Paralinguistics-Aware Speech-Empowered Large Language Models for Natural Conversation

Heeseung Kim[1], Soonshin Seo[2], Kyeongseok Jeong[2], Ohsung Kwon[2], Soyoon Kim[2], Jungwhan Kim[2], Jaehong Lee[2], Eunwoo Song[2], Myungwoo Oh[2], Jung-Woo Ha[2], Sungroh Yoon[1,†], Kang Min Yoo[2,†]

[1]Seoul National University, [2]NAVER Cloud, [†]Corresponding Authors

Motivation

- Dominant Approach in Existing Spoken Dialog Models: **"Cascaded"**

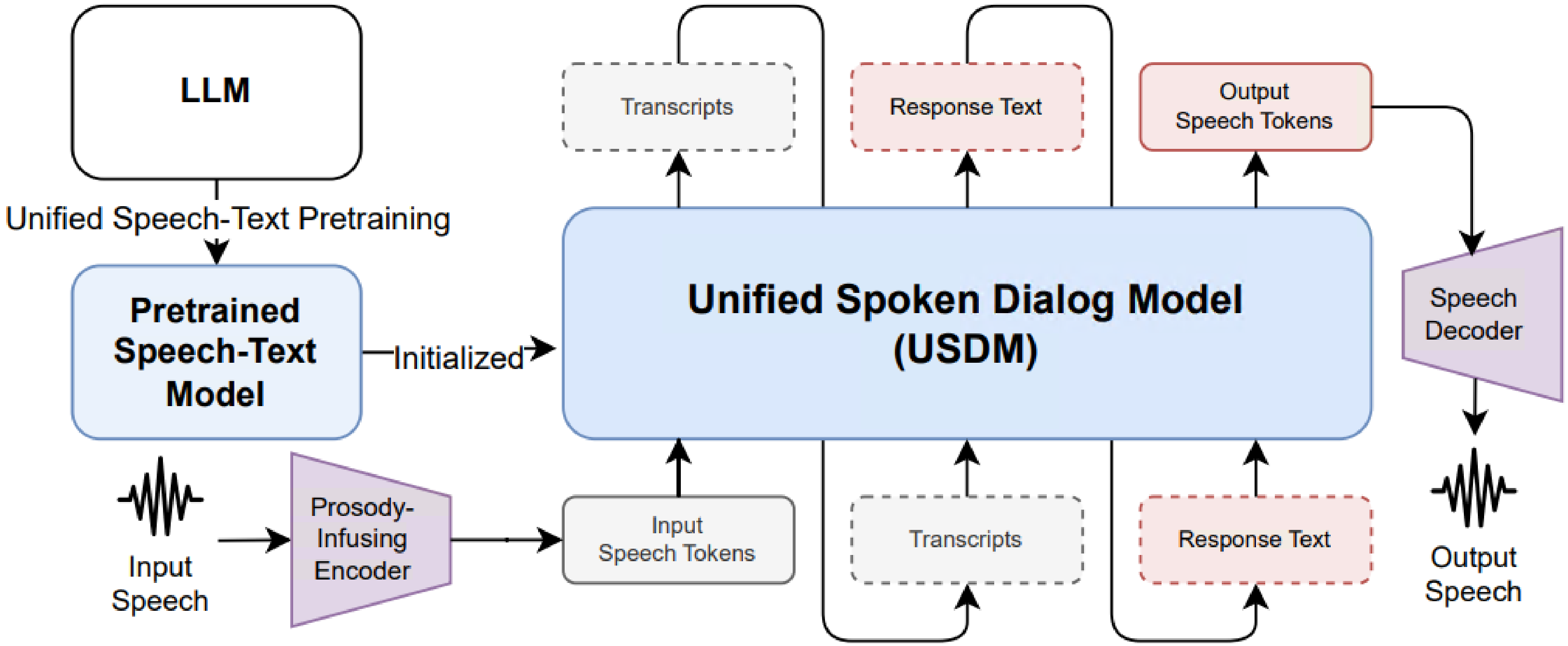
Automatic Speech Recognition (ASR) → LLM-based Chatbot → Text-to-Speech (TTS)

  – Combining ASR, Text-based Chatbot, TTS → **paralinguistic** information in user's speech is lost

- We propose the Unified Spoken Dialog Model (USDM), an **end-to-end** spoken dialog model that is **paralinguistic-aware**.

Overview of USDM

- Constructing USDM follows these three steps:
  1. **Pre-training the LLM** (we used the pre-trained Mistral-7B-v0.1[1])
  2. **Further Training** to expand the LLM for **speech modality**
  3. **Supervised Fine-tuning** for Downstream Tasks (in our work, Spoken Dialog Modeling)



USDM – Speech Tokenization

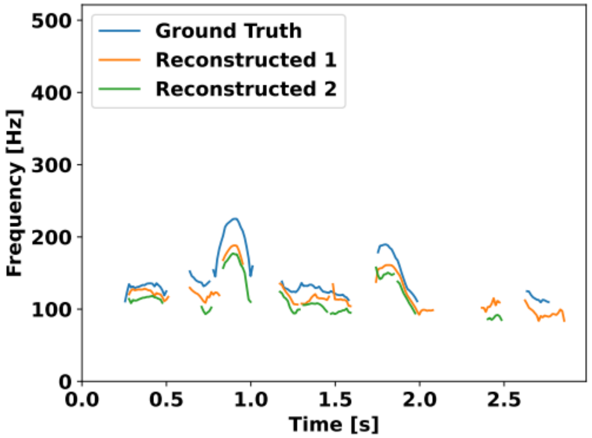
- **Compressing** speech along the **time axis** and converting it into a **discrete token sequence**.

10 399 1022 320 5 …

Speech (16kHz ~ 96kHz)

Speech Token (12.5Hz ~ 75Hz)

  – Observation: speech tokens used in SeamlessM4T[2] contain **pronunciation** and **non-verbal** cues.
    - A 50Hz token with |V| = 10k (k-means clustering to the intermediate representations of XLS-R[3])



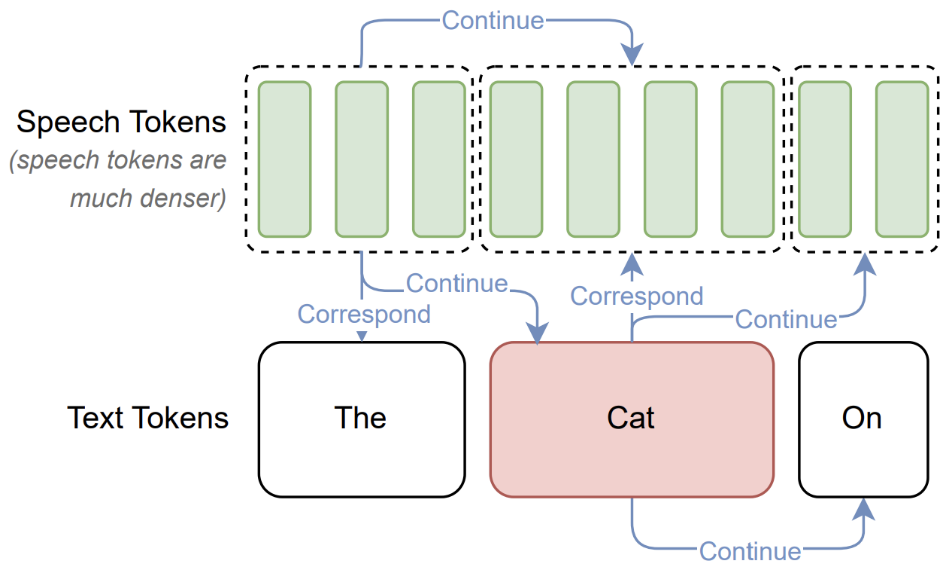
Pitch Contour

Dataset	Class	Guess Acc	Classifier Acc
CREMA-D	Emotion	16.6%	60.8%
TextrolSpeech	Gender	50.8%	83.4%
	Pitch	34.2%	70.9%
	Tempo	63.8%	82.4%
	Energy	38.2%	64.8%

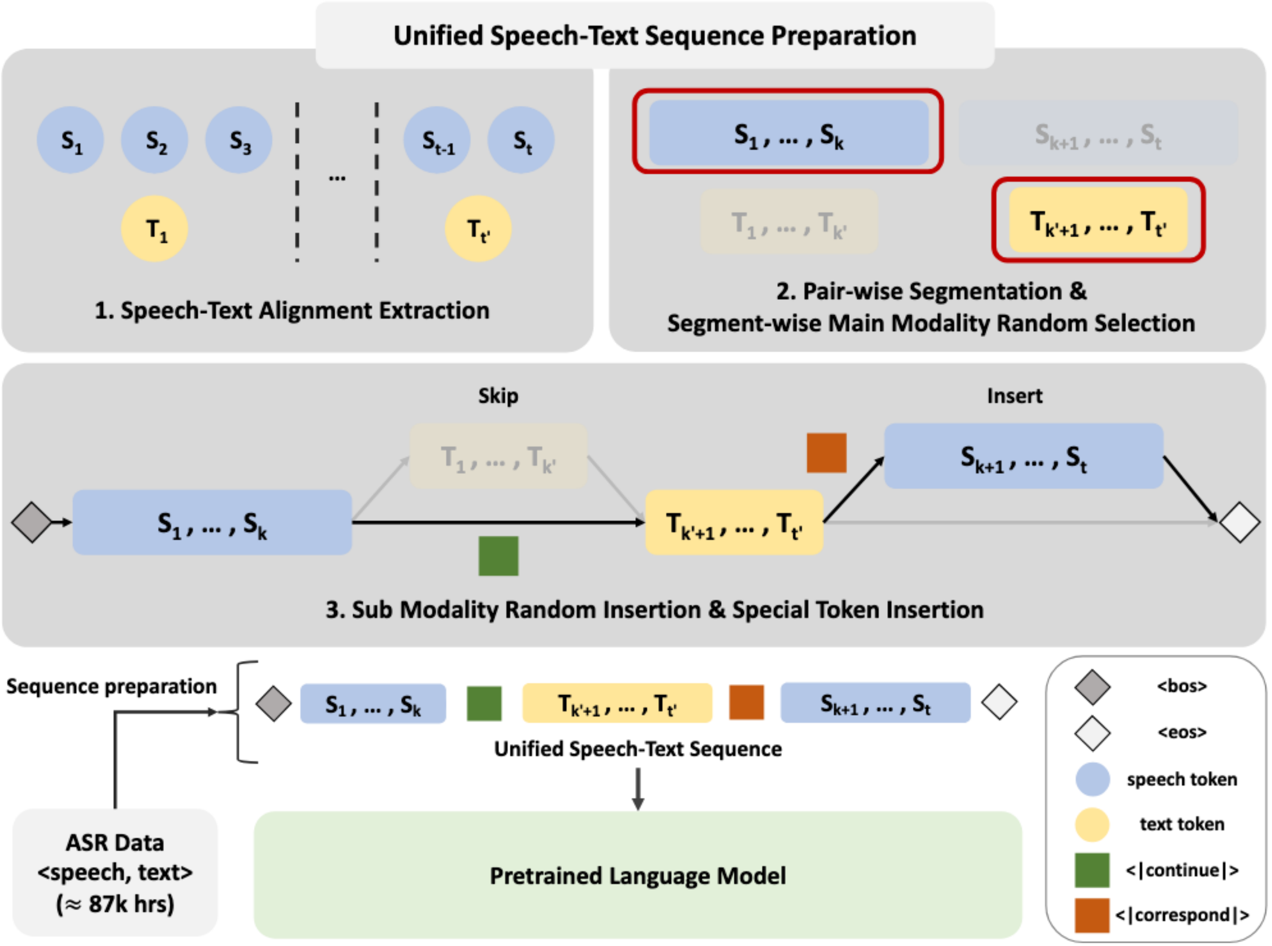
Token Classifier Accuracy Token Classifier Accuracy for Various Classes Related to Non-Verbal Cues

USDM – Unified Speech-Text Pre-training

- To extend a pre-trained LLM to speech modality, we propose an **interleaved sequence processing method** that contains **various relationships between the two modalities**.



We propose a **speech-text** sequence generation method that captures various relationships with two key relations: **correspondence** and **continuation**.



USDM – Supervised Fine-tuning on Spoken Dialog & Speech Reconstruction

- Following previous studies[4], we model spoken responses **indirectly through text** to leverage the **text capabilities of the pre-trained LLM**.

Below is a conversation between the user and the agent. Each turn includes the user's speech and its corresponding transcript, along with the agent's response text and the corresponding speech

User
speech token <|correspond|> text token
Agent
text token <|correspond|> speech token

: part where the loss is calculated.

- The generated speech tokens are converted back to audio using a **Voicebox-based**[5] **token-to-speech reconstruction model**.
  - Leveraging the **pronunciation** and **non-verbal cues** in the speech tokens for reconstruction.
  - **Given reference** speech, it additionally utilizes **timbre** information for personalized reconstruction.

Results (1)

- Comparison with **previous spoken dialog models**

Table 1: Human evaluation results of our model and the baselines. We report the MOS and P-MOS scores with a 95% confidence interval.

Method	**Overall**			**Acoustic**	
	win	***tie***	***lose***	**MOS**	**P-MOS**
Ground Truth	45.9%	8.0%	46.1%	4.51 ± 0.05	4.35 ± 0.05
USDM	–	–	–	4.31 ± 0.07	4.31 ± 0.06
Cascaded	55.3%	4.9%	39.8%	4.26 ± 0.07	4.22 ± 0.07
From Scratch	53.3%	7.6%	39.1%	3.71 ± 0.11	3.65 ± 0.10
SpeechGPT [25]	53.8%	6.9%	39.3%	4.08 ± 0.09	4.04 ± 0.08

Table 2: GPT-4 evaluation and quantitative results of our model and the baselines.

Method	**Semantic**					**WER**	
	win	***tie***	***lose***	**METEOR**	**ROUGE-L**	**STT**	**TTS**
Ground Truth	32.7%	19.6%	47.7%	–	–	–	2.2%
USDM	–	–	–	13.1	15.7	7.4%	2.0%
Cascaded	42.7%	24.6%	32.7%	12.5	15.0	3.8%	1.3%
From Scratch	79.7%	10.1%	10.2%	8.6	10.6	58.1%	64.0%
SpeechGPT [25]	61.0%	13.1%	25.9%	9.9	11.8	12.4%	23.2%

Results (2)

- Ablation studies for **pre-training** & **supervised fine-tuning**

Table 3: Results of the ablation studies on the pretraining and fine-tuning schemes. For PPL, we report the average PPL for each modality across the six combinations described in the text.

	Method	Pretraining		Spoken Dialog Modeling			
		Text PPL	Unit PPL	STT WER	TTS WER	METEOR	ROUGE-L
Pre-training	Ours	6.886	4.813	7.4%	2.0%	13.1	15.7
	Setup 1	14.485	5.261	57.8%	82.1%	8.9	10.6
	Setup 2	31.679	5.619	11.2%	2.5%	12.5	15.1
	Setup 3	21.392	5.146	7.3%	2.0%	12.7	15.4
Fine-tuning	S1 $\rightarrow$ S2	–	–	–	–	6.5	7.7

Conclusion

- USDM is a spoken dialog model that considers not only **content** but also **non-verbal elements**.

- The proposed **cross-modal pre-training** proved effective for spoken dialog and serves as a **foundational model capable of handling various tasks** through fine-tuning with diverse data and templates tailored to specific contexts.

Project Page

Paper

Code & Checkpoints

References

1. Jiang, Albert Q., et al. "Mistral 7B." *arXiv preprint arXiv:2310.06825* (2023).
2. Barrault, Loïc, et al. "SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. " *arXiv preprint arXiv:2308.11596* (2023).
3. Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., ... & Auli, M. (2021). XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.
4. Zhang, Dong, et al. "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities." *arXiv preprint arXiv:2305.11000* (2023).
5. Le, Matthew, et al. "Voicebox: Text-guided multilingual universal speech generation at scale." *Advances in neural information processing systems* 36 (2024).


Thank You!