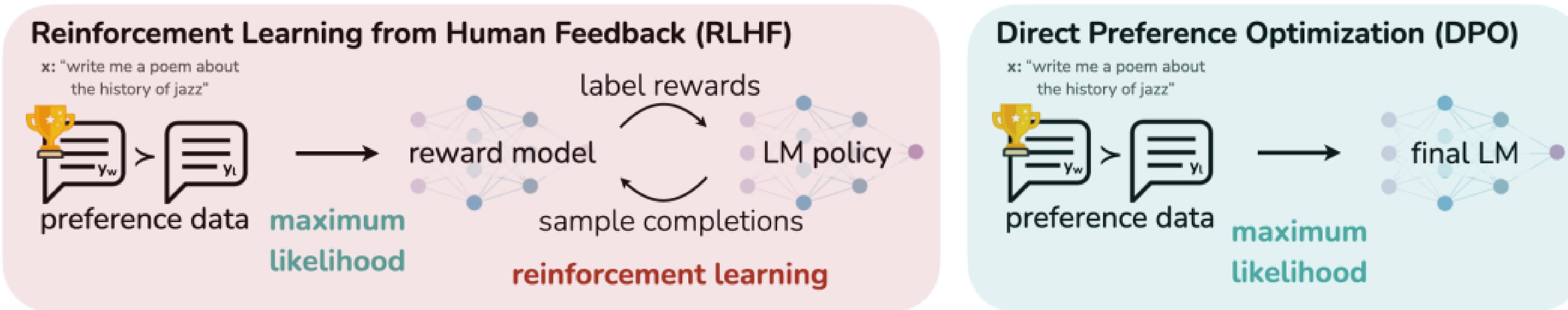
- **Objective Intelligibility:** missing words, wrong words, repetition
- **Subjective Naturalness:** unnatural prosody, tones (human preference)

RLHF is What We Need



- Align LLM's generation with human preference

PPO vs. DPO



- **PPO[1] (OpenAI):** most popular RLHF algorithm, make ChatGPT a success
- **DPO[2] (Stanford):** direct preference optimization without reward model

Reward model in TTS: MOSNet --> Not fine-grained enough

[1] Schulman, et al. "Proximal policy optimization algorithms." *arXiv preprint arXiv:1707.06347* (2017).
[2] Rafailov, et al. "Direct preference optimization: Your language model is secretly a reward model." *NeurIPS 2023*.

DPO in TTS?

Zero-shot TTS: speak a sentence with *cloned voice*

Target Text

I have a dream.

Reference

Speech

Synthesized

Speech

"I have a dream."

Question: If we want to use **DPO**, how to sample binary data?

$$\mathcal{L}_{\mathrm{DPO}}(\pi_\theta, \pi_{\mathrm{ref}}) =$$
$$\mathbb{E}\left[-\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right)\right]$$

Our backbone[3] lacks diversity:

w and l are not distinctive enough

[3] Peng, et al. "VoiceCraft: Zero-Shot Speech Editing and Text-to-Speech in the Wild." *ACL 2024*.

DPO in TTS?

Question: If we want to use **DPO**, how to sample binary data?

$$\mathcal{L}_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}\left[-\log \sigma\left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right]$$

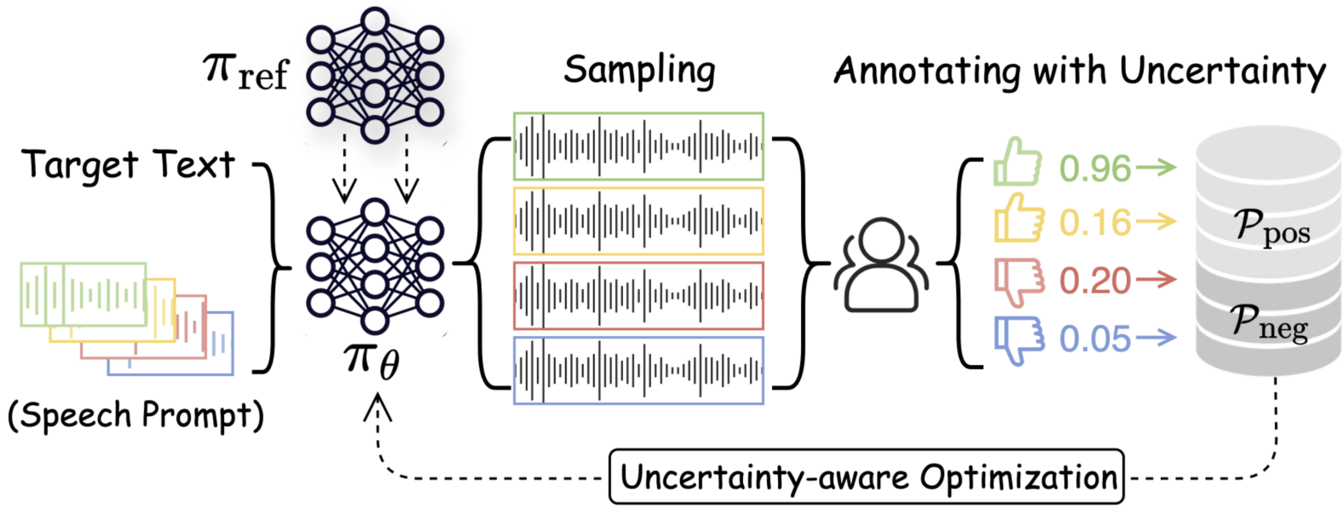
Our backbone[3] lacks diversity:
w and l are not distinctive enough

Solutions:

1. Ground-truth as pos, generated as neg (SpeechAlign[4])
2. Use StyleTTS to produce diversity
3. Change infer hyper-params: top-k, top-p, τ

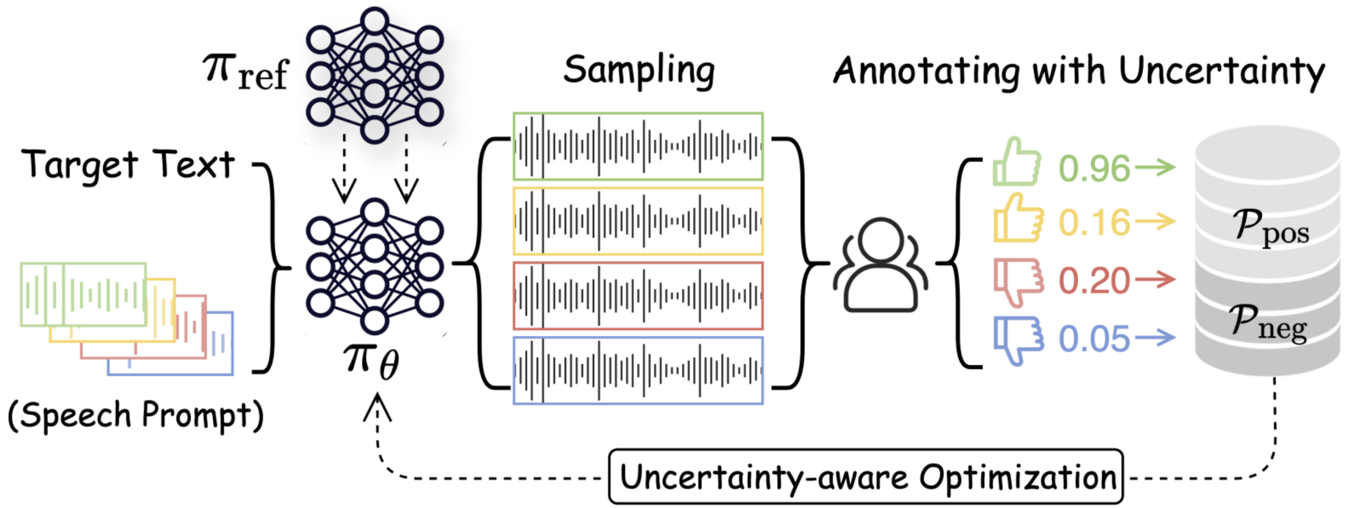
[4] Zhang, et al. "SpeechAlign: Aligning Speech Generation to Human Preferences." *arXiv preprint arXiv:2404.05600* (2024).

Our Method



- Sampling and annotating with MOS evaluators (human or NN)
- Incorporate MOS's uncertainty to reweight training samples
- Encourage TTS model to synthesize **good** speech and avoid **bad** speech

Our Method



- Encourage TTS model to synthesize **good** speech and avoid **bad** speech

$$\mathcal{L}_{\text{TTS}}(\pi_\theta, \pi_{\text{ref}}) = \mathbb{E}_{t,p,s\sim\mathcal{P}_{\text{pos}}\cup\mathcal{P}_{\text{neg}}}(1 - V_{\text{TTS}}(t,p,s;u)).$$

$$V_{\text{TTS}}(t,p,s;u) = \begin{cases} \sigma(u^{-1}\cdot R(t,p,s) - Z_{\text{ref}}), & \text{if } (t,p,s;u) \sim \mathcal{P}_{\text{pos}} \\ \sigma(Z_{\text{ref}} - u^{-1}\cdot R(t,p,s)), & \text{if } (t,p,s;u) \sim \mathcal{P}_{\text{neg}} \end{cases}$$

$$R(t,p,s) = \log\frac{\pi_\theta(s|t,p)}{\pi_{\text{ref}}(s|t,p)}$$

Objective Evaluation

Model	Label	WER↓ (%)	SIM↑ (0,1)	MOS ↑ by I-CNF	MOS ↑ by EDL	MOS ↑ by MOSNet
VoiceCraft (baseline)	-	8.4	0.84	3.51	3.55	3.65
SpeechAlign-DPO	✓	7.2	0.91	$3.70_{+0.19}$	$3.72_{+0.17}$	$3.86_{+0.21}$
SpeechAlign-ODPO	✓	6.9	0.90	$3.73_{+0.21}$	$3.76_{+0.24}$	$3.90_{+0.25}$
PPO-SDP	✗	7.7	0.88	$3.65_{+0.14}$	$3.69_{+0.14}$	$3.85_{+0.20}$
UNO-ICNF	✗	2.6	0.91	$\mathbf{3.93}_{\mathbf{+0.42}}$	$3.90_{+0.35}$	$\mathbf{4.31}_{\mathbf{+0.66}}$
UNO-EDL	✗	**2.4**	**0.92**	$3.88_{+0.37}$	$\mathbf{3.91}_{\mathbf{+0.36}}$	$4.28_{+0.63}$
GroundTruth (upper-bound)	-	2.0	-	4.15	4.19	4.52

- ICNF/EDL: which model to annotate training data
- ICNF/EDL: which model to do evaluation
- MOSNet: external evaluation

Training data: 400 utterances (200 pos, 200 neg)

Subjective Evaluation

Table 2: Results on human evaluation.

Model	MOS *by* Human	MOS *by* MOSNet
VoiceCraft	3.38	3.57
UNO-ICNF	4.06	4.20
UNO-Human	3.98	4.13
GroundTruth	4.55	4.46

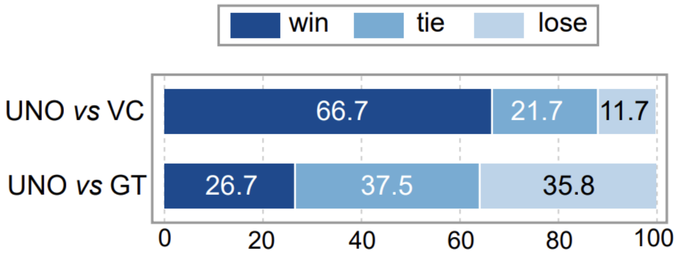


Figure 3: Result of A/B test. "VC" and "GT" denote the "VoiceCraft" and "GroundTruth".

- Human: human annotated training data
- Human: human evaluation

Demos

- **Objective Intelligibility:** missing words

VoiceCraft: Ours:

"If a layman in giving baptism pour the water before saying the words is the child baptized"

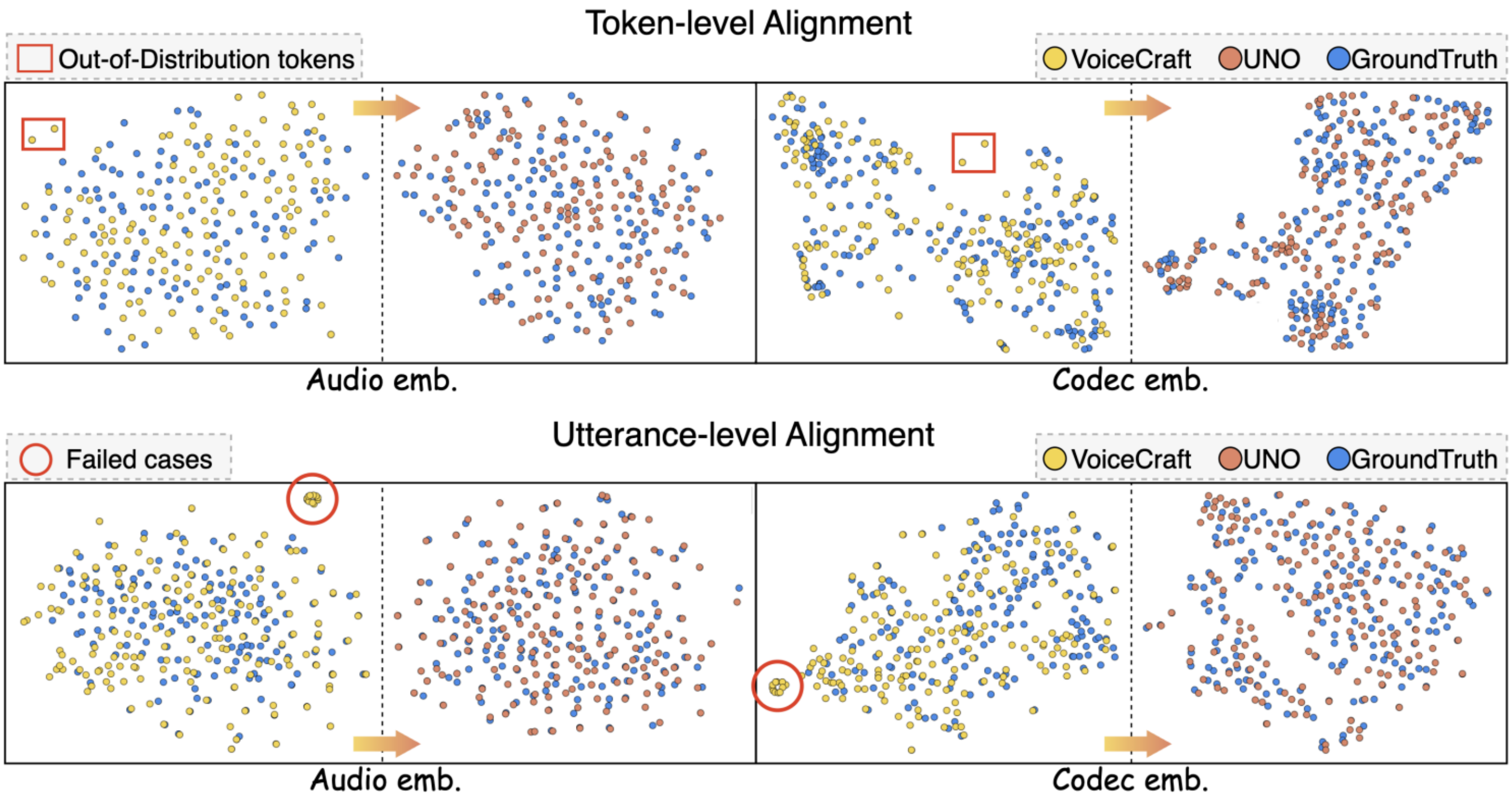
- **Subjective Naturalness:** unnatural rhythm and tones

VoiceCraft: Ours:

"Not only this but on the table I found a small ball of black dough or clay with specks of something which looks like sawdust in it"

https://yuchen005.github.io/UNO-TTS-demos/

Visualization



- Align better to the ground-truth speech
- Remove some bad cases

Scalability

EmotionTTS-Valence						EmotionTTS-Arousal					
$\mathcal{P}_{\text{pos}}$		$\mathcal{P}_{\text{neg}}$		*Valence* ↑		$\mathcal{P}_{\text{pos}}$		$\mathcal{P}_{\text{neg}}$		*Arousal* ↑	
$\bar{v}$	$\bar{m}$	$\bar{v}$	$\bar{m}$	*before*	*after*	$\bar{a}$	$\bar{m}$	$\bar{a}$	$\bar{m}$	*before*	*after*
0.65	4.08	0.36	4.04	0.55	0.67	0.69	4.05	0.48	4.20	0.62	0.71

- We can also use RLHF to improve emotion TTS

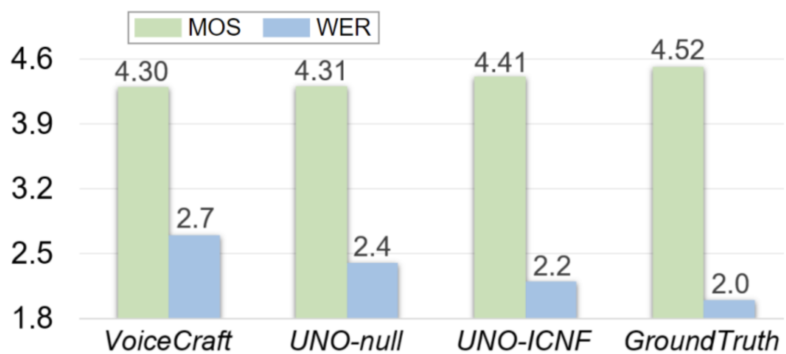
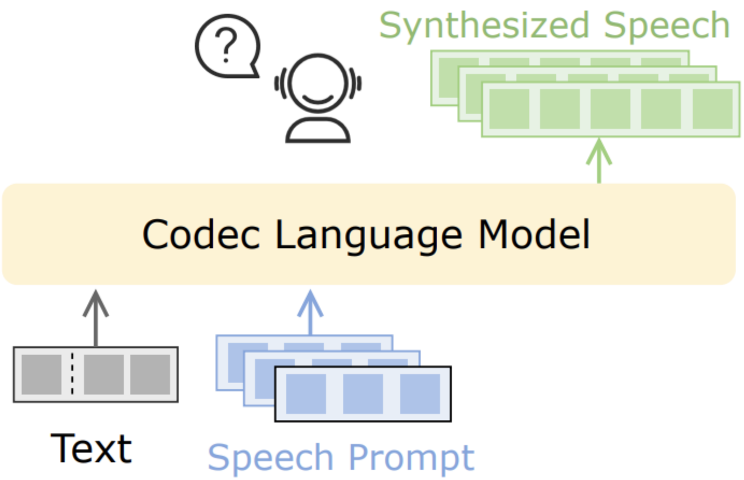


Figure 4: WER and MOS Results on 830M models.

- *VoiceCraft-830M* is already very good
- Our improvements are limited

RLHF on Larger TTS Model

VoiceCraft-830M synthesizes abundant **good** and **non-distinctive** speech:

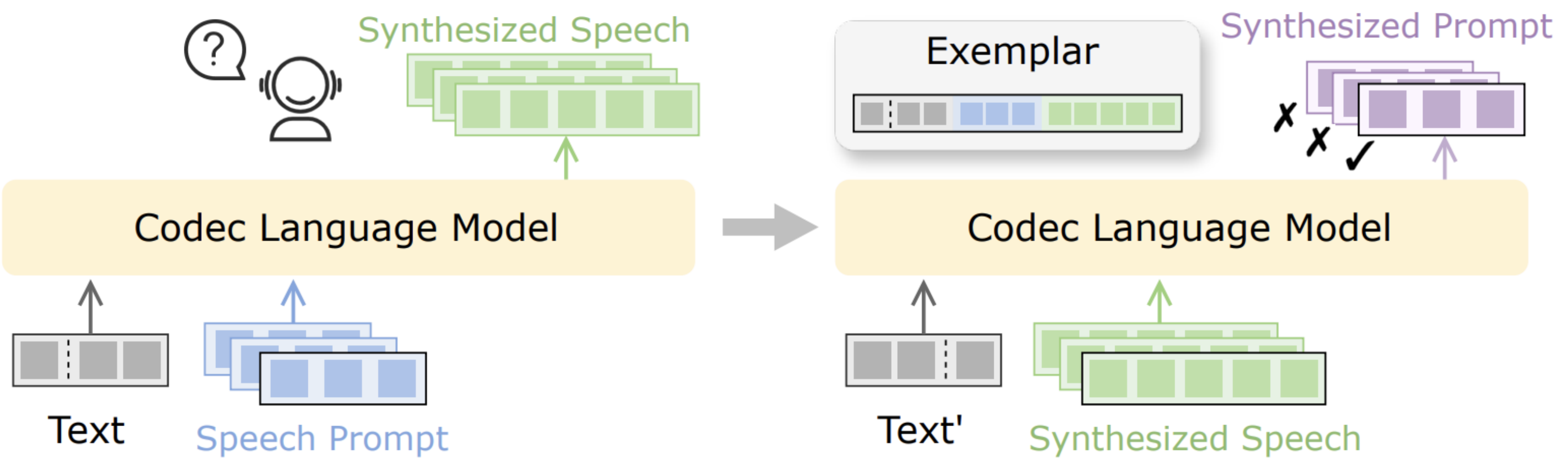


(a) Zero-shot TTS

- Are they really *good* ?
- Does TTS model understand what he is speaking ?

Reverse Inference Optimization

Let TTS model itself perceives the synthesized speech:



(a) Zero-shot TTS

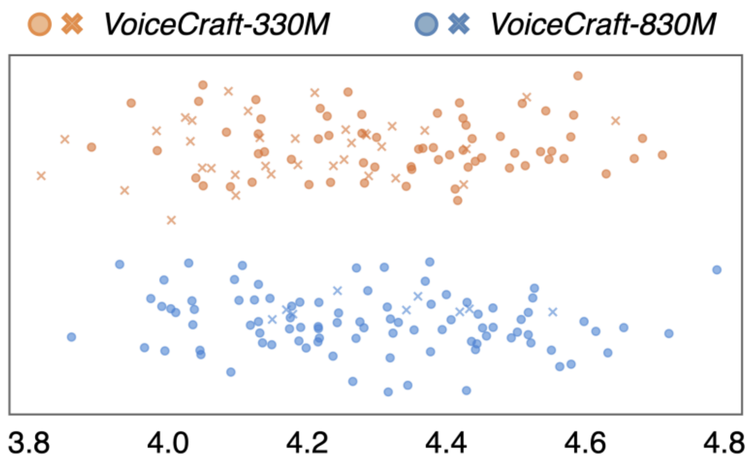
(b) Reverse Inference

Bayes' formula:

$$P(\mathbf{Y}|\mathbf{T}_\mathrm{Y},\mathbf{T}_\mathrm{X},\mathbf{X}) = \frac{P(\mathbf{X}|\mathbf{T}_\mathrm{X},\mathbf{T}_\mathrm{Y},\mathbf{Y})\,P(\mathbf{Y}|\mathbf{T}_\mathrm{Y},\mathbf{T}_\mathrm{X})}{P(\mathbf{X}|\mathbf{T}_\mathrm{X},\mathbf{T}_\mathrm{Y})} = \frac{P(\mathbf{Y}|\mathbf{T}_\mathrm{Y})}{P(\mathbf{X}|\mathbf{T}_\mathrm{X})}P(\mathbf{X}|\mathbf{T}_\mathrm{X},\mathbf{T}_\mathrm{Y},\mathbf{Y})$$

Empirical Observation

Ratio of bad reverse inference



MOS vs. Bad reverse inference

Model	✓✗	✓✓
VoiceCraft-330M	3.79	3.90
VoiceCraft-830M	4.24	4.32

(✓ : good generation, MOS > 3)

- Bad reverse inference results correspond to low TTS robustness
- Human/MOSNet cannot perceive subtle differences in speech, but TTS model can

Evaluation on Longer Test Samples

Model	WER↓ (%)	SIM↑ (0,1)	MOS ↑ by MOSNet	MOS ↑ by Human	Bad Case Ratio ↓ MOS ⩽ 3	Bad Case Ratio ↓ %WER > 20
VoiceCraft	35.3	0.79	3.36	3.22	27%	51%
RIO-DPO	11.3	0.92	$4.11_{+0.75}$	-	5%	17%
RIO-ODPO	9.2	0.93	$4.15_{+0.79}$	-	5%	15%
UNO-null	6.8	0.93	$4.20_{+0.84}$	-	4%	11%
RIO (ours)	**3.4**	**0.96**	$\mathbf{4.40}_{+1.04}$	$\mathbf{4.18}_{+0.96}$	**1%**	**4%**
Ground-Truth	1.3	-	4.48	4.54	0%	0%

Results on 330M models

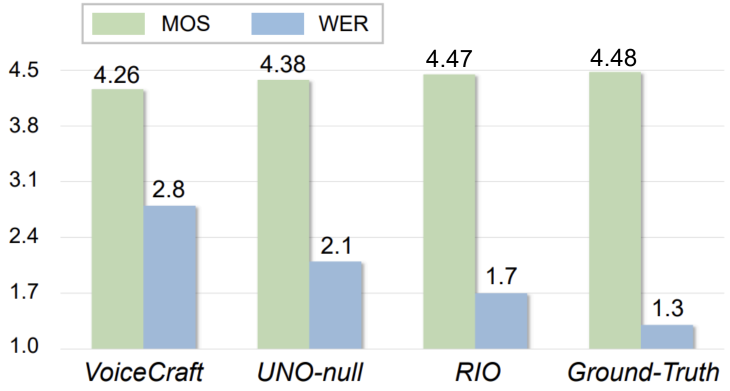


Figure 4: MOS and WER Results on 830M models.

- Less bad cases, more robust TTS
- Approach GT on 830M backbone

Performance on Large TTS Model

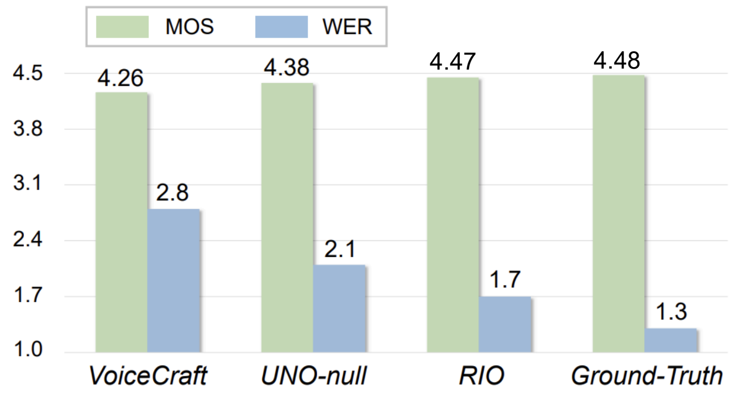


Figure 4: MOS and WER Results on 830M models.

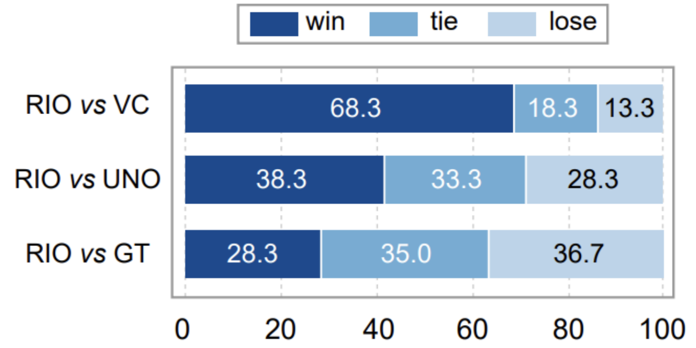
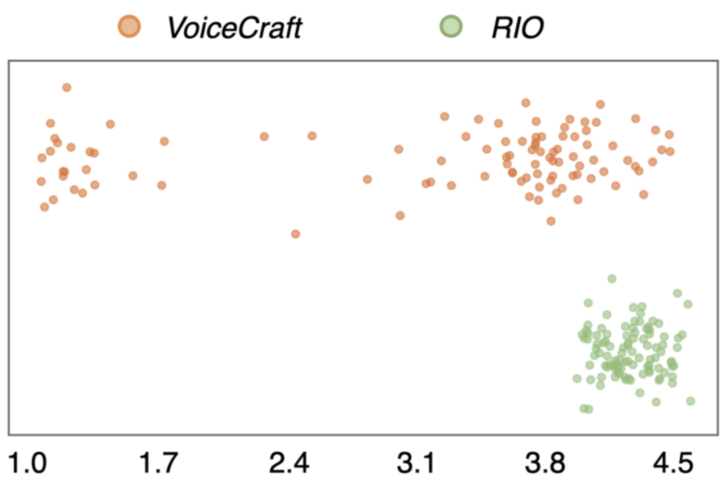


Figure 3: Results of A/B test. "VC" and "GT" denote the "VoiceCraft" and "Ground-Truth".

- High-level sampling can push the limit of RLHF on large TTS models

Analysis

MOS distribution



Ratio of good reverse inference in good zero-shot samples

Model	Baseline	RIO
VoiceCraft-330M	54%	85%
VoiceCraft-830M	80%	97%

- Better MOS distribution: higher mean and lower variance
- More good reverse inference results after post-training --> better robustness

Conclusion & Discussion

RLHF Post-training in TTS

- Improve subjective naturalness and expressiveness
- Correct many objective issues

Annotating is important in RLHF-TTS:

- Positive samples: Good enough, i.e., natural, expressive, correct
- Negative samples: Cover as many error types as possible

Discussion:

- For general speech perception: Model > Human
- TTS model can **not** understand what it predict, similar to Whisper/Qwen-Audio
- Future: Speech/text understanding/generation are unified in one single model

Summary

Pretraining & SFT:

$$\mathcal{L}(x,y) = \sum_{l=1}^{L} -\log \mathcal{P}_{\theta}(y_l | y_{l-1}, \cdots, y_1, x)$$

Learn a distribution --> uncertain

Post-training:

Sampling → Annotating → Learning

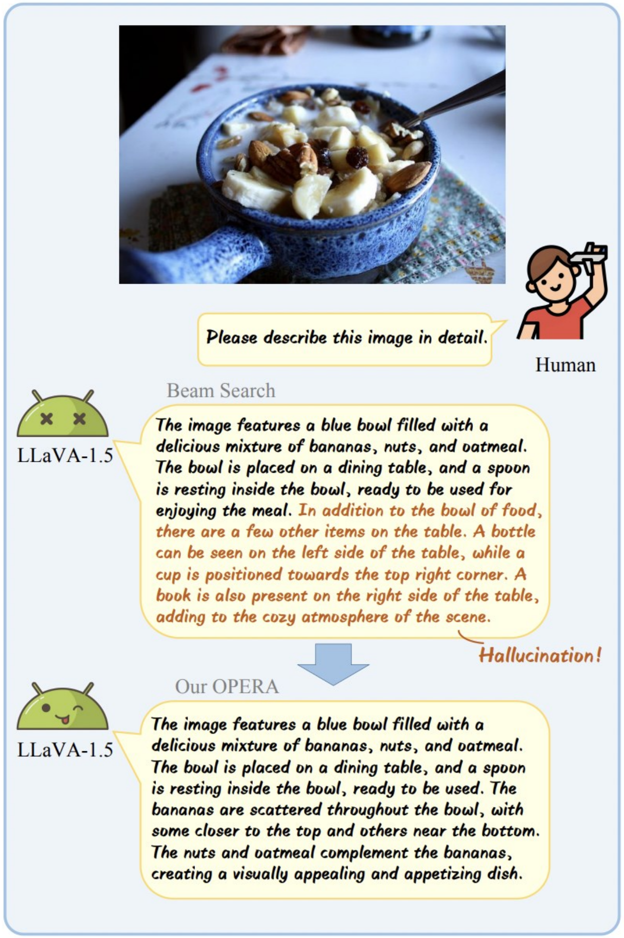
↑

Knowledge

- Appropriate annotating can help efficiently improve the performance!

Thank you! & QA

Appendix: LLM Hallucination



- **Non-Hallucinations:** describes the food (e.g., bananas, nuts, oatmeal) inside the bowel
- **Hallucinations:** imagines the items on the table that is outside the image

***NOTE:** Hallucinations starts with "In addition to …"*

Appendix: LLM Hallucination

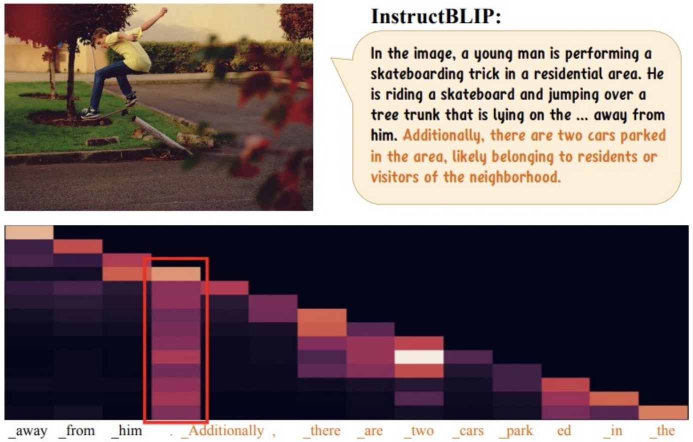


Figure 2. A case of relationship between hallucinations and knowledge aggregation patterns. Hallucinations are highlighted.

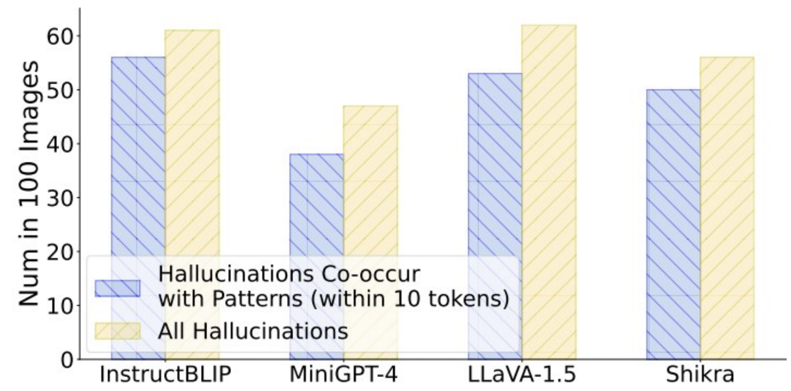


Figure 3. Hallucinations often start within the first 10 tokens after knowledge aggregation patterns.

- Hallucinations are usually triggered by specific tokens (e.g., "*additionally*");
- We can observe a "knowledge aggregation pattern" in self-attention map along with the beginning of hallucinations → *An insightful finding!*

Appendix: LLM Hallucination

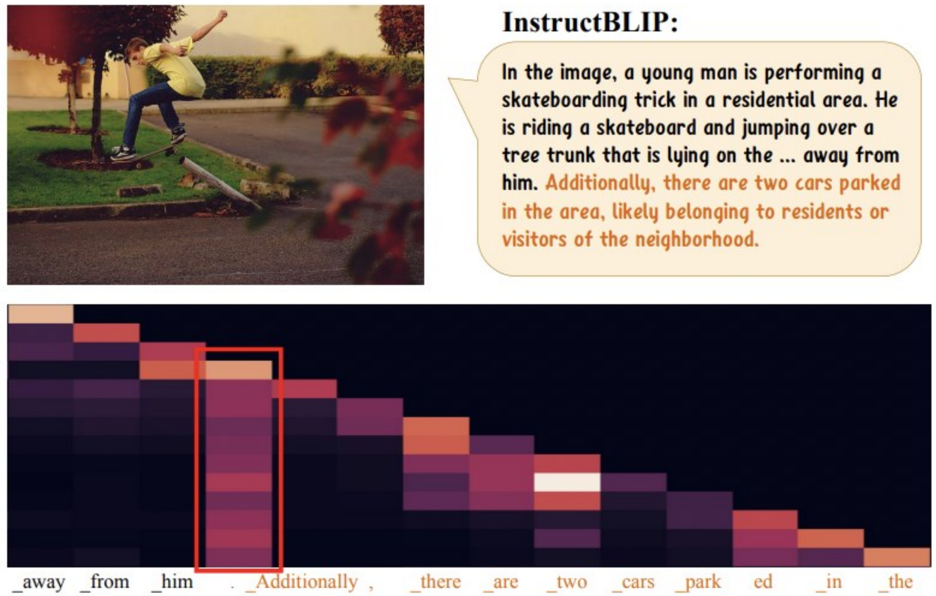


Figure 2. A case of relationship between hallucinations and knowledge aggregation patterns. Hallucinations are highlighted.

All hallucinations are highly related to the starting token "*additionally*" but unrelated to previous normal tokens!