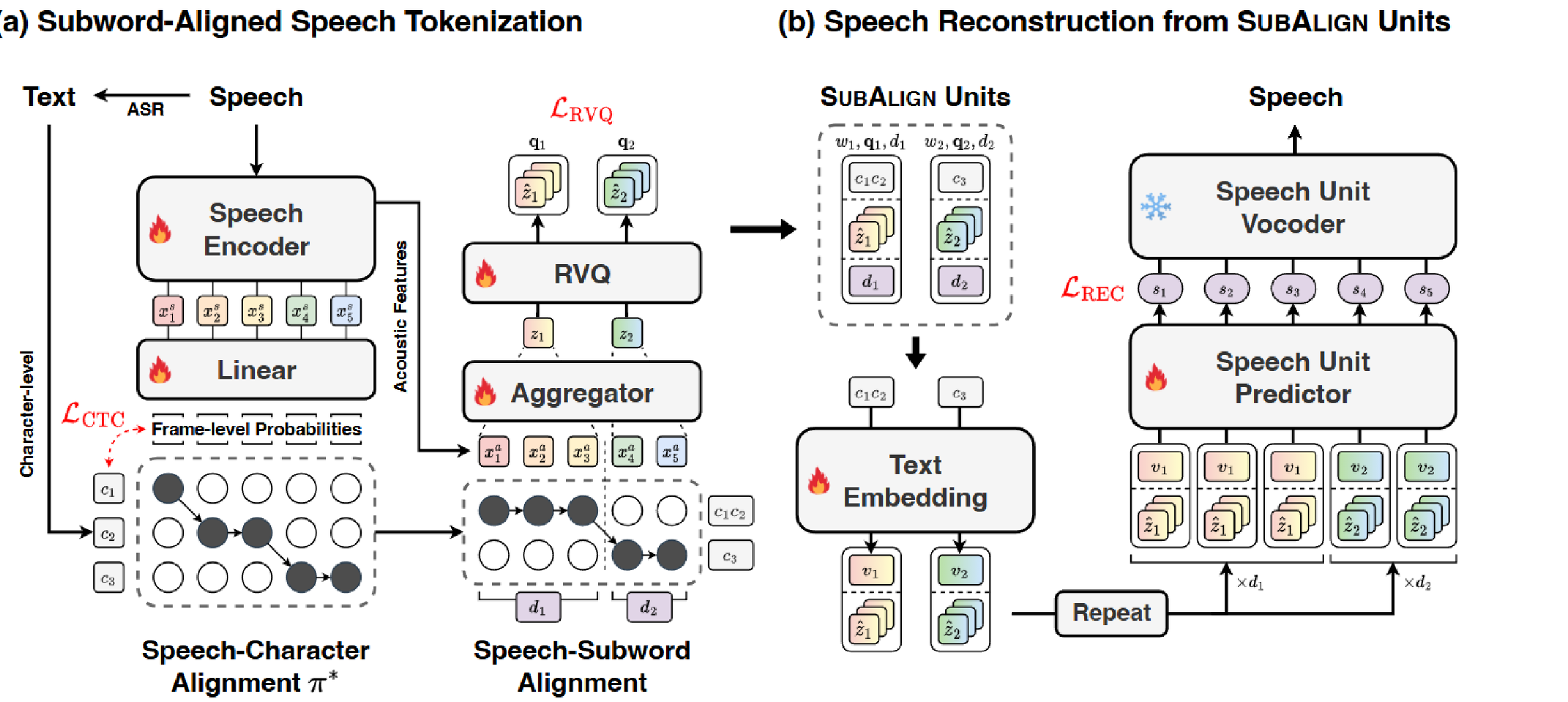
SubAlign: Speech Tokenization Aligned with LLM Vocabularies for Spoken Language Modeling

Anonymous submission

SubAlign Framework

We propose **SubAlign,** the first speech tokenization framework that segments speech at the subword level to better match large language model (LLM) vocabularies.



Speech Reconstruction Results

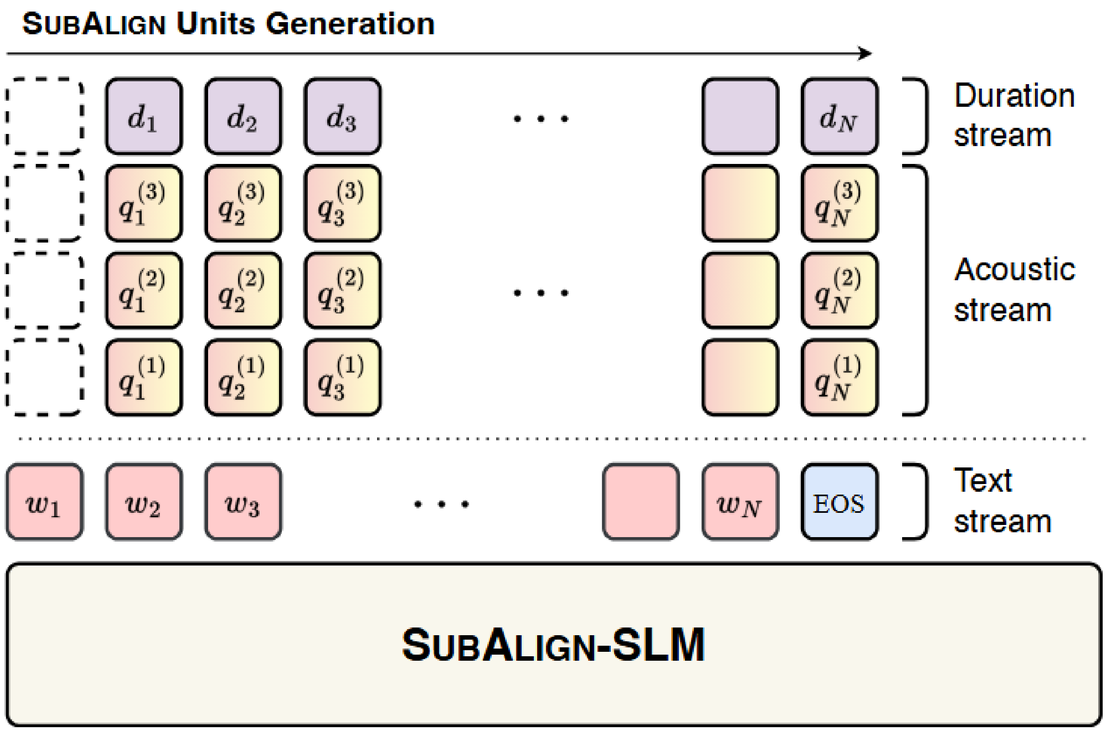
SubAlign encodes both speech and its transcript at a low bitrate, while maintaining the quality and similarity of the reconstructed waveform to the original speech.

Model	Bitrate ↓	Quality		Similarity			
		WER ↓	UTMOS ↑	SECS ↑	F0 RMSE ↓	STOI ↑	DC ↑
Mel + BigVGAN	–	4.23	3.73	0.867	57.13	0.992	0.996
SpeechTokenizer	4000	5.06	3.57	0.820	67.25	0.909	0.982
SpeechTokenizer	2000	7.04	3.28	0.639	71.21	0.866	0.977
SpeechTokenizer	1000	10.90	2.21	0.325	89.49	0.765	0.969
Mimi	1000	7.28	3.37	0.633	66.85	0.897	0.978
WavTokenizer	900	7.82	3.69	0.711	68.08	0.905	0.981
S^3 Tokens	600	6.48	3.77	0.649	71.09	0.779	0.972
Sylber *(w/o quant.)*	–	8.76	3.99	0.369	72.49	0.737	0.958
TASTE	195	9.69	3.94	0.585	79.00	0.431	0.928
Ours *(w/o quant.)*	–	7.01	4.01	0.610	71.90	0.740	0.969
Ours	193	5.66	4.05	0.587	73.89	0.686	0.959

Table 1: **Comparison of reconstruction and perceptual metrics.** We report Bitrate, WER, UTMOS, SECS, F0 RMSE, STOI, and DC. Lower is better for ↓ metrics; higher is better for ↑ metrics. '*w/o quant.*' indicates models using continuous tokens.

SubAlign-SLM

Building on this framework, we present **SubAlign-SLM**, a spoken language model trained on SubAlign units, and demonstrate the effectiveness of SubAlign on downstream tasks.



Likelihood-based Evaluation

SubAlign-SLM demonstrates robust performance on SALMon and the spoken StoryCloze benchmark, indicating that SubAlign effectively balances acoustic and semantic representational capabilities.

			Acoustic					Semantic		
Method	**Backbone**	**Size**	**Sentiment**	**Speaker**	**Gender**	**Energy**	**Avg.**	**sSC**	**tSC**	**Avg.**
ASR + LLM	Llama-3.2	1B	–	–	–	–	–	66.2	<u>90.3</u>	<u>78.2</u>
ASR + LLM	Qwen3	1.7B	–	–	–	–	–	65.6	85.7	75.7
TWIST	OPT	1.3B	61.5	69.0	69.5	60.0	65.0	52.4	70.6	61.5
TWIST	Llama	7B	61.5	71.0	70.0	61.5	66.0	55.3	74.1	64.7
SpiRit-LM	Llama-2	7B	54.5	69.5	67.0	<u>61.5</u>	63.1	61.0	82.9	72.0
SpiRit-LM (*expr.*)	Llama-2	7B	**73.5**	**81.0**	**85.0**	49.0	**72.1**	56.9	75.4	66.2
TASLM (*embed.*)	Llama-3.2	1B	57.5	67.0	75.5	–	–	64.0	89.5	76.7
TASLM (*token*)	Llama-3.2	1B	59.0	68.0	70.5	50.0	61.9	64.2	88.9	76.5
Ours										
SUBALIGN-SLM	Llama-3.2	1B	<u>67.0</u>	69.5	77.0	**62.0**	68.9	63.9	89.0	76.5
SUBALIGN-SLM (*punc*)	Llama-3.2	1B	65.5	<u>75.0</u>	77.5	61.0	<u>69.8</u>	**67.7**	**91.5**	**79.6**
SUBALIGN-SLM	Qwen3	1.7B	65.5	66.5	<u>78.5</u>	58.0	67.1	<u>66.5</u>	87.7	77.1

Table 2: **Results of different SLMs on SALMon and StoryCloze.** We report likelihood-based accuracy on SALMon (acoustic aspect) and StoryCloze (semantic aspect). The best scores are highlighted in **bold**, and the second-best scores are <u>underlined</u>.

Speech Continuation Results

SubAlign-SLM also demonstrates strong performance in speech continuation evaluation.

Method	**Backbone**	**Size**	**GPT-4o**	**UTMOS**	**SECS**	**Human Eval**
ASR + LLM + TTS	Llama-3.2	1B	2.54 ± 0.20	$\mathbf{3.60 \pm 0.14}$	0.609 ± 0.021	4.020 ± 0.120
ASR + LLM + TTS	Qwen3	1.7B	2.40 ± 0.20	3.58 ± 0.13	0.596 ± 0.022	3.717 ± 0.140
TWIST	OPT	1.3B	1.96 ± 0.12	3.58 ± 0.12	–	–
TWIST	Llama	7B	2.23 ± 0.16	3.38 ± 0.16	–	–
SpiRit-LM	Llama-2	7B	2.45 ± 0.22	3.30 ± 0.05	–	–
SpiRit-LM (*expr.*)	Llama-2	7B	1.87 ± 0.14	3.20 ± 0.08	–	–
TASLM (*token*)	Llama-3.2	1B	2.73 ± 0.18	3.54 ± 0.11	0.556 ± 0.024	3.220 ± 0.160
Ours						
SUBALIGN-SLM	Llama-3.2	1B	$\mathbf{3.07 \pm 0.19}$	3.38 ± 0.14	$\mathbf{0.642 \pm 0.025}$	$\mathbf{4.187 \pm 0.105}$
SUBALIGN-SLM	Qwen3	1.7B	2.97 ± 0.21	3.31 ± 0.14	0.636 ± 0.024	4.003 ± 0.120

Table 3: **Speech continuation results across different SLMs.** We report scores for semantic quality (GPT-4o), acoustic quality (UTMOS), and consistency with the prompt waveform (SECS and Human Evaluation) of the continuation. Higher values indicate better performance. For SECS and Human Evaluation, results are only available for methods with access to prompt-consistent generation.