APSIPA ASC 2026

Speak Better via Simultaneous Listening and Speaking in Unified Decoder-Only LLMs

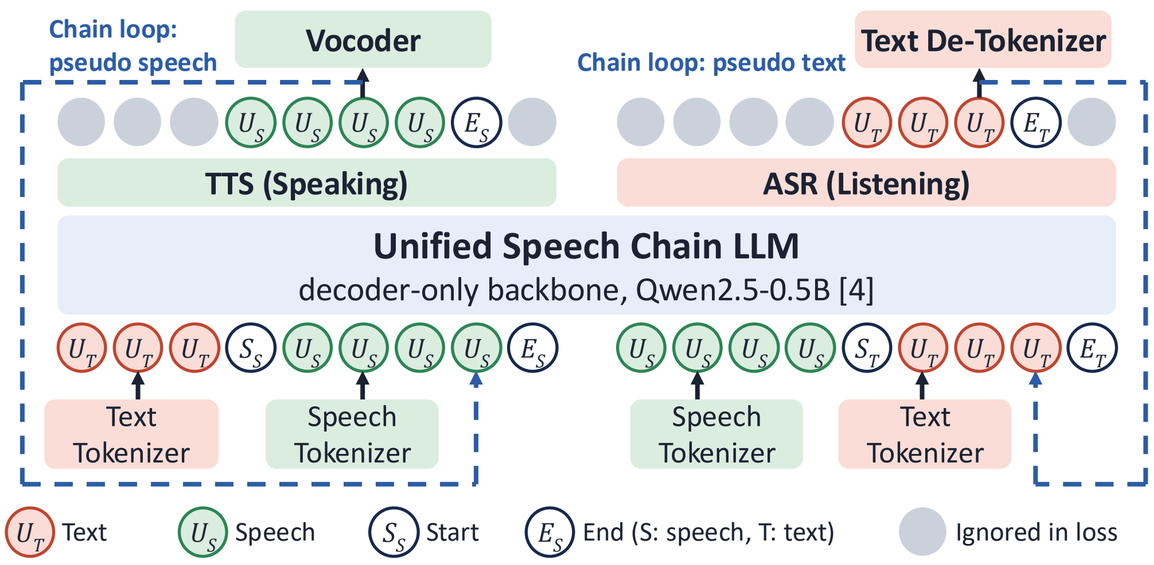
Chung Tran, Bagus Tris Atmaja, Sakriani Sakti

Nara Institute of Science and Technology, Japan

1 Motivation

- LLM-TTS such as CosyVoice2 [1] and Spark-TTS [2] need over 100,000 hours of training data, which limits low-resource languages.
- Machine Speech Chain [3] couples TTS with ASR to use unlabeled data, but needs two models and complex training.
- Idea: train speaking with listening in one decoder-only model.

2 Unified Speech Chain (USC)



Pseudo-Algorithm: Training USC-LLM model

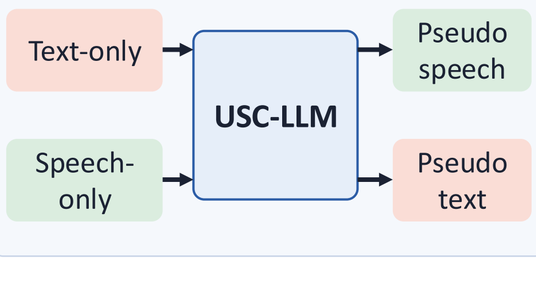
Initialize USC-LLM model
while not converged **do**
  Sample a batch of data B_{tts} and B_{asr}
  $L_{tts} \leftarrow$ USC-LLM(B_{tts}) *// Speaking*
  $L_{asr} \leftarrow$ USC-LLM(B_{asr}) *// Listening*
  $L_{total} \leftarrow L_{tts} + L_{asr}$
  Back-propagation to update parameters
end

3 Multi-Stage Chain Loop (MSCL)

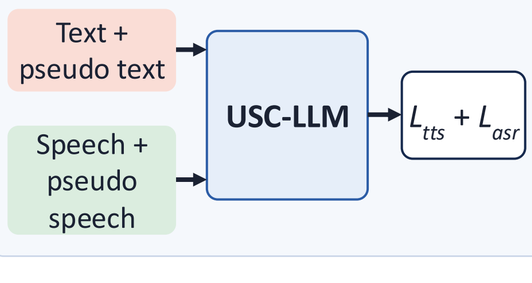
Stage 1: Supervised Training

- Text (BPE) and speech (mHuBERT [5]) share one token stream.
- Paired data trains both tasks together.

Stage 2: generate pseudo-data



Stage 3: chain loop



4 Experimental setup

- Monolingual: LJSpeech [6], 24 h of English, split 80/10/10.
- Multilingual: 7 languages, 6–24 h each, from LJSpeech, CSS10 [7], VIVOS [8] and INDSpeech [9].
- Metrics: WER and CER on Whisper-Large V2 [10] transcripts; UTMOS [11] for naturalness.

5 Same data, listening helps speaking

Data	Depth	TTS WER ↓	USC WER ↓	TTS CER ↓	USC CER ↓
50%	12L	**70.08**	71.05	**57.43**	58.69
50%	24L	26.35	**20.94**	18.67	**14.96**
75%	12L	40.95	**30.70**	32.90	**24.28**
75%	24L	13.44	**12.90**	8.94	**8.77**
100%	12L	24.24	**20.51**	18.75	**15.98**
100%	24L	10.27	**10.06**	6.66	**6.56**

USC wins 5 of 6 settings on LJSpeech. At full data, 24L USC has UTMOS 3.92, and VITS [12] has WER 13.74, CER 3.59, UTMOS 3.87.

6 Impact of MSCL on speech synthesis

Model	Labeled	PsData	ASR PsData		TTS PsData		ASR-TTS PsData	
			WER ↓	CER ↓	WER ↓	CER ↓	WER ↓	CER ↓
24L (LB)	50%	–	20.94	14.96	20.94	14.96	20.94	14.96
24L	50%	25%	15.50	10.70	22.00	16.35	19.65	14.48
24L	50%	50%	13.76	9.56	20.68	15.27	20.36	15.19
24L (UB)	100%	–	10.06	6.56	10.06	6.56	10.06	6.56
24L (LB)	75%	–	12.90	8.77	12.90	8.77	12.90	8.77
24L	75%	15%	12.20	8.64	11.57	7.83	11.70	7.83
24L	75%	25%	11.58	7.55	10.04	6.64	10.64	6.85
24L (UB)	100%	–	10.06	6.56	10.06	6.56	10.06	6.56

LB/UB: lower/upper bound; PsData: pseudo-data.

7 MSCL for ASR

Labeled	Pseudo	USC WER ↓	USC CER ↓	MSC [3] CER ↓
50%	–	13.78	7.86	–
50%	50%	13.16	7.21	3.5
75%	–	10.39	5.22	–
75%	25%	9.34	4.41	–
100%	–	8.82	4.08	3.1

MSCL improves WER of ASR task.

8 Multilingual low-resource results

	12L WER ↓		12L CER ↓		24L WER ↓		24L CER ↓	
Lang	TTS	USC	TTS	USC	TTS	USC	TTS	USC
ZH	37.62	**36.65**	17.32	**16.28**	36.54	**36.12**	15.86	**15.53**
EN	13.64	**10.58**	9.08	**6.70**	10.97	**10.15**	7.00	**6.19**
FR	20.38	**18.00**	10.30	**8.39**	18.04	**17.82**	8.32	**7.97**
ID	43.03	**35.62**	15.40	**11.98**	**35.06**	35.73	12.78	**12.28**
JA	31.03	**30.31**	13.65	**13.36**	**30.67**	31.59	**13.15**	13.91
ES	10.75	**9.09**	4.66	**3.84**	8.92	**8.40**	3.46	**3.35**
VI	**22.93**	23.82	**12.51**	13.06	23.76	**22.64**	13.01	**12.53**

Shaded cells mark the lower error.

Conclusion

USC unifies listening and speaking in one decoder-only model. With MSCL it learns from paired, unpaired and unlabeled data, improving intelligibility while keeping naturalness on par with strong baselines.

[1] Z. Du et al., CosyVoice 2, 2024. [2] X. Wang et al., Spark-TTS, 2025.
[3] A. Tjandra et al., Machine Speech Chain, IEEE/ACM TASLP, 2020.
[4] Qwen, Qwen2.5 Technical Report, 2025. [5] M. Z. Boito et al., mHuBERT-147, Interspeech 2024.
[6] K. Ito and L. Johnson, The LJ Speech Dataset, 2017. [7] K. Park and T. Mulc, CSS10, Interspeech 2019.
[8] H.-T. Luong et al., VIVOS, 2016. [9] S. Sakti et al., Indonesian LVCSR within A-STAR, TCAST 2008.
[10] A. Radford et al., Robust Speech Recognition via Large-Scale Weak Supervision, ICML 2023.
[11] K. Baba et al., UTMOSv2, IEEE SLT. [12] J. Kim et al., Conditional VAE with Adversarial Learning for TTS, ICML 2021.

Part of this work was supported by JSPS KAKENHI JP21H05054, JP23K21681, JP24K0296, JP25H01139 and JST NEXUS (JPMJNX25C1).