SubAlign: Speech Tokenization Aligned with LLM Vocabularies for Spoken Language Modeling

Kang-wook Kim, Sehun Lee, Sang Hoon Woo, Gunhee Kim

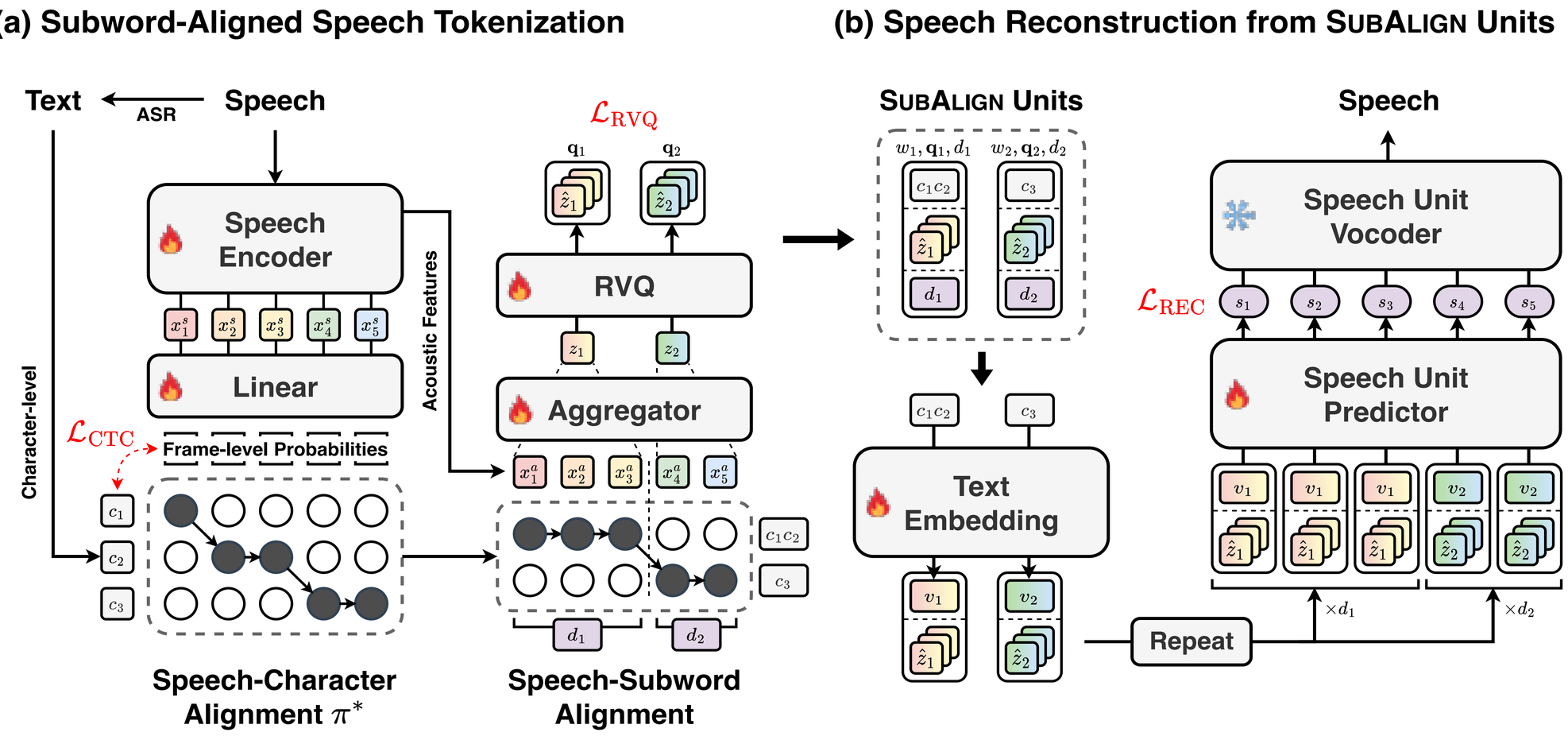
Personal Website:

SEOUL NATIONAL UNIV. VISION & LEARNING

Bridge speech and text seamlessly:
LLM subword-aligned speech tokenization for SLMs

SubAlign Framework

- **First speech tokenization** framework that segments speech at the **subword level** aligned with LLM vocabularies



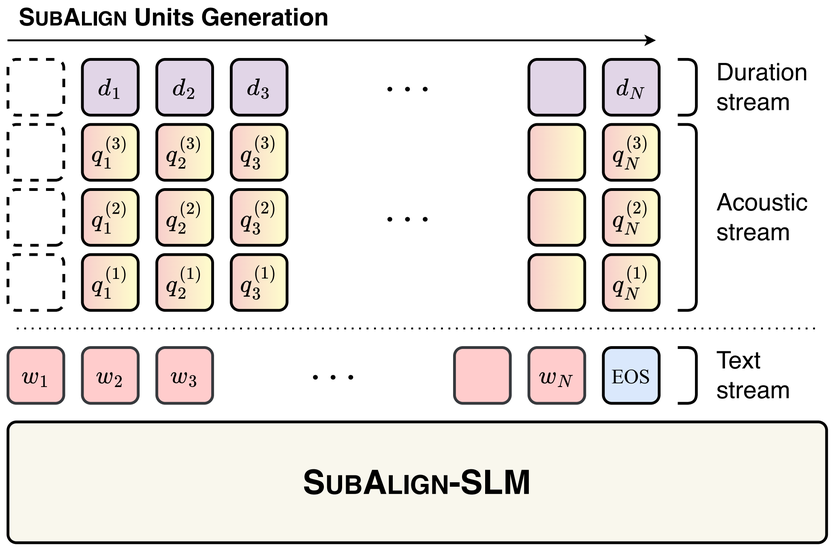
Speech Reconstruction

- Compresses speech to **193 bps** *(LibriTTS)*
- Achieves the highest perceptual quality *(UTMOS)* while maintaining acoustic details

Model	Bitrate ↓	Quality		Similarity		
		WER ↓	UTMOS ↑	Spk Sim. ↑	F0 RMSE ↓	Duration ↑
Mel + BigVGAN	–	4.23	3.73	0.87	57.1	1.00
SpeechTokenizer	4000	5.06	3.57	0.82	67.3	0.98
Mimi	1000	7.28	3.37	0.63	66.9	0.98
Sylber *(w/o quant.)*	–	8.76	3.99	0.37	72.5	0.96
TASTE	195	9.69	3.94	0.59	79.0	0.93
CosyVoice Tokens	600	6.48	3.77	0.65	71.1	0.97
Ours *(w/o quant.)*	–	7.01	4.01	0.61	71.9	0.97
Ours	193	5.66	4.05	0.59	73.9	0.96

Spoken Language Modeling

- **SubAlign-SLM:** a spoken language model trained on SubAlign units



Method	Size	Likelihood Evaluation		Continuation Evaluation			
		Acoustic	Semantic	GPT-4o	UTMOS	SECS	Human Eval
ASR + LLM (Llama-3.2)	1B	–	78.2	2.54	**3.60**	0.609	4.020
ASR + LLM (Qwen3)	1.7B	–	75.7	2.40	3.58	0.596	3.717
TWIST	1.3B	65.0	61.5	1.96	3.58	–	–
TWIST	7B	66.0	64.7	2.23	3.38	–	–
SpiRit-LM	7B	63.1	72.0	2.45	3.30	–	–
SpiRit-LM *(expr.)*	7B	**72.1**	66.2	1.87	3.20	–	–
TASLM *(token)*	1B	61.9	76.5	2.73	3.54	0.556	3.220
SubAlign-SLM *(Ours)*	1B	69.8	**79.6**	**3.07**	3.38	**0.642**	**4.187**

- Likelihood-based classification → excels in **both acoustic and semantic** tasks
- Speech continuation
  - Strong **semantic coherence** (GPT-4o as a judge: 3.07 vs. 2.54 cascaded)
  - Highest speaker consistency & human evaluation scores

References

Tseng, Liang-Hsuan, et al. "TASTE: Text-Aligned Speech Tokenization and Embedding for Spoken Language Modeling." *arXiv preprint arXiv:2504.07053* (2025).
Maimon, Gallil, Amit Roth, and Yossi Adi. "Salmon: A suite for acoustic language model evaluation." *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025.
Hassid, Michael, et al. "Textually pretrained speech language models." *Advances in Neural Information Processing Systems* 36 (2023): 63483-63501.