SpeechGPT Series

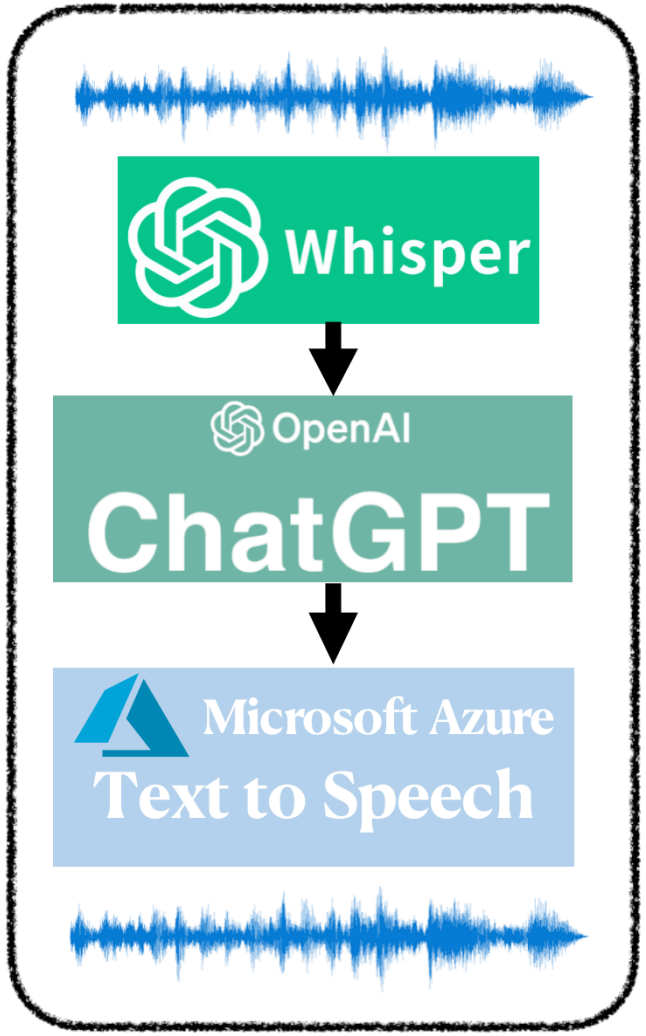
张栋

2024.06

https://0nutation.github.io/

Spoken Dialogue System Before GPT-4o

- **Cascaded System**
  - ASR + LLM + TTS
  - Easy to build 😊
  - Information loss & High latency 😫
- **End-to-end System**
  - Speech language model
  - Lower performance 😫



GPT-4o as an End-to-End Spoken Chatbot

Prior to GPT-4o, you could use Voice Mode to talk to ChatGPT with latencies of 2.8 seconds (GPT-3.5) and 5.4 seconds (GPT-4) on average. To achieve this, Voice Mode is a pipeline of three separate models: one simple model transcribes audio to text, GPT-3.5 or GPT-4 takes in text and outputs text, and a third simple model converts that text back to audio. This process means that the main source of intelligence, GPT-4, loses a lot of information—it can't directly observe tone, multiple speakers, or background noises, and it can't output laughter, singing, or express emotion.

With GPT-4o, we trained a single new model end-to-end across text, vision, and audio, meaning that all inputs and outputs are processed by the same neural network. Because GPT-4o is our first model combining all of these modalities, we are still just scratching the surface of exploring what the model can do and its limitations.

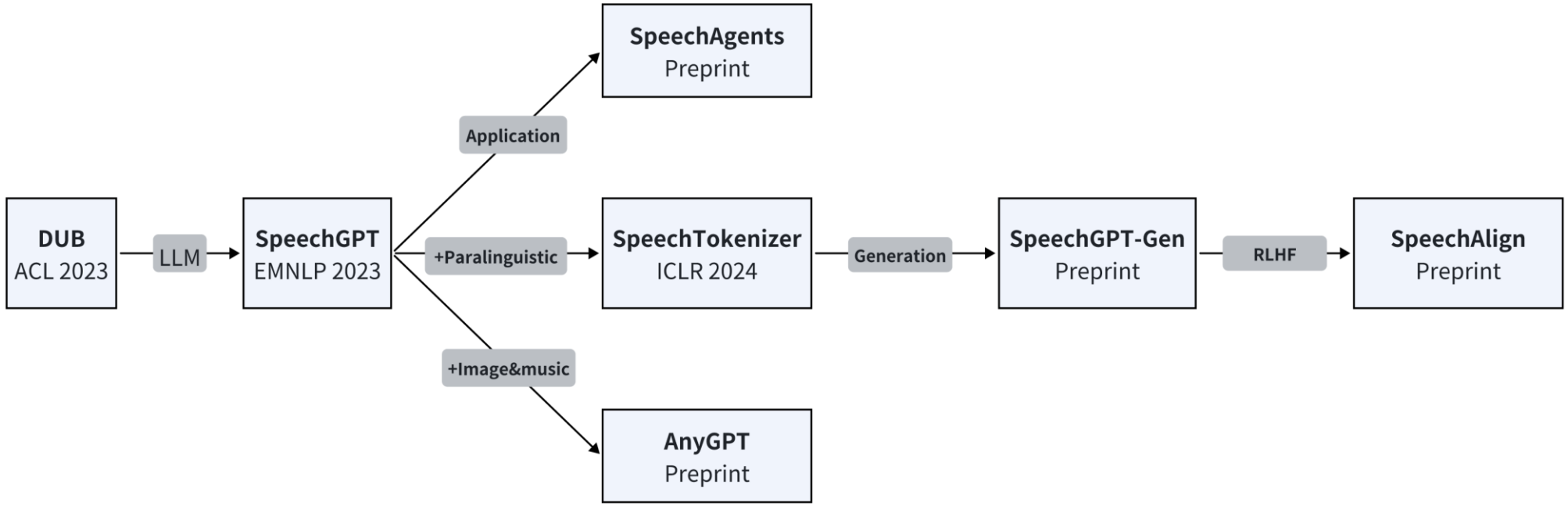
GPT-4o as a Human-Like Spoken Chatbot

- Emotional Intelligence
  - Emotion-awareness & Emotion-expression
- Conversational Expressions
  - Laughter, Sign, Back-channels, ...
- Low latency
  - Less than 300 ms

GPT-4o as a Conversational Speech Toolbox

- Human-Instruction Following
- Speech-to-speech Translation
- Voice Conversion
- ASR
- TTS
- ...

SpeechGPT series

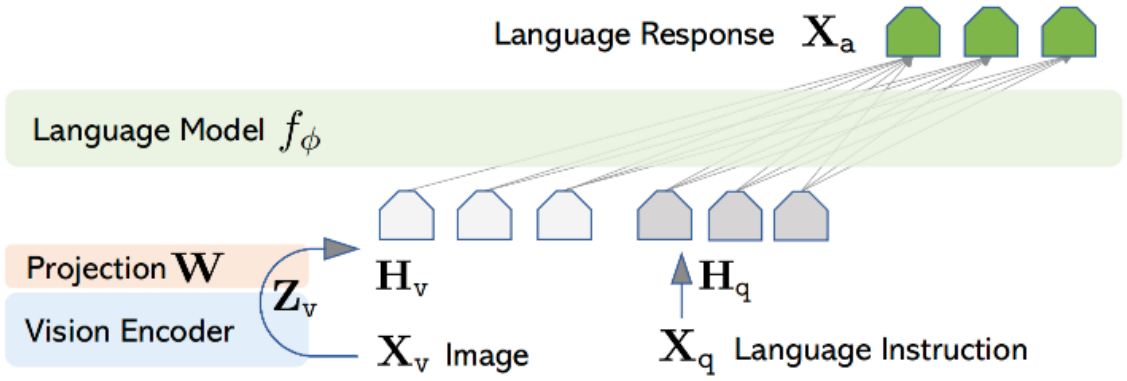


SpeechGPT:
Empowering Large Language Models with Intrinsic Cross-Modal Conversational Abilities

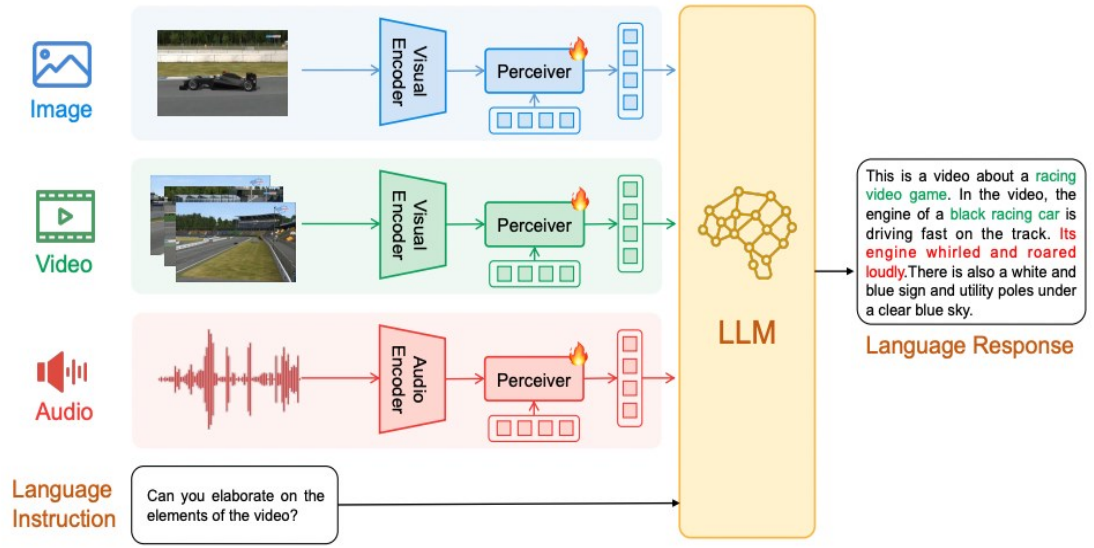
Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang,
Yaqian Zhou, Xipeng Qiu

EMNLP 2023

Background: Current Multi-modal Large Language Models



LLaVA[1]



ChatBridge[2]

- Modality-specific encoder
- Multi-modal in & text out
- Text instructions only
- Continuous representation

★ Lack of multimodal generative abilities

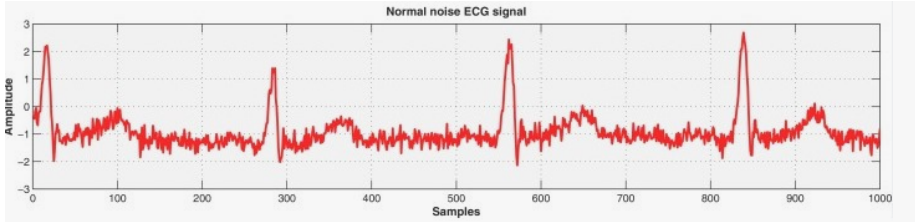
Background: Gap between General Modalities and LLM

Continuous Signals

Image/Video

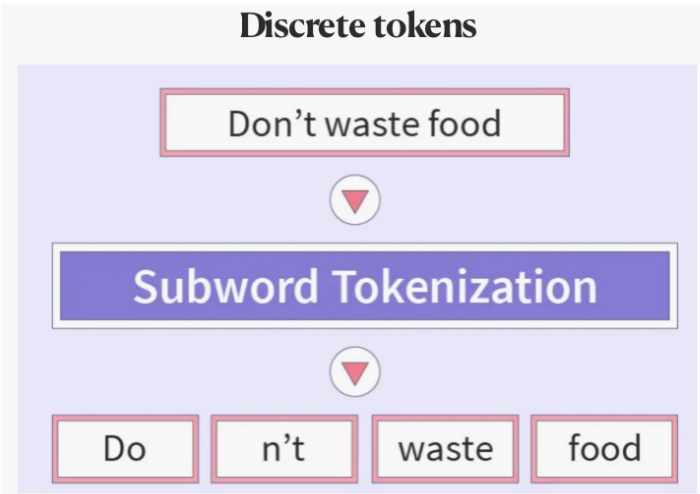
Sound

ECG signal



VS

LLM input&output



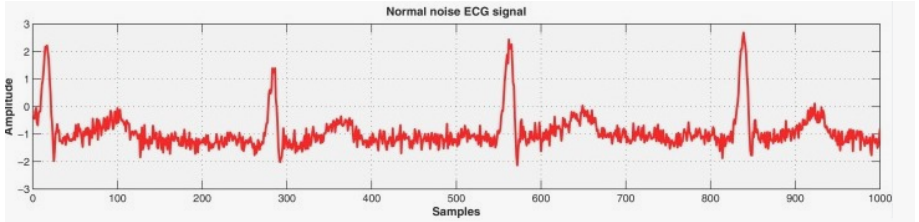
Background: Gap between General Modalities and LLM

Continuous Signals

Image/Video

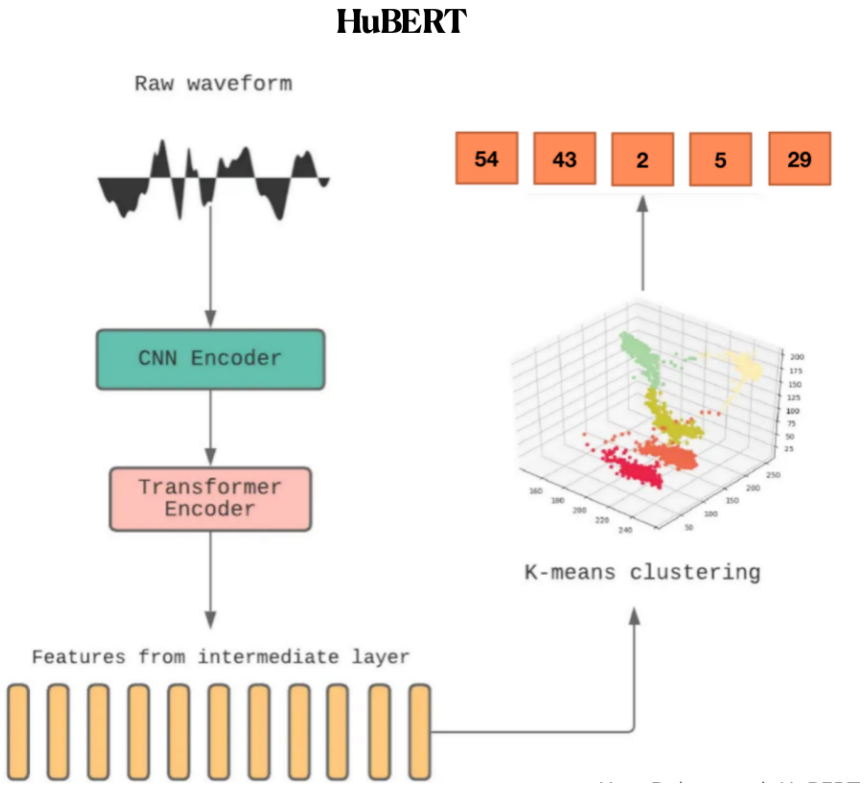
Sound

ECG signal



Discrete tokens ?

Background: Self-supervised Representation HuBERT

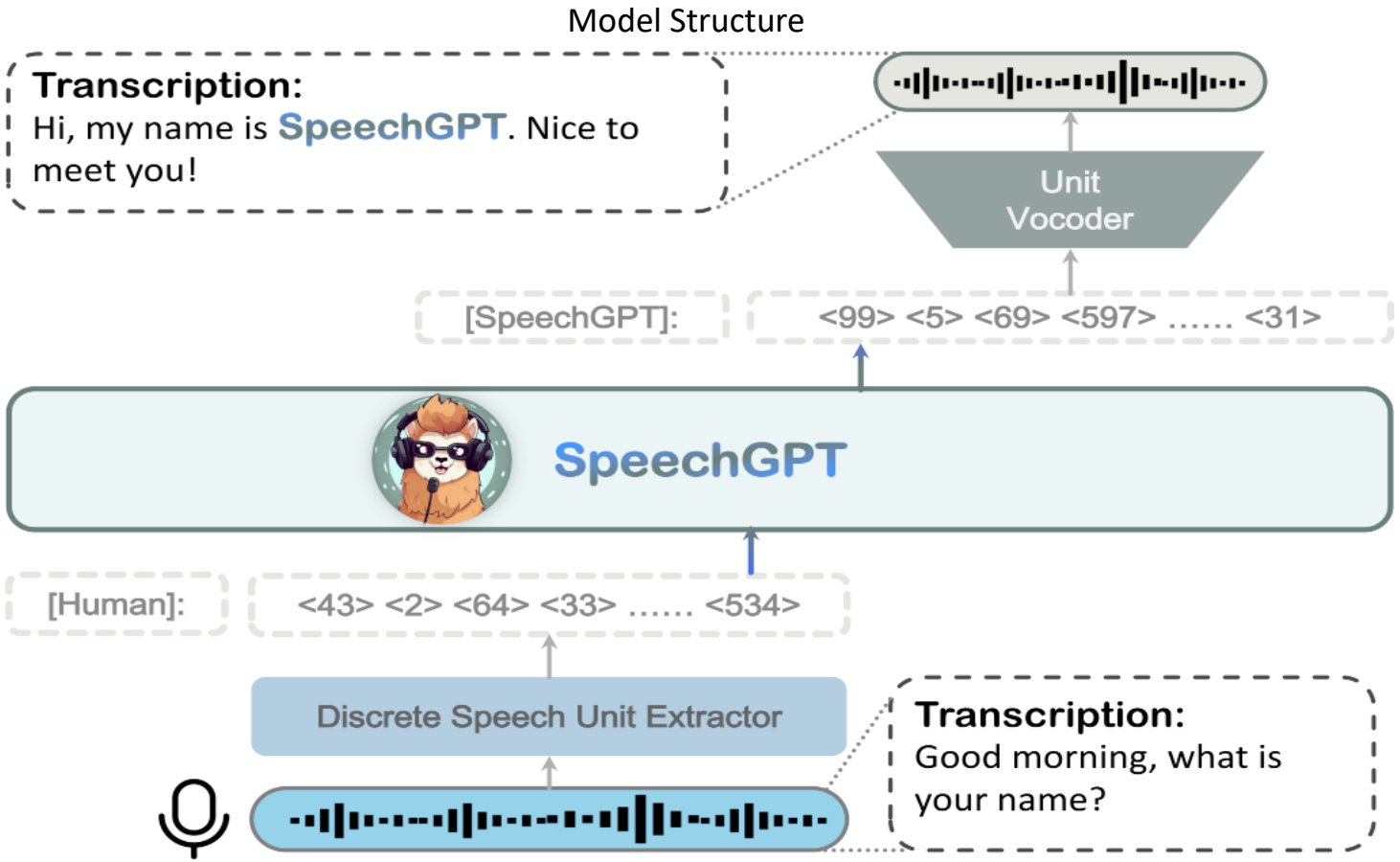


- Generate **discrete unit**:
  - k-means on the representation of the 9th Transformer layer.

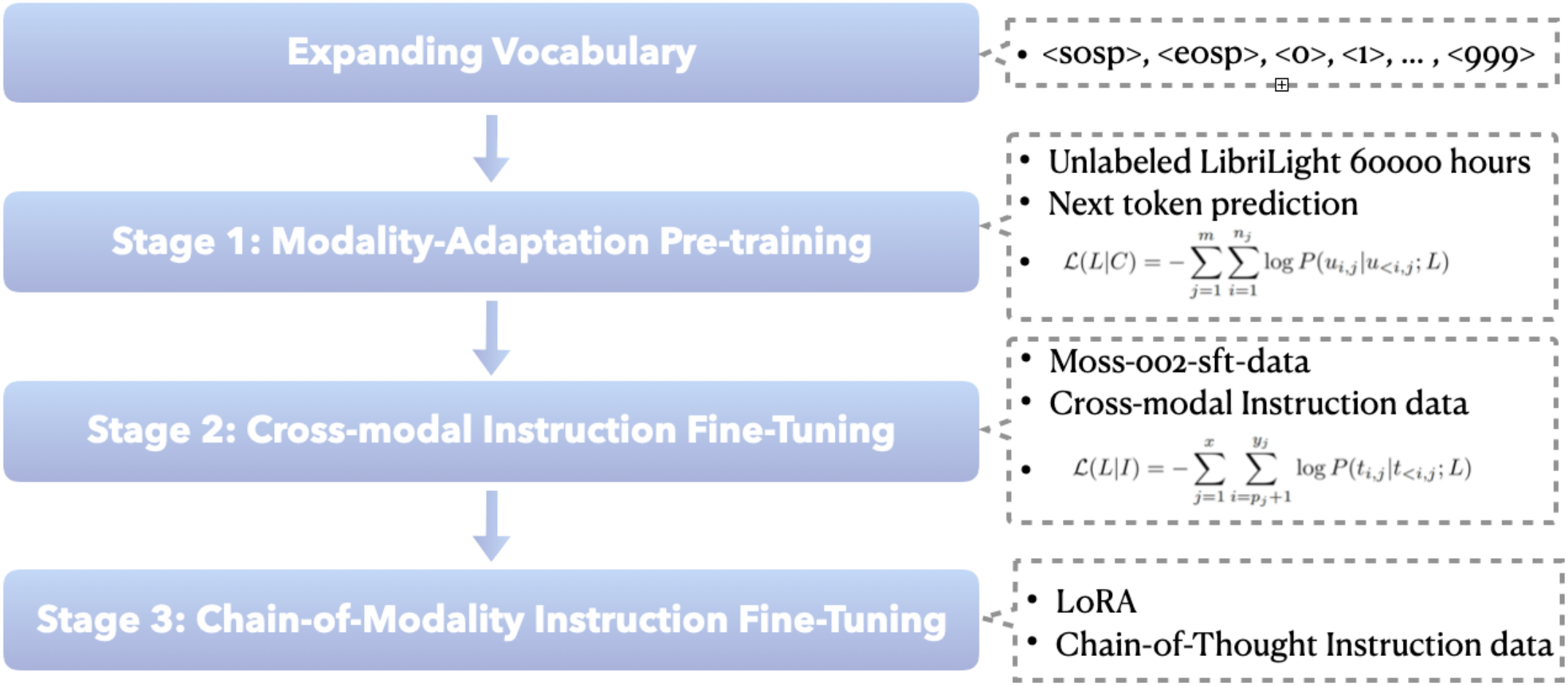
Speech as a foreign language !

SpeechGPT Overview

- Discrete Speech Unit Extractor
  - Speech to unit
  - Fixed HuBERT
- Large Language Model
  - From LLaMA-13B
  - Perceive & generate
  - Text & unit
- Unit Vocoder
  - Unit to speech
  - Stack of CNN layers



SpeechGPT Training



SpeechGPT demo

Cross-modal Instruction Following:

Instruction	Input	Output
Can you transcribe the speech into a written format?		I'm afraid there are no signs here said he.
Would you mind speaking these words as naturally as possible?	Today is a sunny day and I'm happy to be here.	

Spoken Dialogue:

Input	Output	Text output
		The capital of France is Paris. It is located in the northern part of the country, along the Seine River.
		The sun is out, the sky is blue, The flowers bloom and the birds sing too, The days are long, the nights are short, The spring is here and it's so sweet and warm!
		Sure, the correct pronunciation of apple is [app-ull].

SpeechGPT limitations

SpeechGPT can't support speech tasks related to **paralinguistic information**, such as speaker recognition, emotion recognition, etc.

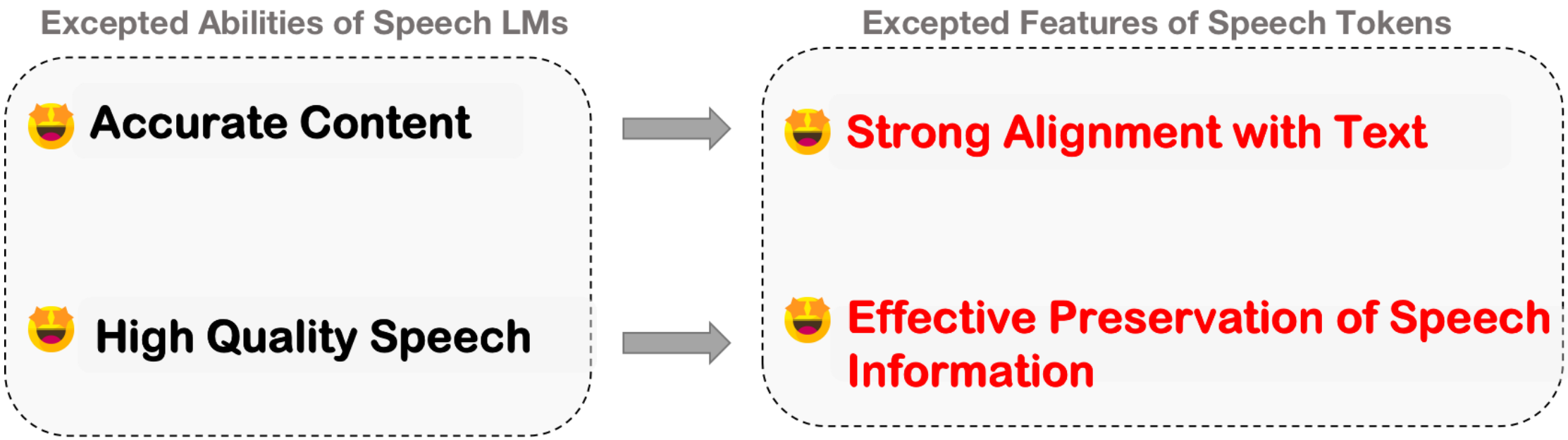
caused by

The HuBERT tokens **lose a lot of speaker information, emotional information, prosody information**, etc

next step

Better speech tokenizer !

Discussion: Ideal Speech Tokenizer for Speech Language Model



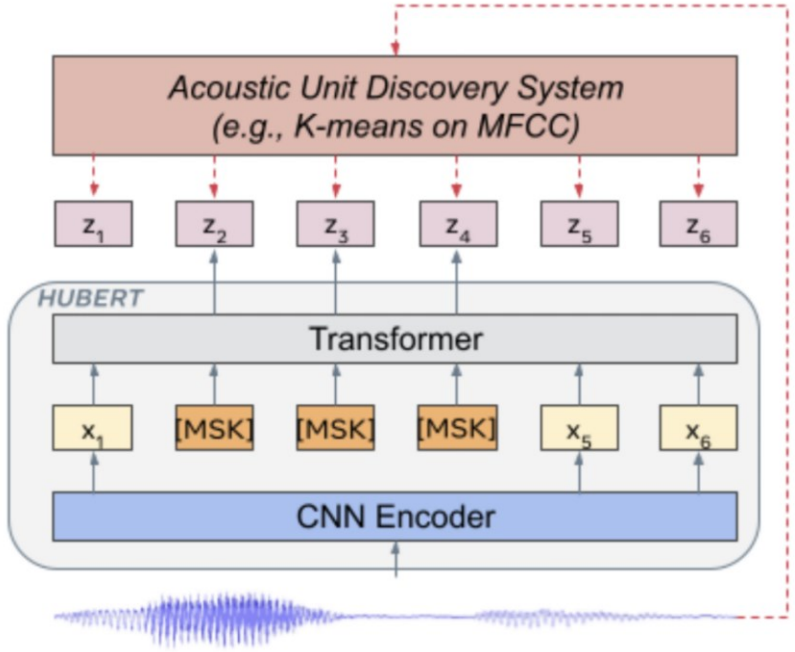
SpeechTokenizer: Unified Speech Tokenizer for Speech Language Models

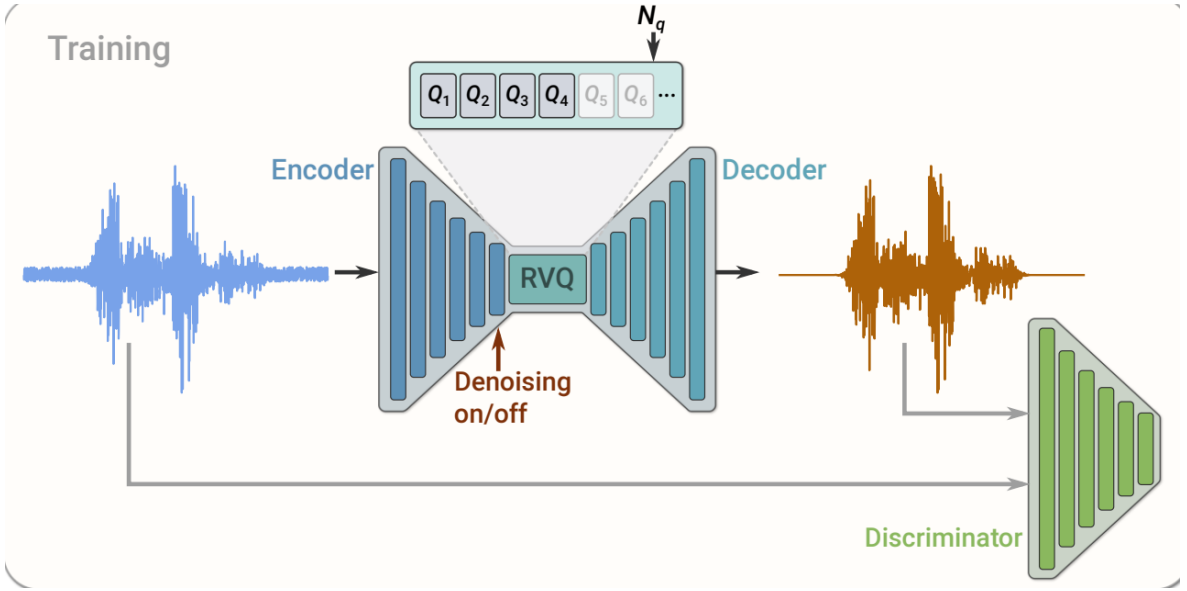
Dong Zhang[*], Xin Zhang[*(random order)], Shimin Li,
Yaqian Zhou, Xipeng Qiu

ICLR 2024

Background: Discrete Speech Representations

- Semantic Tokens (Hubert, w2v-bert)
  - Architecture: Transformer/Conformer
  - Training Objective: MLM/Contrastive Learning
  - Information: Long-term semantic information
- Acoustic Tokens (SoundStream, Encodec)
  - Architecture: RVQ-GAN, CNN based Encoder-Decoder
  - Training Objective: Speech Reconstruction
  - Discretization: RVQ
  - Information: Local acoustic information





How far are we from unified speech tokenizer?

SLMTokBench: Speech Language Model Token Benchmark

- **Text Alignment Evaluation**
  - Mutual information estimation between speech tokens & text
    - vCLUB mutual information estimation
  - Downstream CTC-ASR performance
    - Speech tokens as inputs
    - Two-layer 1024-unit BiLSTM optimized by CTC loss
- **Information Preservation Evaluation**
  - Content & Timbre
    - WER & Speaker Similarity between Raw speech and Reconstructed speech

How far are we from unified speech tokenizer?

SLMTokBench: Speech Language Model Token Benchmark

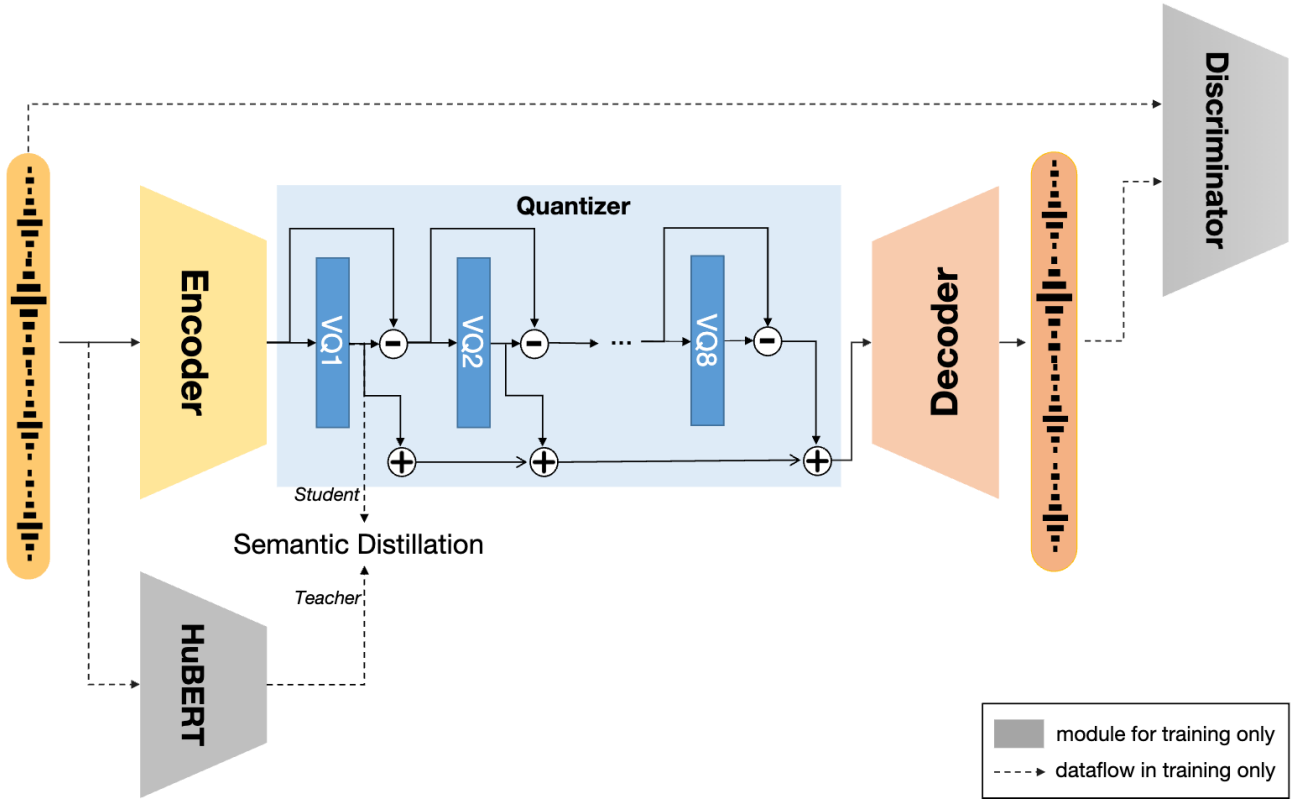
Tokenizer		Teacher	Text Alignment MI↑	Text Alignment WER† ↓	Information Preservation WER* ↓	Information Preservation SIM↑
Groundtruth			-	-	4.58	1.0
HuBERT	KM500	-	31.2	9.88	16.26	0.77
EnCodec	RVQ-1:8	-	23.6	30.91	5.11	0.98

- Semantic tokens align with text well but lose timbre information.
- Acoustic tokens preserve all information but align with text badly.

SpeechTokenizer: First Unified Speech Tokenizer

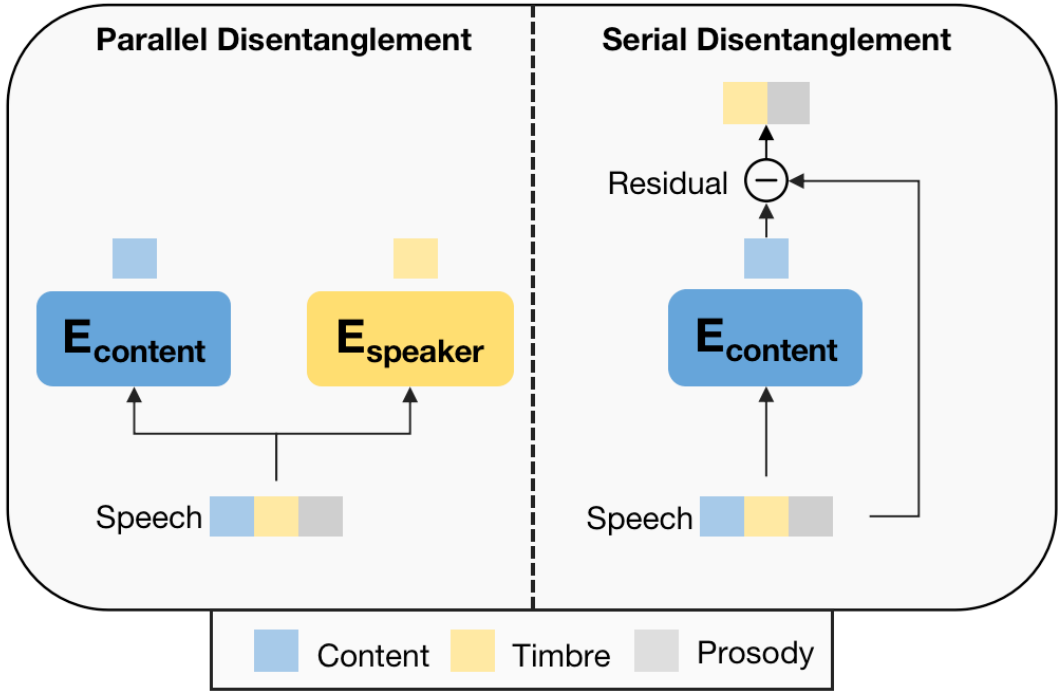
SpeechTokenizer: Disentangle different speech information hierarchically across different RVQ layers

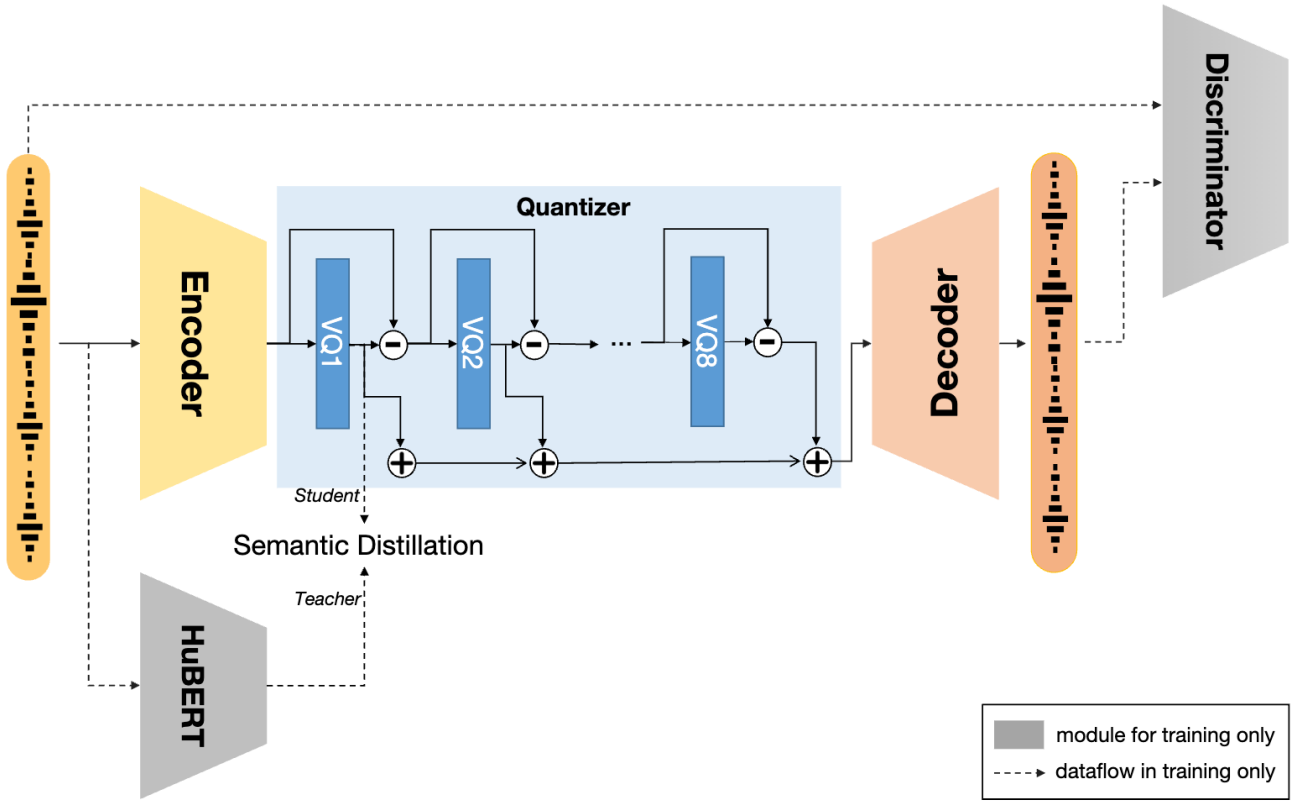
- **RVQ-GAN Architecture**
  - Convolutional-based encoder-decoder network from EnCodec
  - Resiudal vector quantization
  - Dicriminators from Hifi-Codec
- **Semantic Distillation on the first RVQ layer**
  - HuBERT as semantic teacher
  - Continuous representation distillation
  - Pseudo-label prediction
- **Serial Disentanglement:**
  - RVQ-1 contains content information
  - Subsequent RVQ-2:8 complement the remaining paralinguistic information



SpeechTokenizer: First Unified Speech Tokenizer

SpeechTokenizer: Disentangle different speech information hierarchically across different RVQ layers

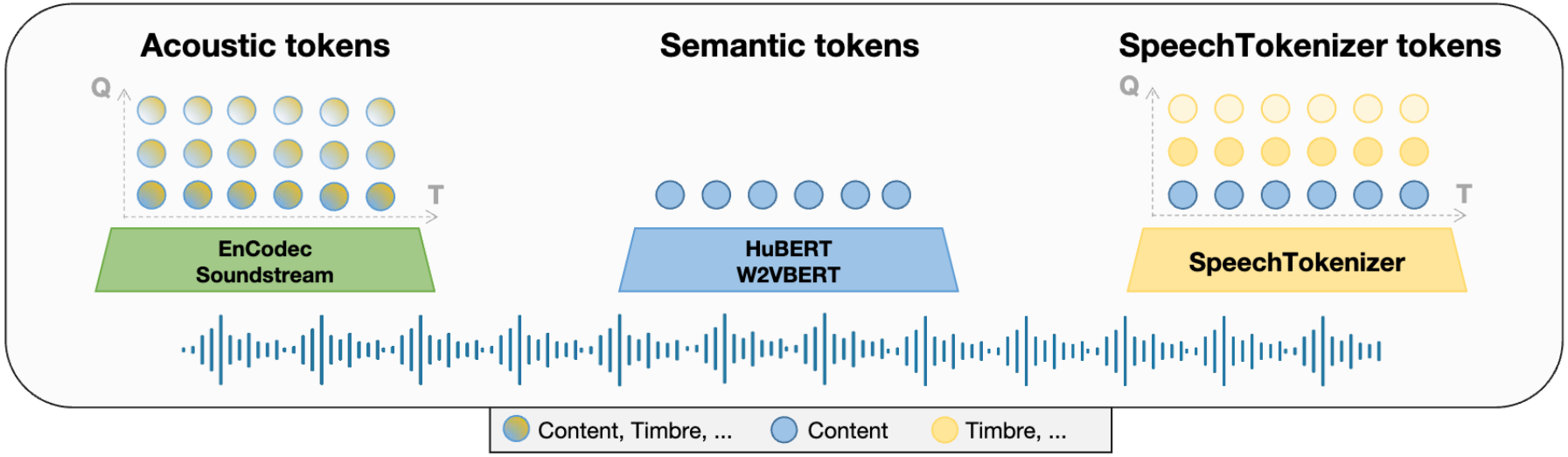




- **Serial Disentanglement:**
  - RVQ-1 contains content information
  - Subsequent RVQ-2:8 complement the remaining paralinguistic information

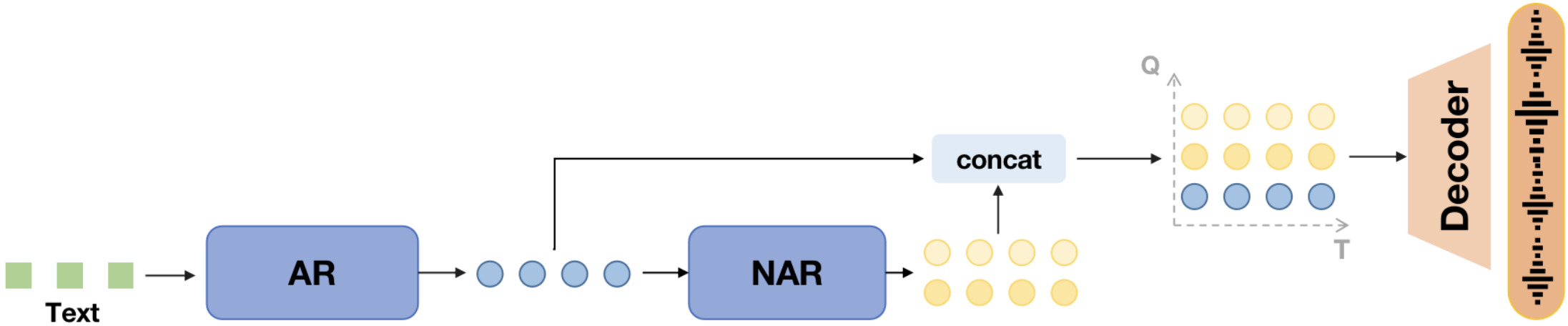
SpeechTokenizer: First Unified Speech Tokenizer

SpeechTokenizer: Disentangle different speech information hierarchically across different RVQ layers



USLM: Unified Speech Language Model

- Based on SpeechTokenizer, we unify Speech Language Model.
  - Autoregressive LM: text -> RVQ-1 tokens
  - Non-autoregressive LM: RVQ-1 tokens -> RVQ-2:8 tokens



SpeechTokenizer Evaluation

- **Speech Reconstruction Evaluation**
  - **Test set:**
    - 300 random samples from LibriSpeech test set
  - **Objective Metrics**
    - ViSQOL
    - WER
  - **Subjective Metrics**
    - MUSHRA

Tokenizer	Objective		Subjective
	WER↓	VISQOL↑	MUSHRA↑
Groundtruth	4.58	-	91.46
EnCodec	5.11	4.37	79.86
SpeechTokenizer	5.04	4.30	90.55

Tokenizer	Raw speech	RVQ-1	RVQ-2:8	RVQ-1:8
EnCodec				
SpeechTokenizer				

SpeechTokenizer Evaluation

➢ **Performance on SLMTokBench**

Tokenizer		Teacher	Text Alignment MI↑	Text Alignment WER† ↓	Information Preservation WER* ↓	Information Preservation SIM↑
Groundtruth			-	-	4.58	1.0
HuBERT	KM500	-	31.2	9.88	16.26	0.77
EnCodec	RVQ-1	-	16.5	61.52	38.34	0.92
EnCodec	RVQ-1:8	-	23.6	30.91	5.11	0.98
Ablations						
SpeechTokenizer	RVQ-1	HuBERT avg	30.9	15.58	9.57	0.74
SpeechTokenizer	RVQ-1:8	HuBERT avg	29.7	16.03	5.04	0.97
SpeechTokenizer	RVQ-1	HuBERT L9	32.9	12.68	14.17	0.73
SpeechTokenizer	RVQ-1:8	HuBERT L9	31.6	13.12	5.31	0.97
SpeechTokenizer	RVQ-1	HuBERT units	24.2	34.13	20.02	0.72
SpeechTokenizer	RVQ-1:8	HuBERT units	25.1	30.71	5.84	0.95

- **SpeechTokenizer can align with text well and effectively preserve speech information.**

USLM Evaluation

- **Zero-shot Text-to-Speech**
  - **Test set:**
    - VCTK test set
  - **Objective Metrics**
    - WER
    - Speaker Similarity
  - **Subjective Metrics**
    - MOS
    - Similarity MOS

Model	Tokenizer	Objective		Subjective	
		WER↓	SIM↑	MOS↑	SMOS↑
Groundtruth		1.9	0.93	4.5	3.96
VALL-E	EnCodec	7.9	0.75	3.08	3.31
USLM	SpeechTokenizer	**6.5**	**0.84**	**3.63**	**3.45**

Table 4: Results of zero-shot TTS

Text	Speaker Prompt	GroundTruth	Vall-E	USLM
She also defended the lord chancellors existing powers.				

Effectiveness of Serial Information Disentanglement

- **Voice Conversion by Token Mixing**

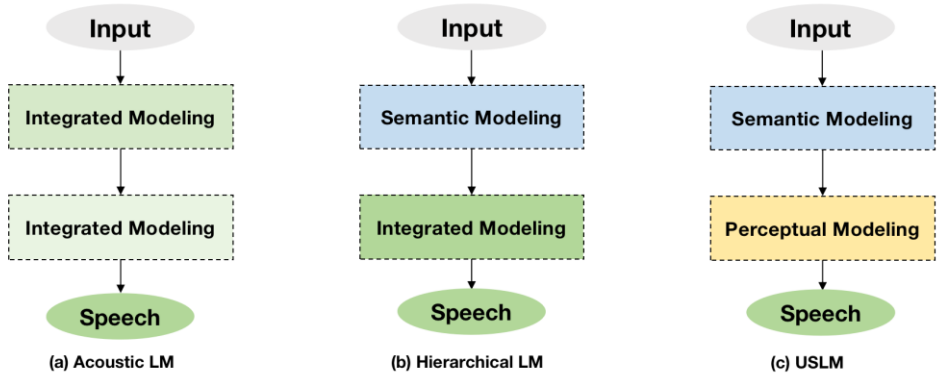
  1. Transform the source speech and reference speech into SpeechTokenizer token matrices
  2. Concatenate the RVQ-1 tokens of source token matrix with RVQ-2:8 tokens of the reference token matrix (truncation or circular padding)
  3. Pass this combined token matrix to the decoder

Source	Reference	WER↓	SIM↑
Groundtruth		0.4	0.93
RVQ-1	RVQ-2	2.6	0.72
RVQ-1	RVQ-2:4	11.7	0.80
RVQ-1	RVQ-2:8	35.4	0.82

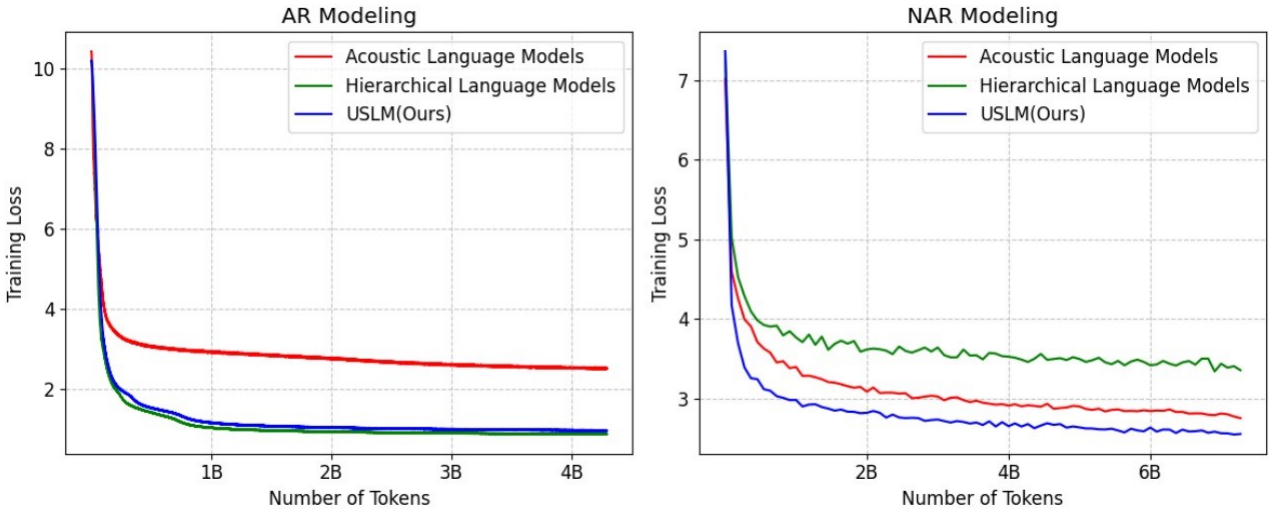
Source Speech	Reference Speech	Converted Speech
		

Effectiveness and Efficiency of USLM

- **Zero-shot Text-to-Speech**
  - **Data:**
    - Mulitlingual LibriSpeech
  - **Model Structure:**
    - AR + NAR (SoundStorm)
  - **Acoustic Language Models**
    - Tokenizer: Encodec
    - AR: text toEnCodec RVQ-1
    - NAR: EnCodec RVQ-1 to EnCodec RVQ-2:8
  - **Hierarchical Language Models**
    - Tokenizer: HuBERT & Encodec
    - AR: text toHuBERT tokens
    - NAR: HuBERT tokens to EnCodec RVQ-1:8
  - **USLM**
    - Tokenizer: SpeechTokenizer
    - AR: text toSpeechTokenizer RVQ-1
    - NAR: SpeechTokenizer RVQ-1 to SpeechTokenizer RVQ-1:8

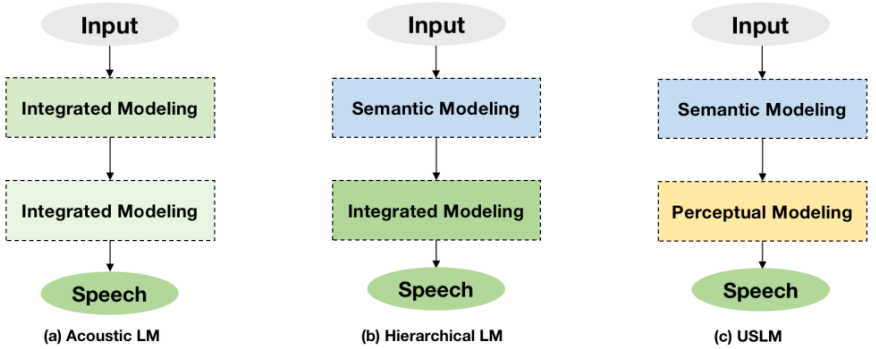


Model Convergence Comparsion

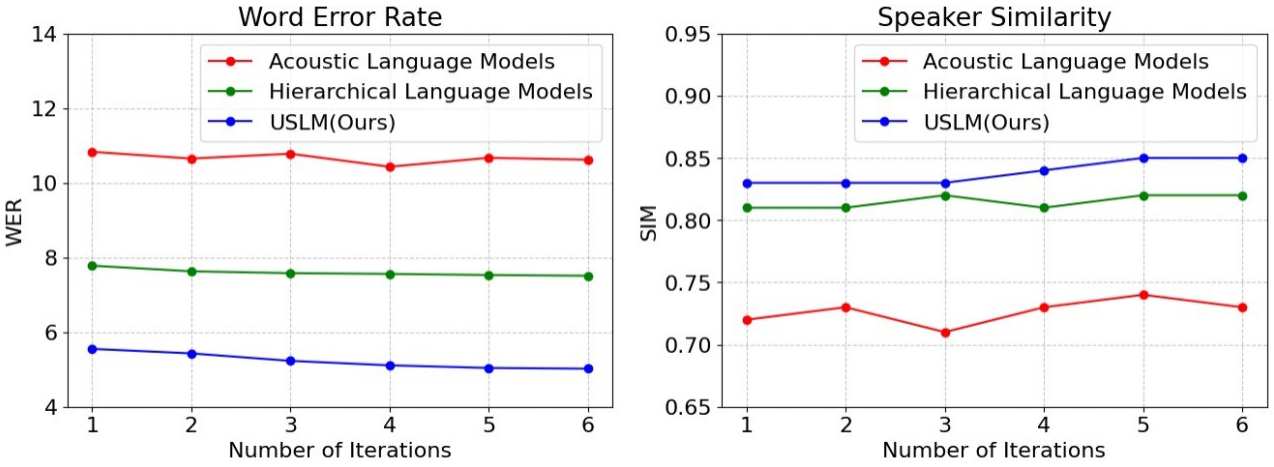


Effectiveness and Efficiency of USLM

- **Zero-shot Text-to-Speech**
  - **Data:**
    - Mulitlingual LibriSpeech
  - **Model Structure:**
    - AR + NAR (SoundStorm)
  - **Acoustic Language Models**
    - **Tokenizer**: Encodec
    - **AR**: text toEnCodec RVQ-1
    - **NAR**: EnCodec RVQ-1 to EnCodec RVQ-2:8
  - **Hierarchical Language Models**
    - **Tokenizer**: HuBERT & Encodec
    - **AR**: text toHuBERT tokens
    - **NAR**: HuBERT tokens to EnCodec RVQ-1:8
  - **USLM**
    - **Tokenizer**: SpeechTokenizer
    - **AR**: text toSpeechTokenizer RVQ-1
    - **NAR**: SpeechTokenizer RVQ-1 to SpeechTokenizer RVQ-1:8

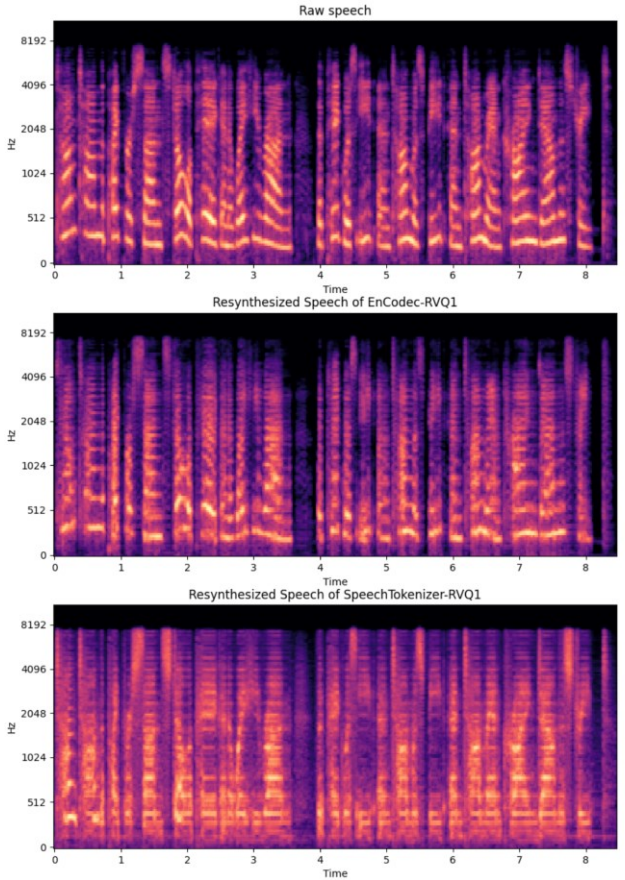


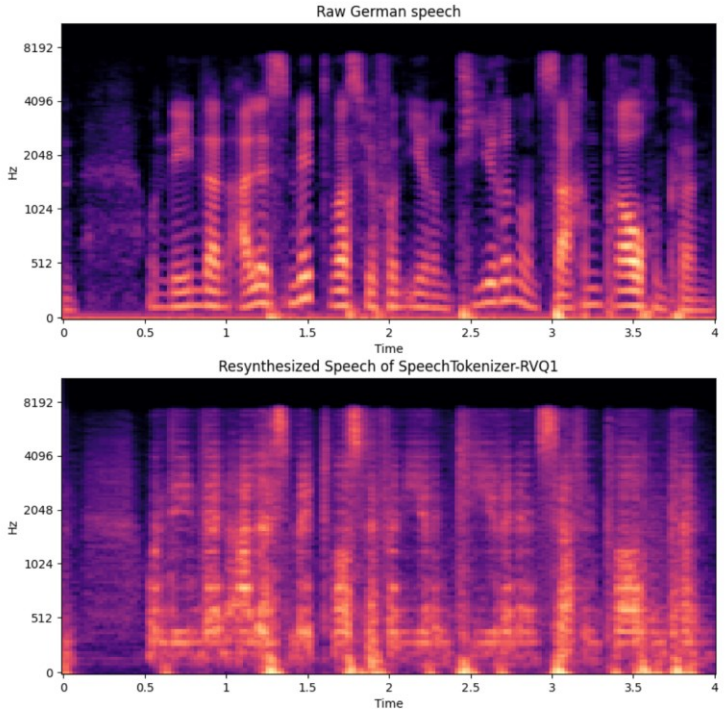
Downstream Performance Comparsion



Analysis Experiment: Extension to Unseen Language & MelSpectorgram Analysis

Language	Raw speech	RVQ-1	RVQ-1:8
Chinese			
German			



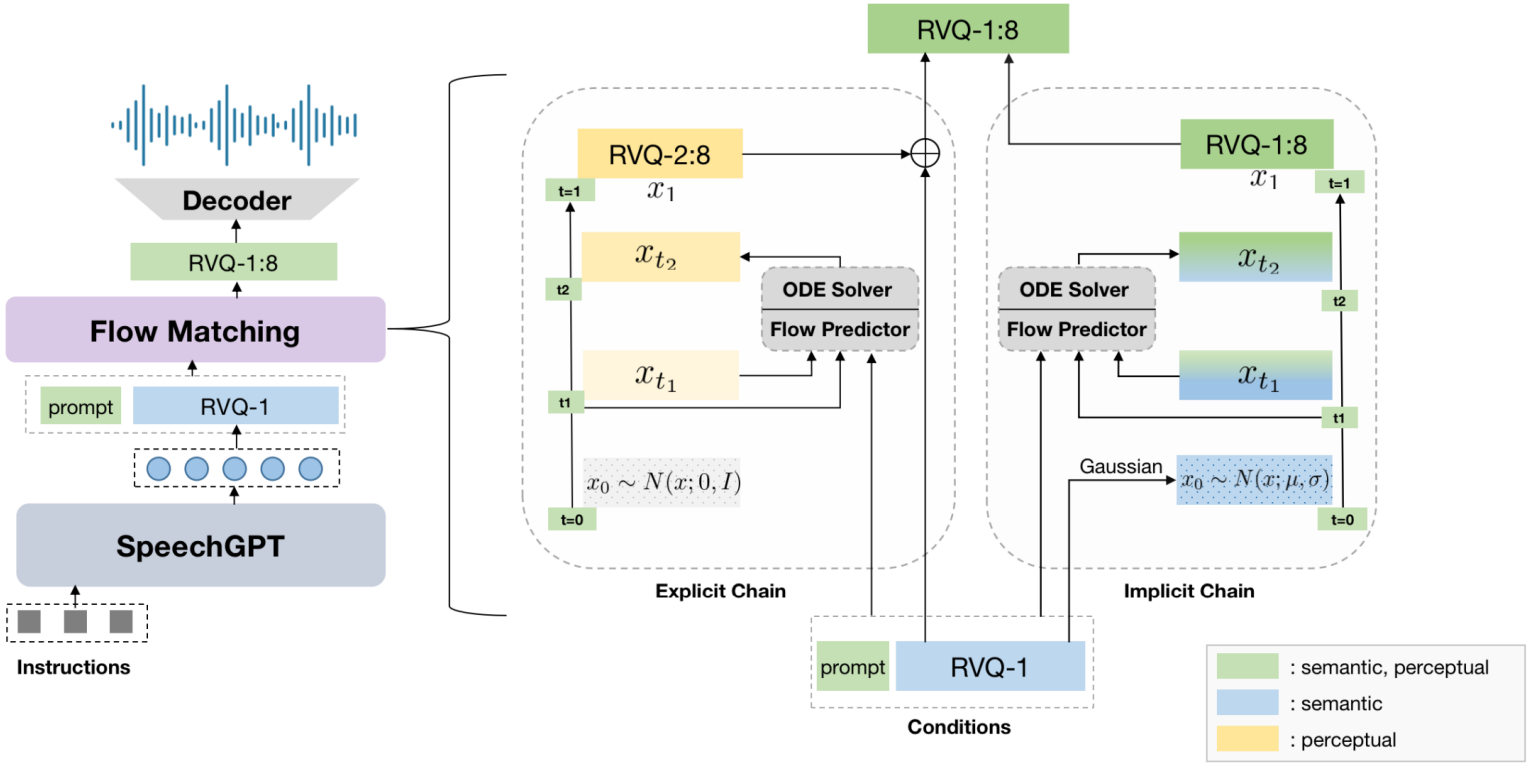


SpeechGPT-Gen: Scaling Chain-of-Information Speech Generation

Dong Zhang[*], Xin Zhang[*], Jun Zhan, Shimin Li,
Yaqian Zhou, Xipeng Qiu

SpeechGPT-Gen: Scaling USLM to 8 Billion Parameters

- **Tokenizer:**
  - SpeechTokenizer
- **AR:**
  - SpeechGPT (7B)
- **NAR:**
  - Flow Matching (1B)

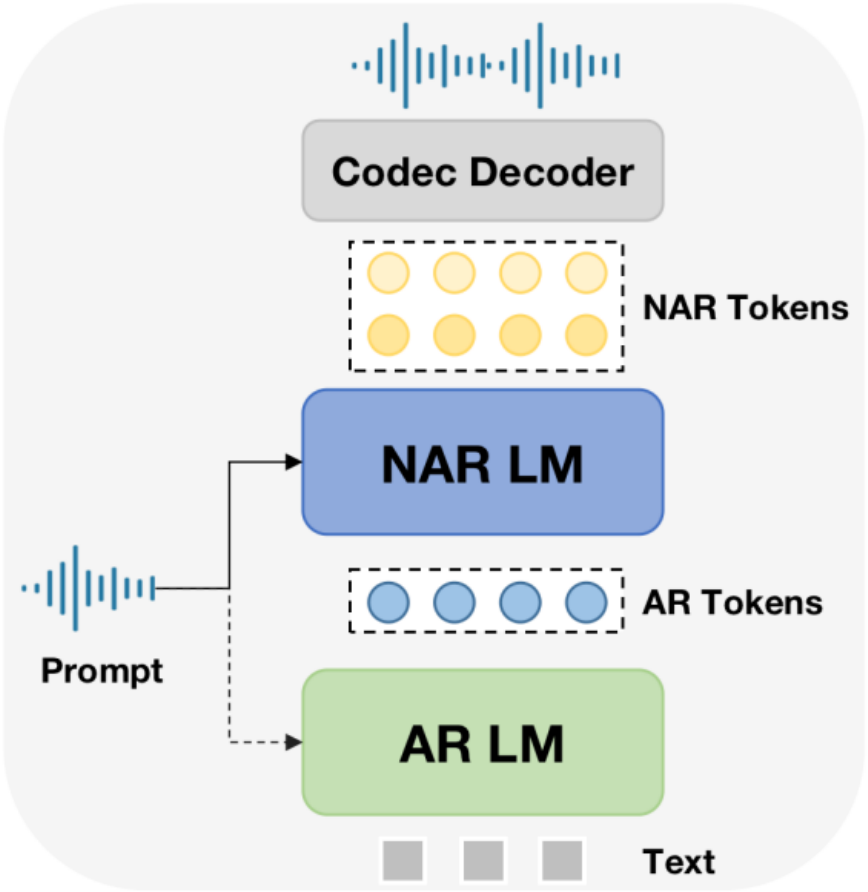


Speech Instruction	Speech Prompt	Speech Response

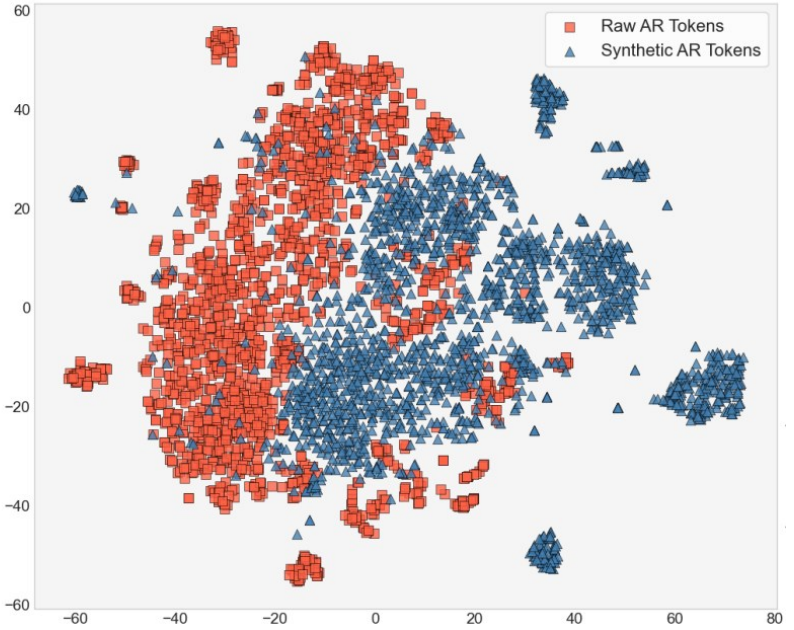
SpeechAlign: Aligning Speech Generation to Human Preferences

Dong Zhang[*], Zhaowei Li[*], Shimin Li, Xin Zhang, Pengyu Wang,
Yaqian Zhou, Xipeng Qiu

Observation: Distribution Gap Degrades Performance of Codec LM



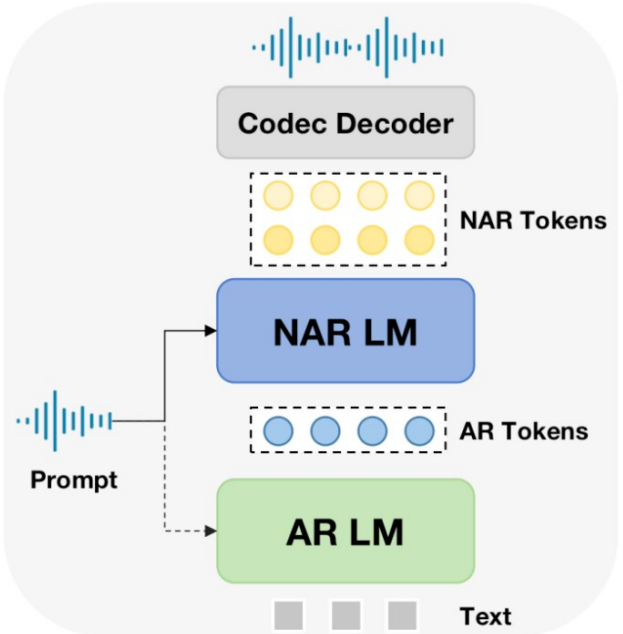
(a) Inference Process of Codec Language Models



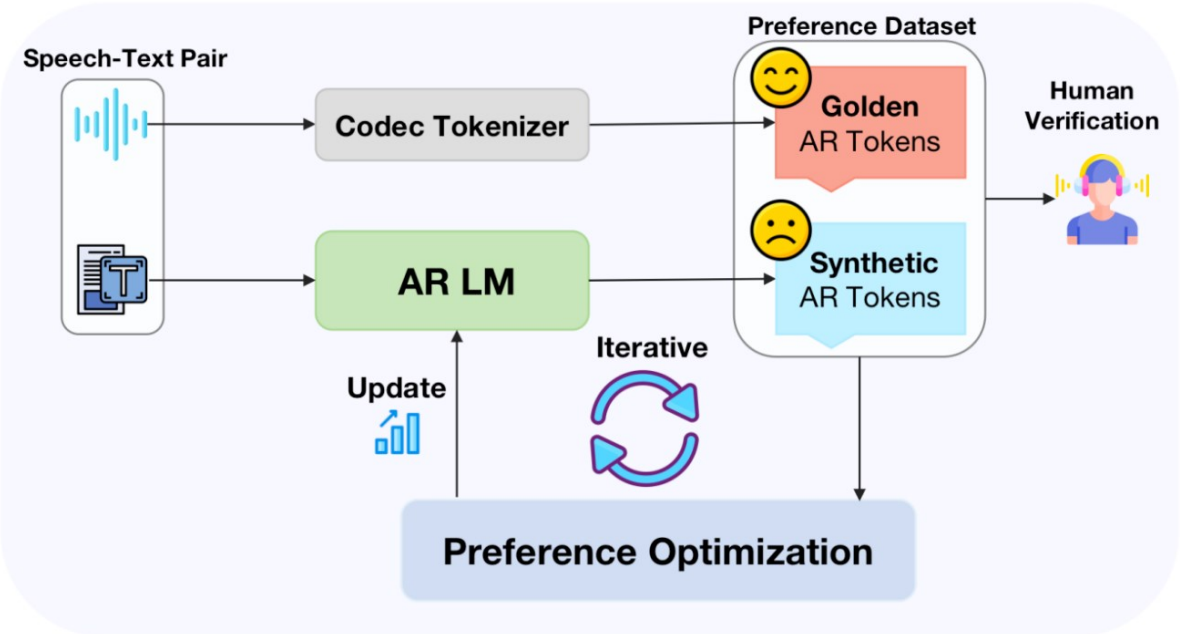
(a) Raw AR Tokens and Synthetic AR Tokens

Input	WER(↓)	SIM(↑)
Groundtruth	3.4	-
Golden AR tokens	5.9	0.93
Synthetic AR tokens	7.2	0.87

SpeechAlign: Align Speech Generation to Human Preferences



(a) Inference Process of Codec Language Models



(b) SpeechAlign

SpeechAlign: Align Speech Generation to Human Preferences

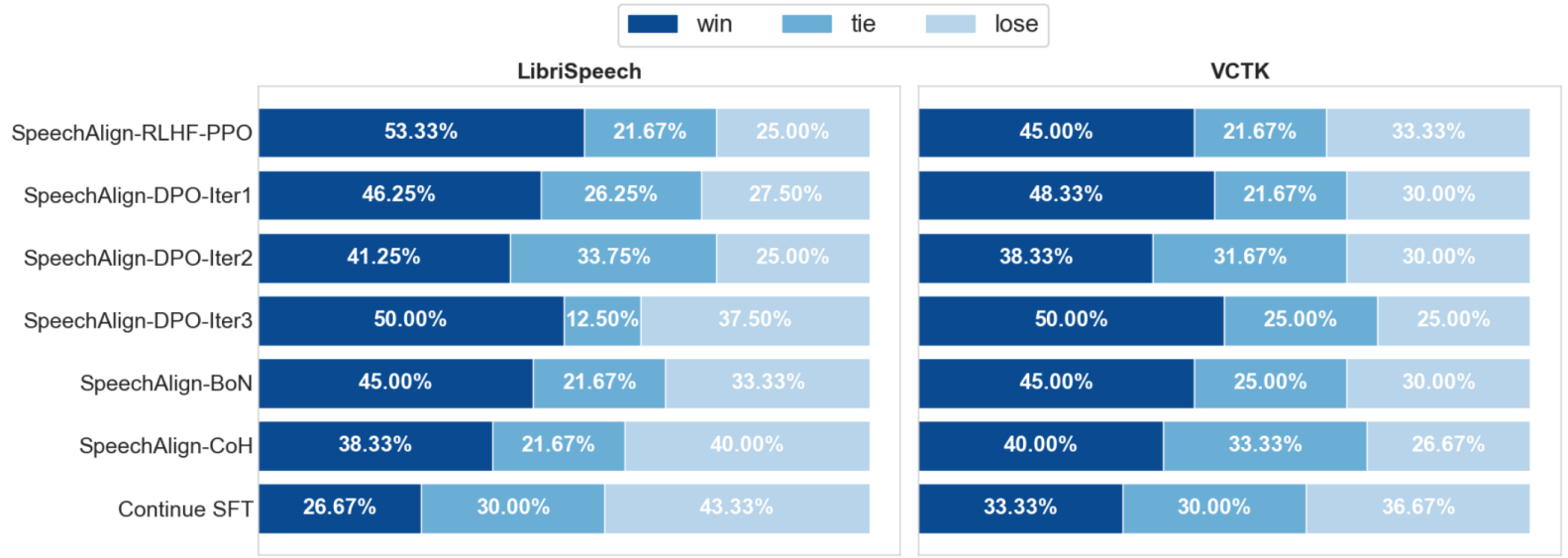
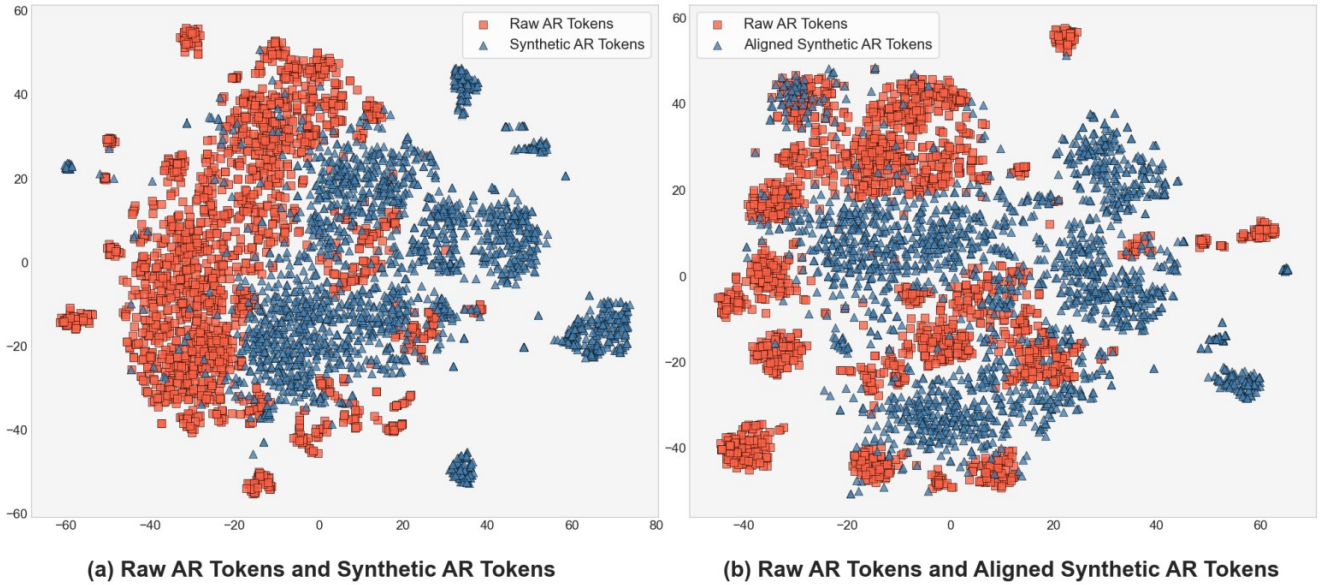


Figure 1: Qualitative side-by-side comparsion results of preference optimized models versus the baseline SFT model on zero-shot text-to-speech performance. SpeechAlign-RLHF-PPO denotes models optimized by RLHF using PPO algorithm. SpeechAlign-DPO-Iter1 denotes models optimized by Direct Preference Optimization method at the first iteration. SpeechAlign-DPO-Iter2 and SpeechAlign-DPO-Iter3 denote the models optimized at the second and third iterations, respectively. SpeechAlign-CoH represents models optimized by Chain-of-Hindsight strategy. SpeechAlign-BoN refers to baseline SFT model employing Best-of-N sampling method. SpeechAlign-BoN, SpeechAlign-RLHF-PPO and SpeechAlign-DPO series models significantly outperform baseline model on both LibriSpeech and VCTK dataset.

SpeechAlign: Align Speech Generation to Human Preferences



(a) Raw AR Tokens and Synthetic AR Tokens

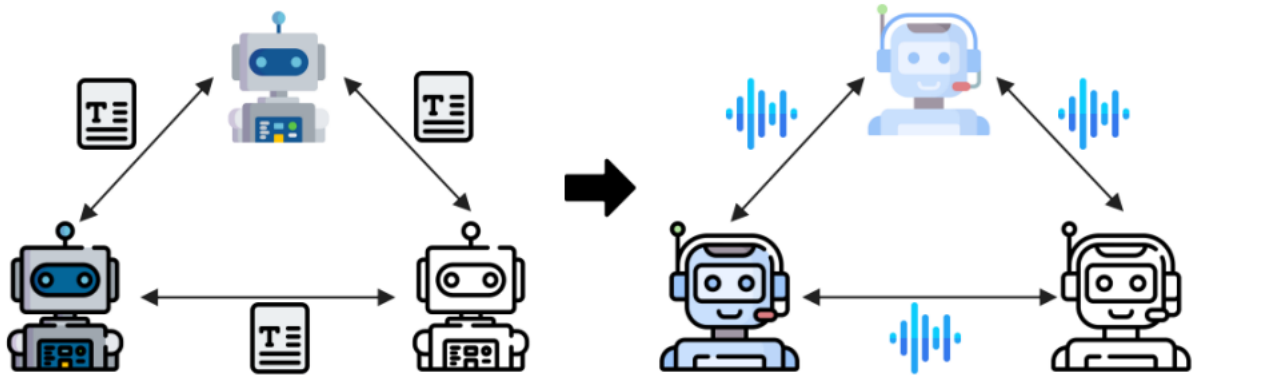
(b) Raw AR Tokens and Aligned Synthetic AR Tokens

Figure 2: T-SNE visualization of representations of different AR tokens. **Left**: Golden AR tokens and synthetic AR tokens. **Right**: Golden AR tokens and aligned synthetic AR tokens.

SpeechAgents: Human-Communication Simulation with Multi-Modal Multi-Agent Systems

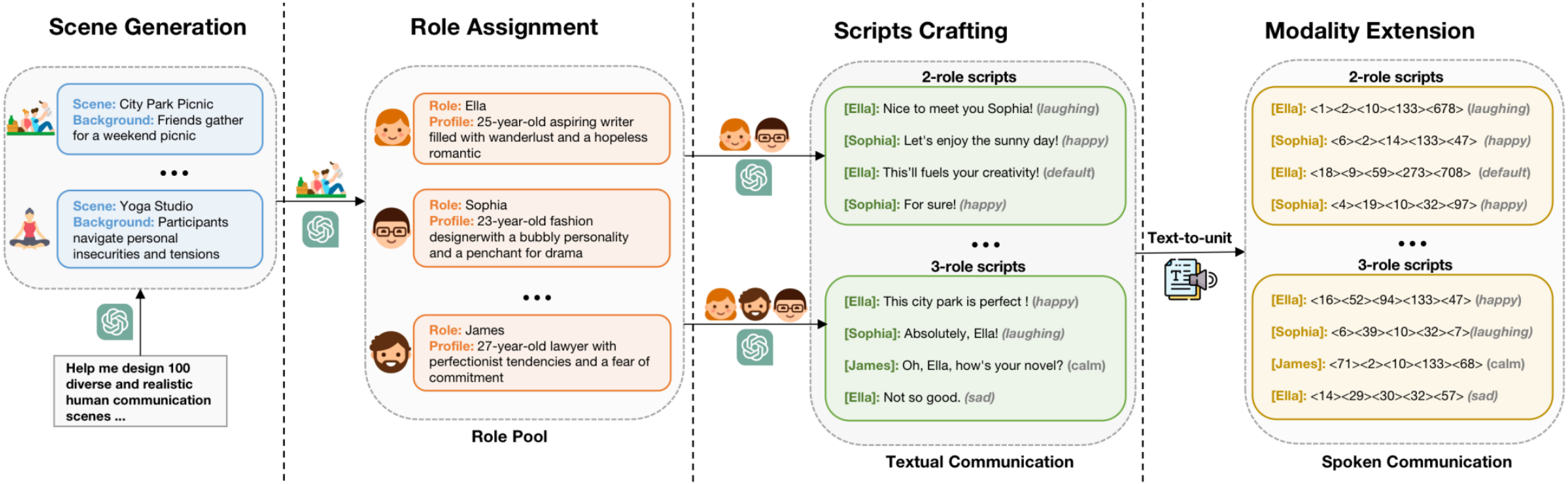
Dong Zhang, Zhaowei Li, Pengyu Wang, Xin Zhang,
Yaqian Zhou, Xipeng Qiu

SpeechAgents: Human-Communication Simulation with Multi-Modal Multi-Agent Systems

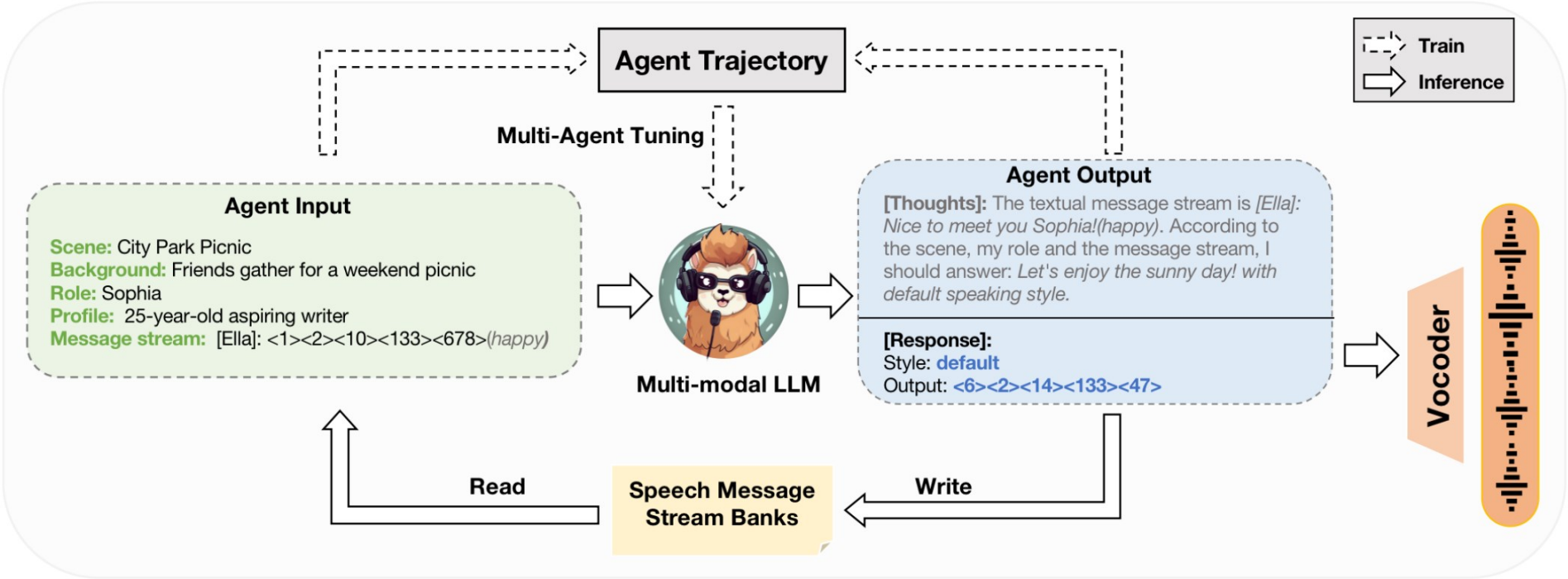


(a) Multi-Agent System (b) Multi-Modal Multi-Agent System

SpeechAgents: Human-Communication Simulation with Multi-Modal Multi-Agent Systems



SpeechAgents: Human-Communication Simulation with Multi-Modal Multi-Agent Systems



Scene: Summer Community BBQ

Background: Community members organize a summer barbecue, bringing people together for grilled food, music, and outdoor fun, fostering a sense of unity and camaraderie.

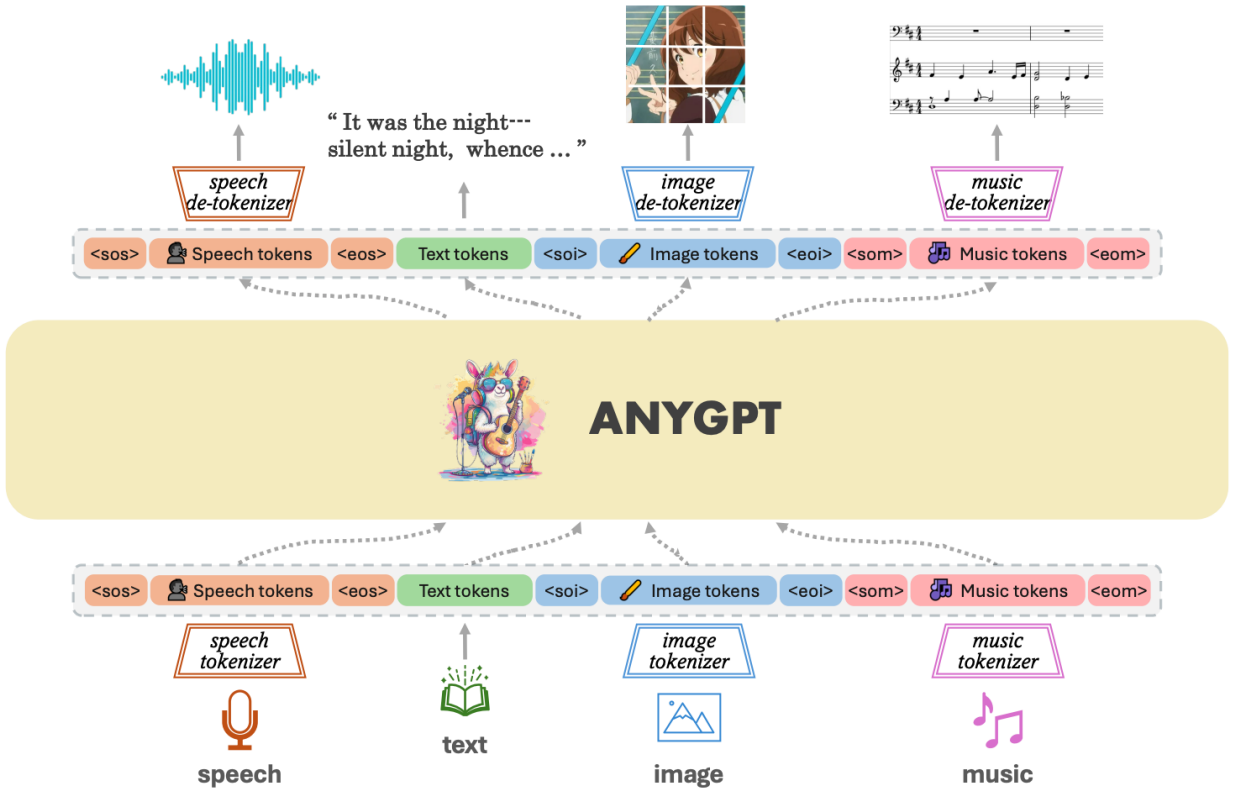
AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling

Jun Zhan*, Junqi Dai*, Jiasheng Ye*, Yunhua Zhou, **Dong Zhang**, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yugang Jiang, Xipeng Qiu

ACL 2024

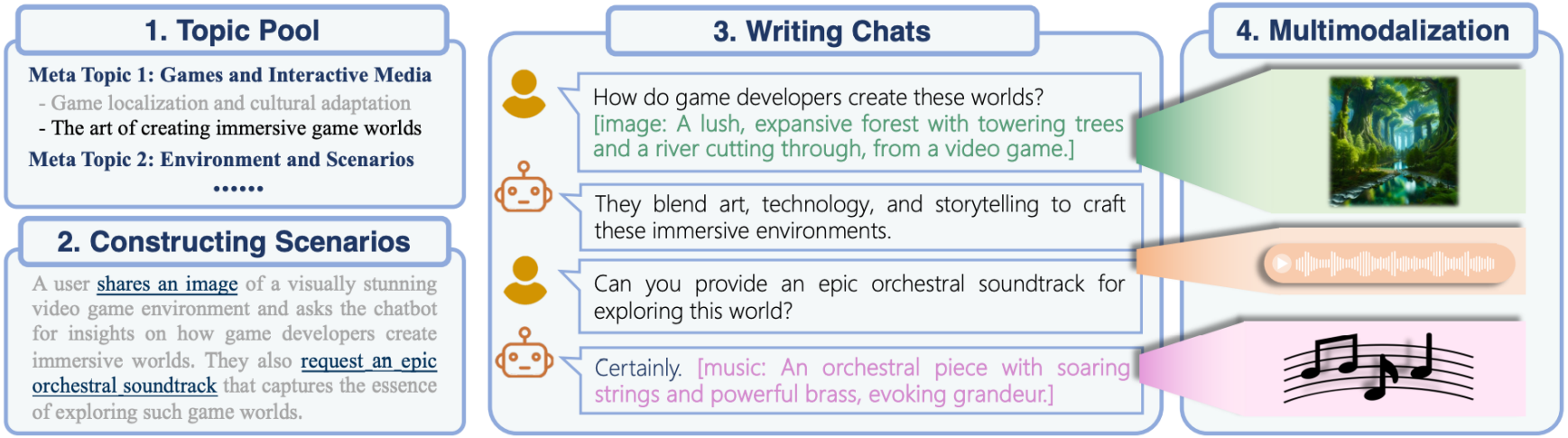
AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling

- Multi-modal Tokenizer
  - Image: Seed
  - Speech: SpeechTokenizer
  - Music: Encodec



Pretraining Dataset

AnyInstruct



Resources

- **SpeechGPT github: https://github.com/0nutation/SpeechGPT (950+ stars!)**
- **SpeechTokenizer github: https://github.com/ZhangXInFD/SpeechTokenizer (300+ stars!)**
- **USLM github: https://github.com/0nutation/USLM (100+ stars!)**
- **AnyGPT github: https://github.com/OpenMOSS/AnyGPT (590+ stars!)**

Thanks