香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

SCHOOL OF
DATA SCIENCE
數據科學學院

Unified Voice Generation with Preference Alignment

Xueyao Zhang

Ph.D. Thesis Defense

Supervisor: Zhizheng Wu

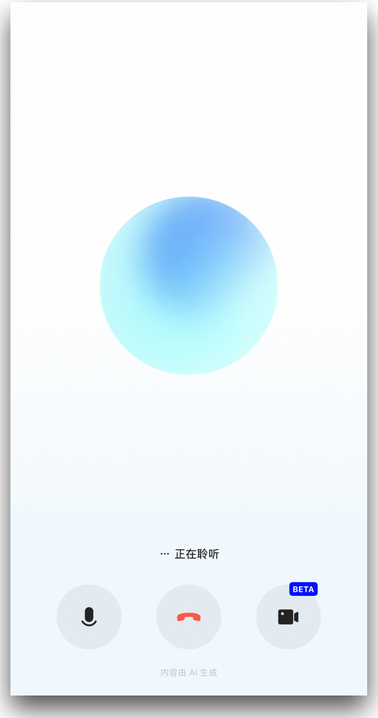
2026/05/14

Contents

Contents

Background: Why Voice Generation Matters?

Conversational AI



Siri *GPT-4o*

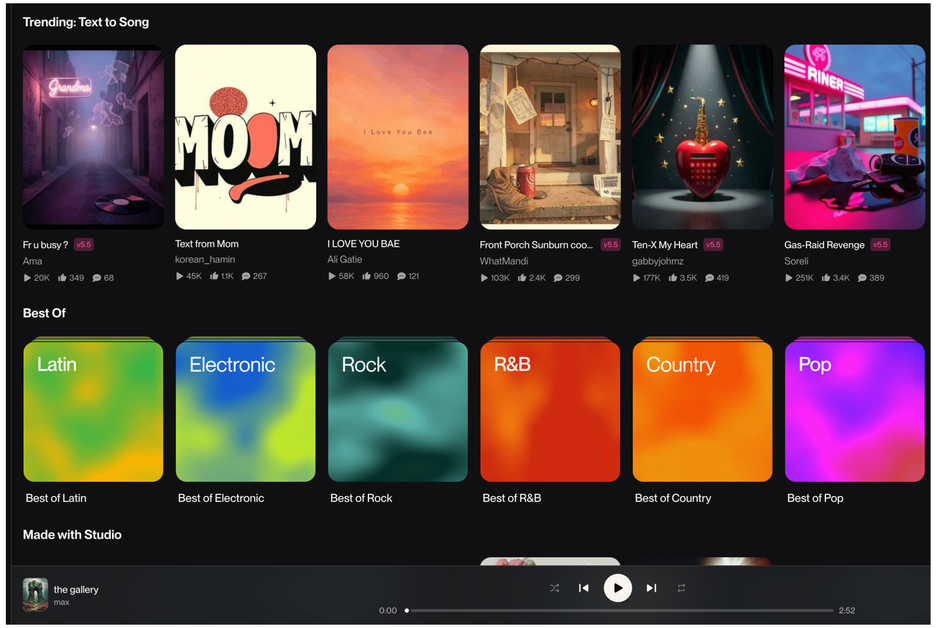
Doubao …

Content Localization

HeyGen *ElevenLabs*

Jianying …

Creative Applications

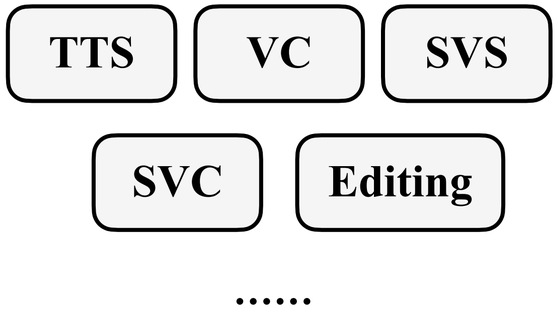


ACE Studio *Suno*

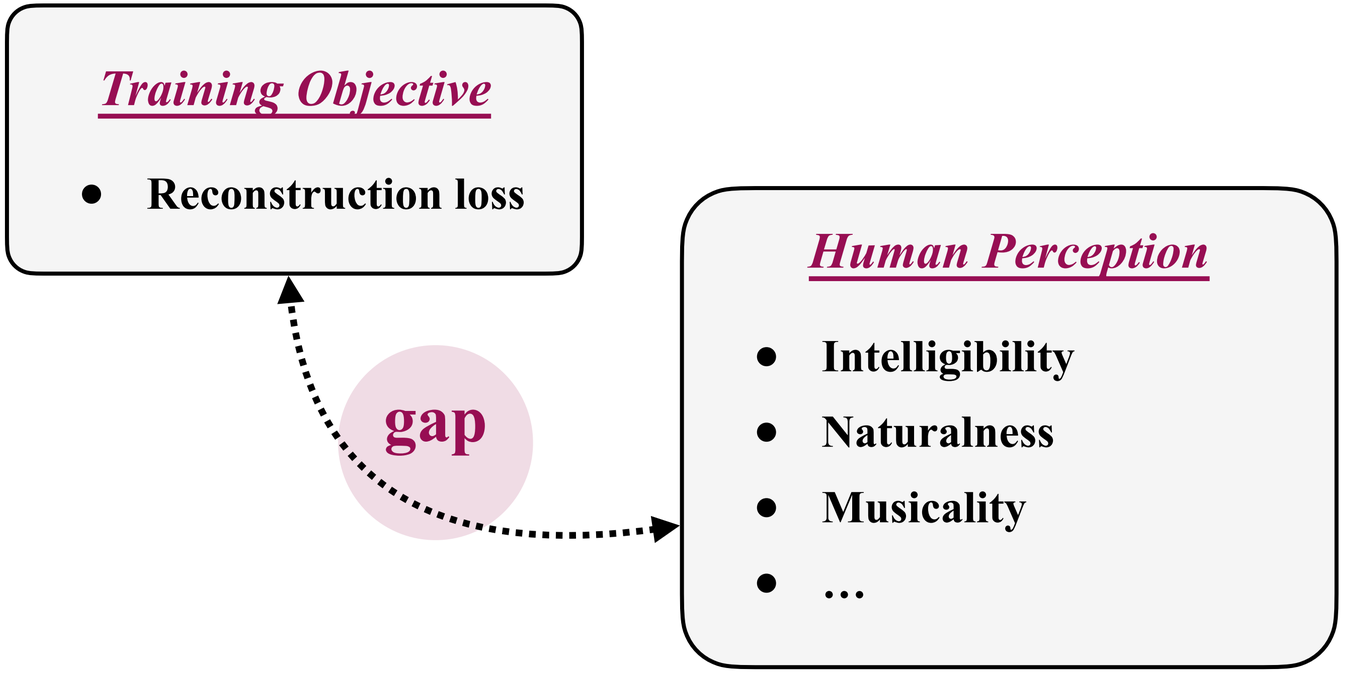
Synthesizer V …

Voice is becoming a general-purpose interface for interaction and content creation.

Limitations of Existing Voice Generation Research



- Different models
- Different datasets
- Different training recipes



1 Fragmented Across Tasks

2 Misaligned with Human Perception

These limitations motivate the need for unified and human-aligned voice generation

Focus: Three Core Scientific Questions

Scientific Question 1

Controllable Voice Generation

Scientific Question 2

Unified Voice Modeling

Scientific Question 3

Human Preference Alignment

Challenges in Controllable Voice Generation

1. What to Control?

- **Content — what is said or sung**
  - **Linguistic information** such as phonemes, words, sentences, and lyrics.
- **Prosody — how the voice evolves over time**
  - Suprasegmental cues such as **pitch**, **duration**, **rhythm**, **stress**, and **loudness**.
- **Melody — musically structured prosody**
  - A **pitch-duration contour** constrained by **musical notes**, rhythm, and key; especially important for singing.
- **Style — how it is expressed**
  - **Accent**, **emotion**, **speaking habits**, **singing techniques**, genre, and performance nuances.
- **Timbre/Speaker Identity — who is speaking or singing**
  - **Acoustic characteristics** and **personal vocal traits** that make a speaker or singer recognizable.

Voice Attributes for Controllable Generation [1-4]

2. How to Control?

3. Why is controllability challenging?

- **Entangled attributes:** Different voice attributes are highly entangled in the acoustic signals.
- **Limited supervision:** Parallel data are scarce. Attribute labels are expensive to collect.

How to learn disentangled, reference-controllable voice representations for various attributes under limited supervision?

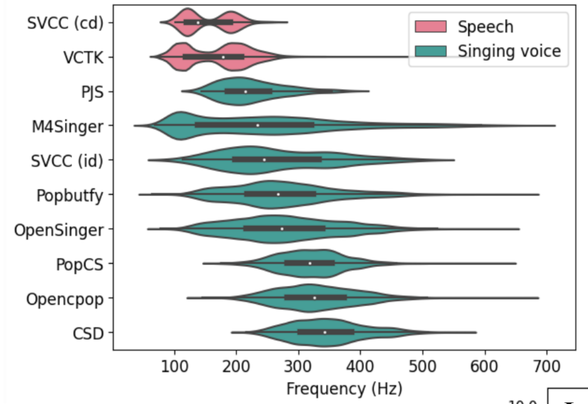
[1] Sundberg, J. and Rossing, T. D. (1990). The science of singing voice.
[2] Tan, X. (2023). Neural Text-to-Speech Synthesis. Springer.
[3] Taylor, P. (2009). Text-to-speech synthesis. Cambridge university press.
[4] Umbert, M., Bonada, J., Goto, M., Nakano, T., and Sundberg, J. (2015). Expression control in singing voice synthesis: Features, approaches, evaluation, and challenges. IEEE Signal Process. Mag., 32(6):55–73.

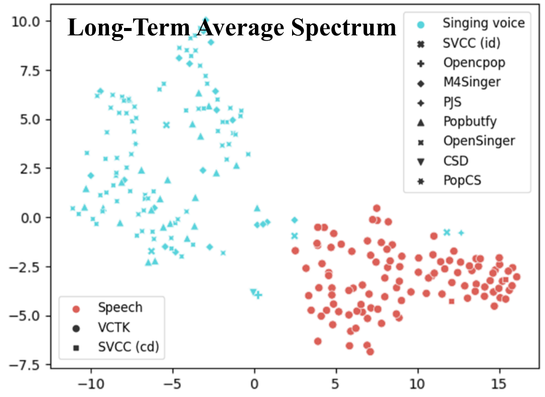
Challenges in Unified Voice Modeling

Diverse Tasks, Heterogeneous Requirements

Task	Input Conditions	Main Manipulation
Text to Speech	○ Text ○ Target speaker reference	○ Generate correct content ○ Imitate timbre, style, and prosody
Voice Conversion	○ Source speech ○ Target speaker reference	○ Preserve content ○ Convert speaker
Singing Voice Synthesis	○ Lyrics ○ Musical score ○ Target singer reference	○ Generate correct lyrics and melody ○ Imitate timbre, style, and prosody
Singing Voice Conversion	○ Source singing ○ Target singer reference	○ Preserve lyrics and melody ○ Convert singer
Speech Editing	○ Original speech ○ Edited text	○ Modify only the text
Singing Lyric Editing	○ Original singing ○ Edited lyric	○ Modify only the lyric

Speech-Singing Distribution Gap





*There are distinct **F0**, **energy**, and **timbre** patterns between speech and singing voices.* [1]

How to design a single formulation that can unify diverse tasks?

How to design shared representations that can bridge speech-singing distribution gaps?

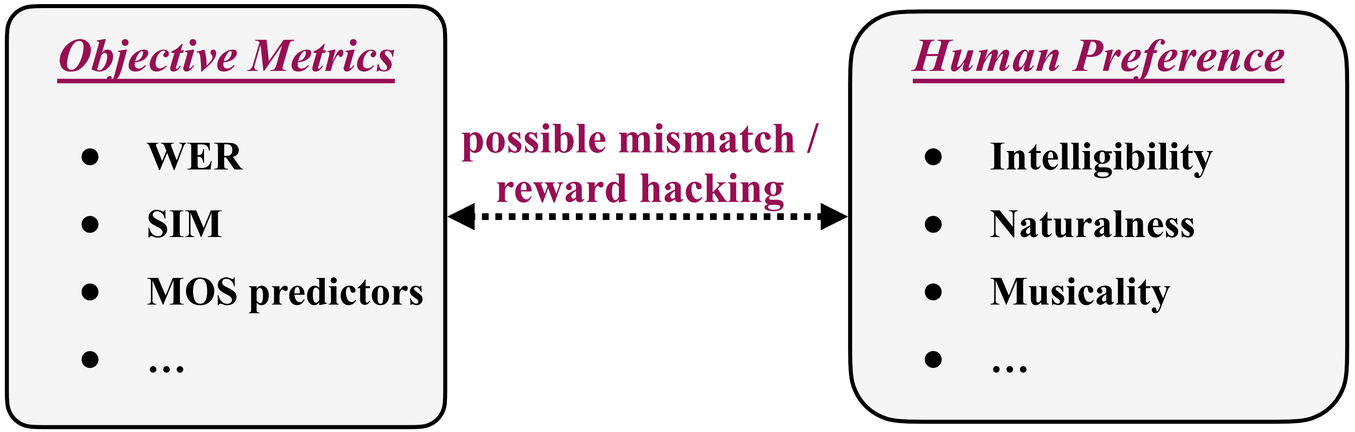
[1] **Xueyao Zhang**, et al. Leveraging Content-based Features from Multiple Acoustic Models for Singing Voice Conversion. Machine Learning for Audio Workshop, NeurIPS 2023.

Challenges in Human Preference Alignment

Challenges in Reward Signals *(i.e., What to Align?)*

Challenges in Algorithms *(i.e., How to Align?)*

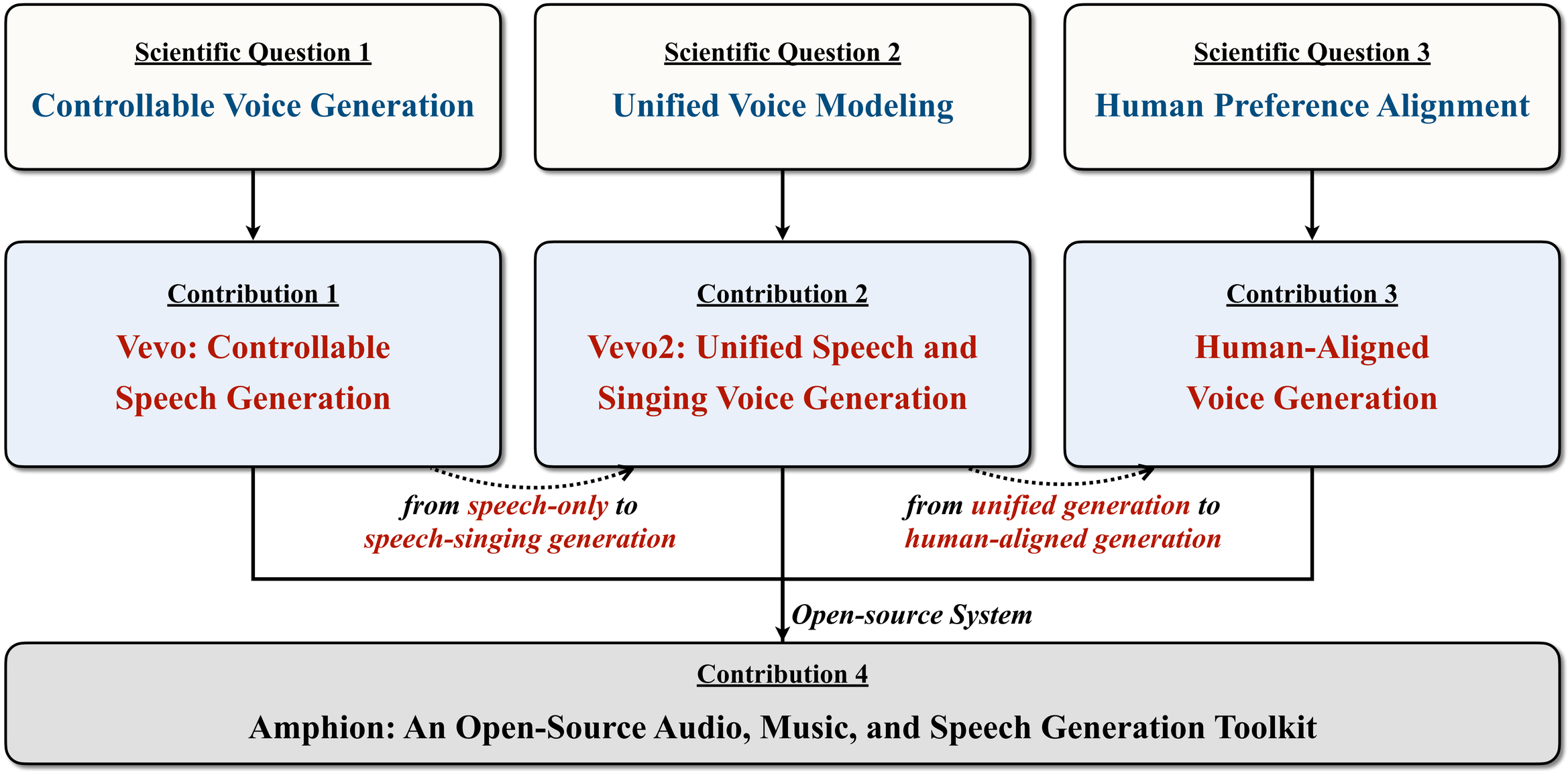
Limitations of metric-based rewards



Human feedback is ideal but expensive

- **Diverse generative architectures**
  - Auto-regressive based
  - Flow-Matching based
  - Masked Generative Model based
- **Multiple preference objectives**
  - Speech emphasizes intelligibility and naturalness.
  - Singing additionally requires melody quality and pitch accuracy.
  - A key challenge is ***multi-objective alignment*** without harmful trade-offs.

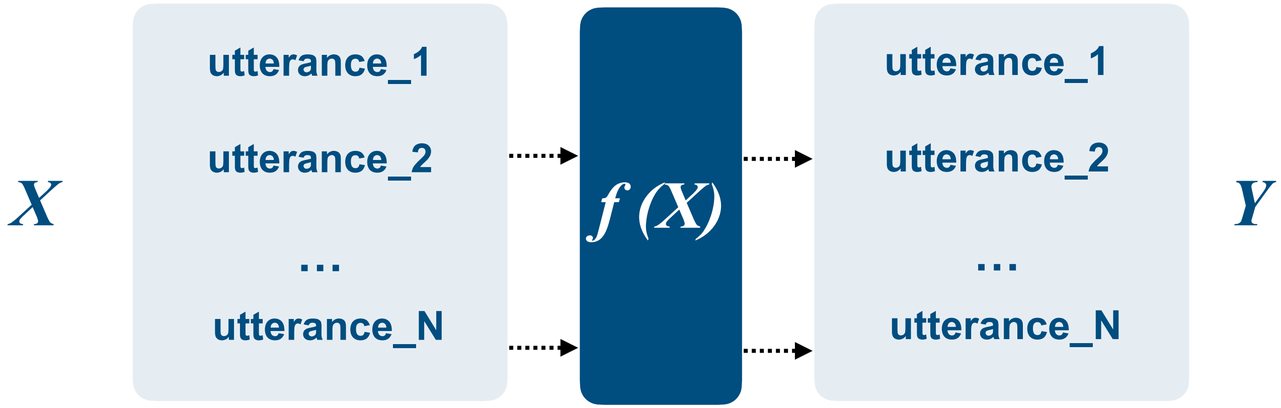
Research Aim: Human-Aligned Unified Voice Generation

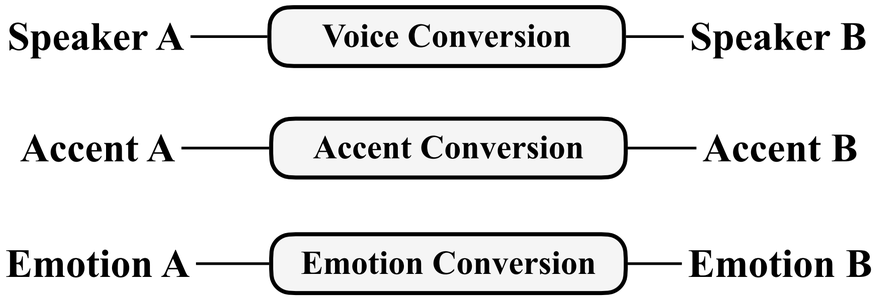


Contents

How to achieve controllable speech generation?

Naive Approach: Learning Mappings from Parallel Corpus

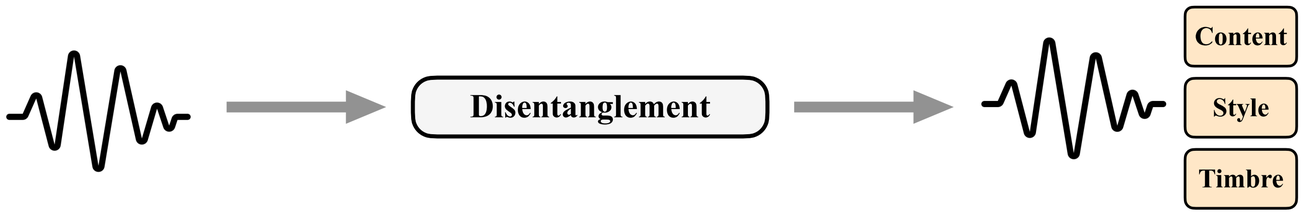




Parallel corpus is expensive to collect and not scalable.

Prior Art: Disentanglement Enables Controllability

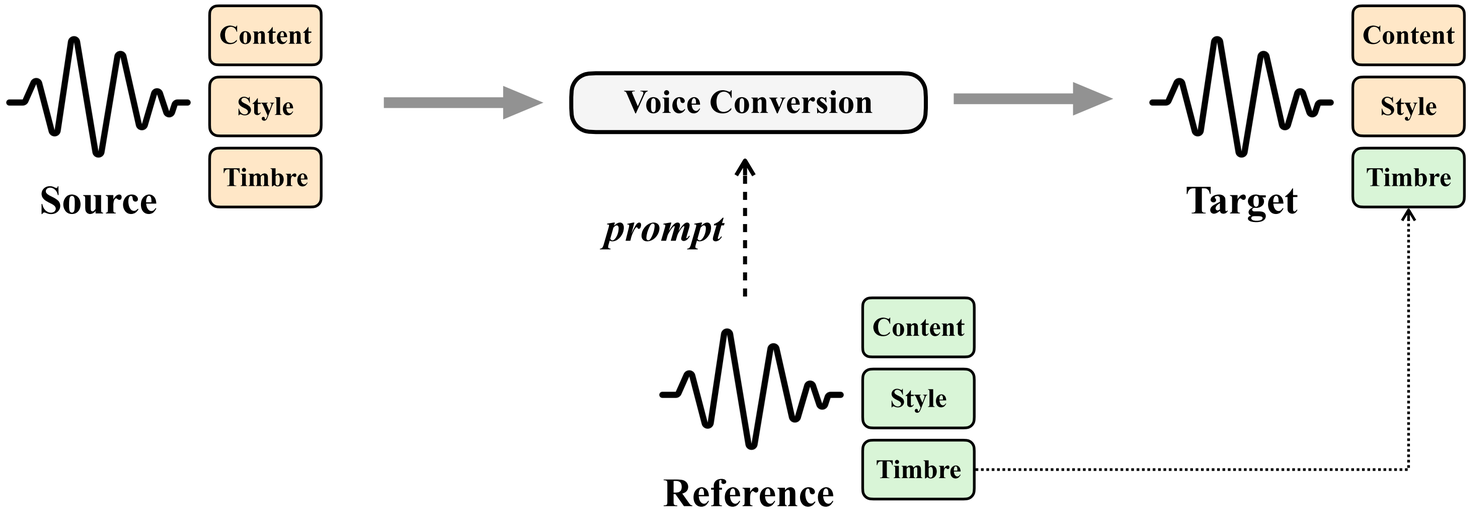
Training Stage: Attribute Disentanglement



Speech attributes in Vevo:

- **Content:** what is said
- **Style:** how it is expressed
- **Timbre:** who is speaking

Inference Stage: Substituting Attributes for Precise Control

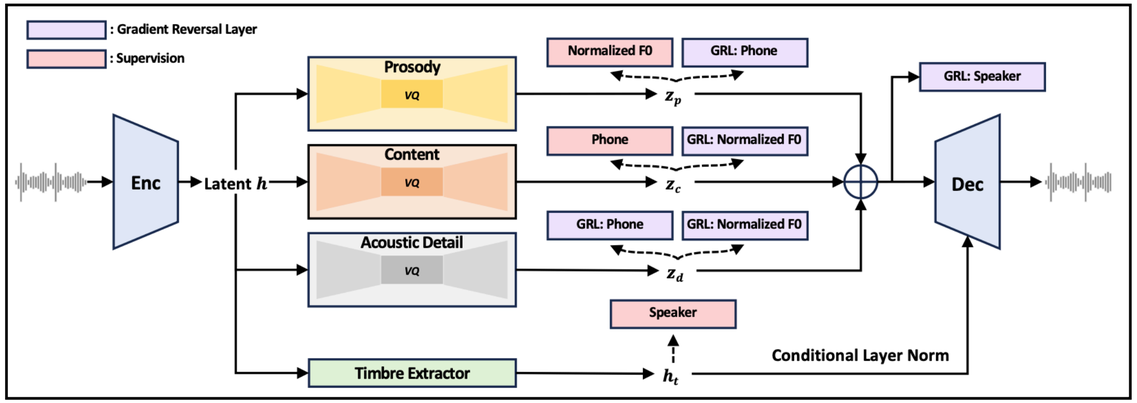


How to achieve the disentanglement?

Existing works

1. **Knowledge Distillation**
   - Auxiliary tasks such as ASR, F0 Prediction, or Speaker Verification (*FACodec* [1]).
2. **Perturbation-based Training**
   - Signal-based perturbation (*NANSY* [2]), Adversarial learning (*FACodec* [1])

Supervised Disentanglement



FACodec (ICML 2024)[1]

Strengths

Domain knowledge driven, High interpretability

Weaknesses

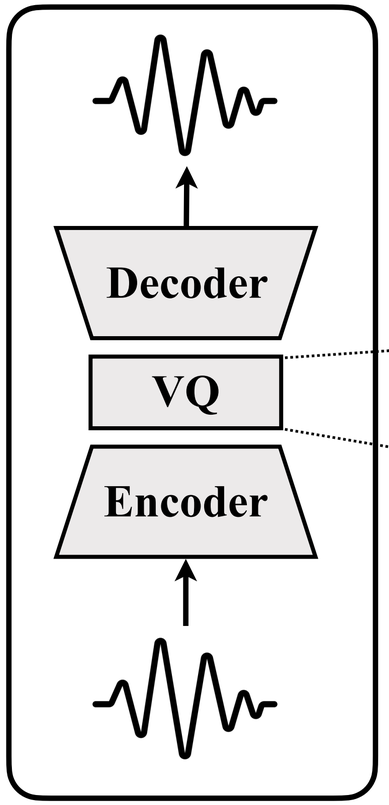
(1) Dependence on annotated data

(2) Multiple losses, unstable training

[1] Zeqian Ju, et al. Naturalspeech 3: Zero-shot speech synthesis with factorized codec and diffusion models. ICML 2024.
[2] Hyeong-Seok Choi, et al. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. NeurIPS 2021.

Vevo: Self-Supervised Disentanglement without Annotations

Existing Speech Tokenizers



Reconstruction-based Self Supervised Training

	Representations	#Vocab	Intelligibility WER (↓)	Speaker Similarity S-SIM (to ref) (↑)	Speaker Similarity S-SIM (to src) (↓)	F0 Correlation FPC (to src) (↑)
	Ground Truth	-	5.526	0.762	0.087	1.000
Starting point of information filtering	24th layer features	-	5.706	0.266	0.400	0.768
	18th layer features	-	5.324	0.250	0.505 ↑	0.824
	12th layer features	-	5.348	0.200	0.626 ↑	0.805
	PPG features	-	6.143	0.449	0.157	0.741
	ASR tokens	29	7.836	0.463	0.125	0.698
	K-means tokens	1024	11.493	0.398	0.150	0.734
Content-style Tokens	VQ-VAE tokens	16384	6.807	0.398	0.306	0.826
		4096	6.908 ↑	0.403	0.236 ↓	0.797 ↓
		1024	6.967 ↑	0.418	0.249	0.764 ↓
		32	9.731 ↑	0.426	0.161 ↓	0.706 ↓
Content Tokens		16	13.169 ↑	0.441	0.146 ↓	0.672 ↓
		8	21.813 ↑	0.392	0.109 ↓	0.675

HuBERT Features [1] — Content-style Tokens — Content Tokens

Timbre is the first to be filtered out

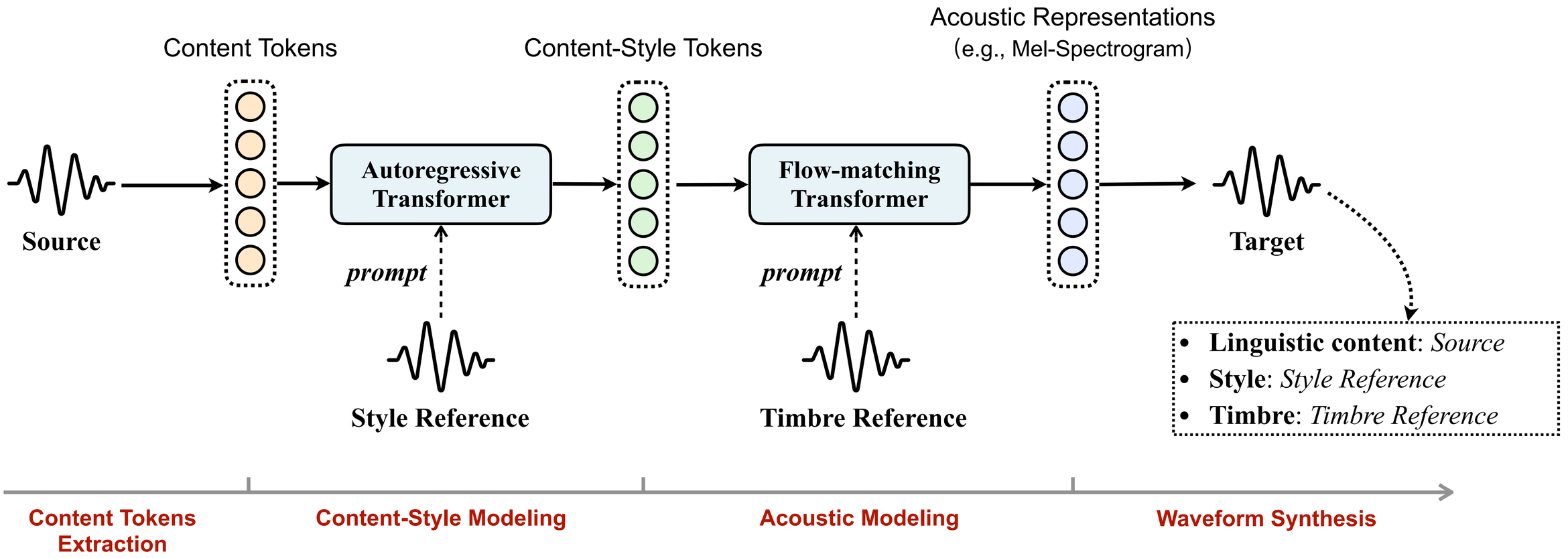
Content is the last to be filtered out

Acoustic-level Information → Semantic-level Information

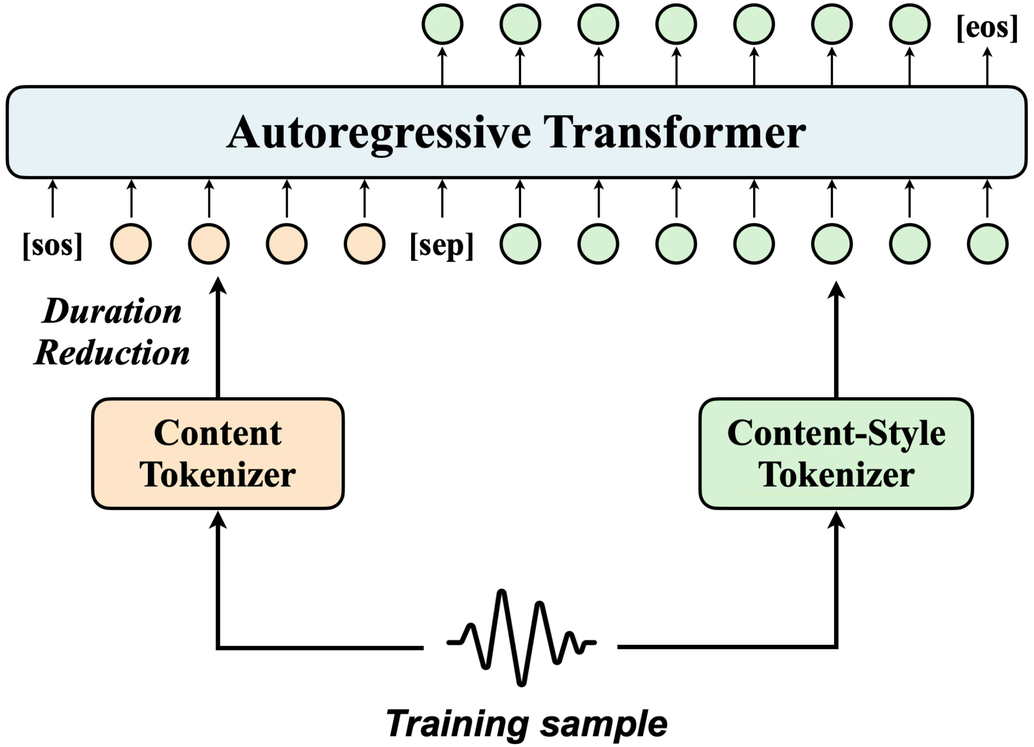
Key Findings: Codebook size is a disentanglement bottleneck for speech tokenizers

[1] Wei-Ning Hsu, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. TASLP 2021.

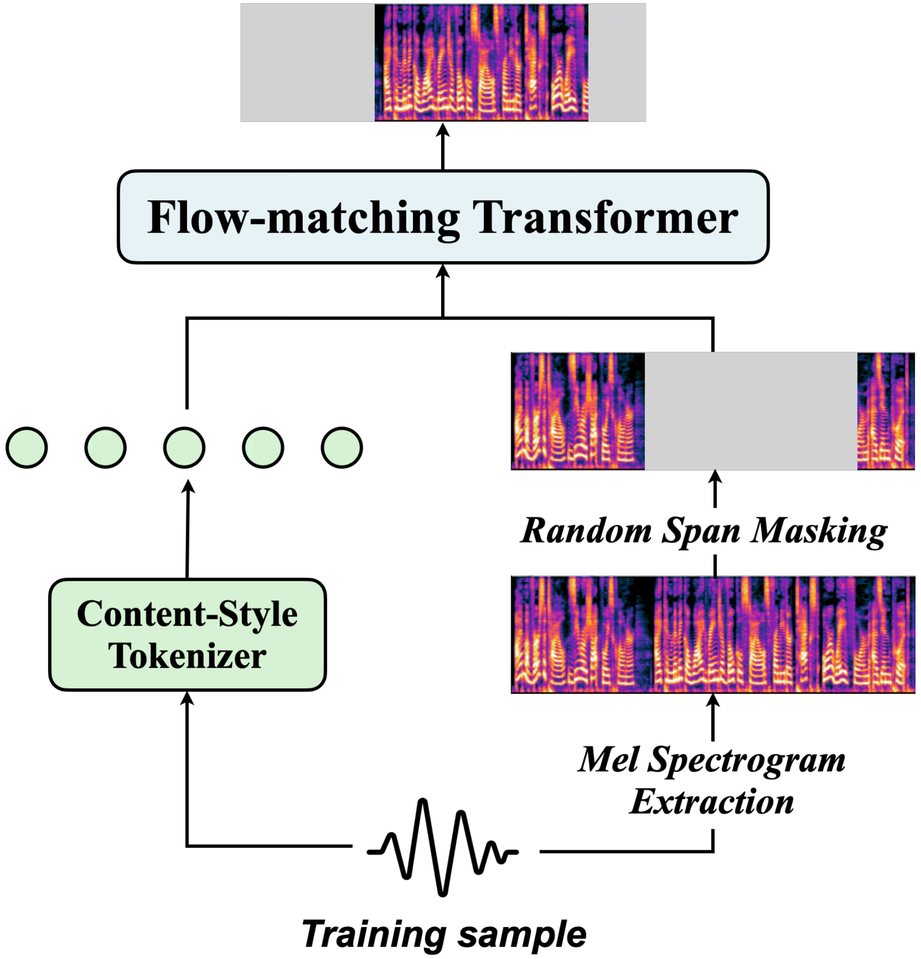
Vevo: From Disentangled Speech Tokens to Controllable Speech Generation



Training: Self-Supervised, In-Context Learning



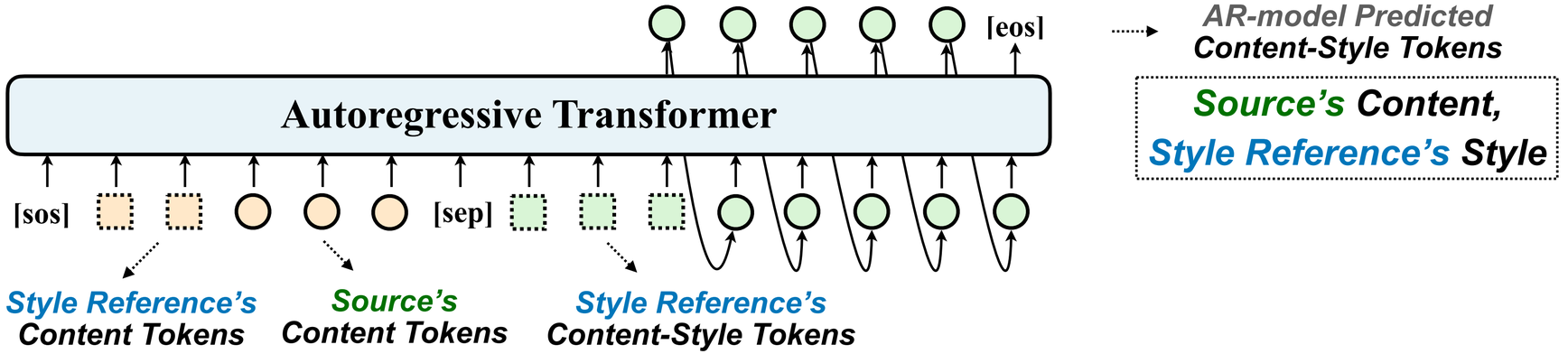
Content-Style Modeling: Next Token Prediction



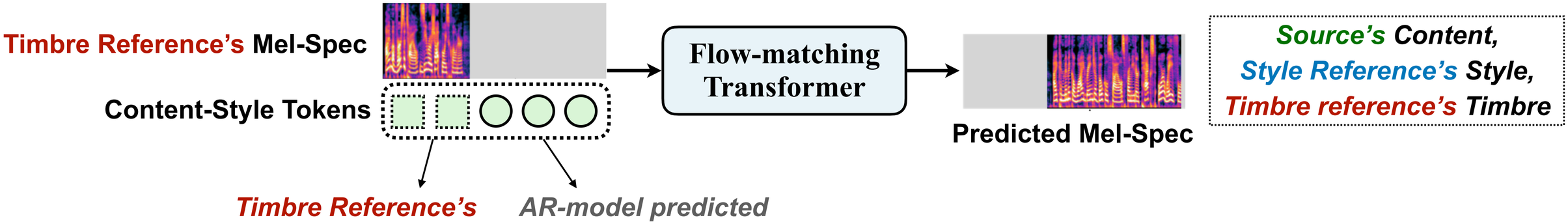
Acoustic Modeling: Masked Mel-Spec Prediction

Inference: Reference-based Controllability

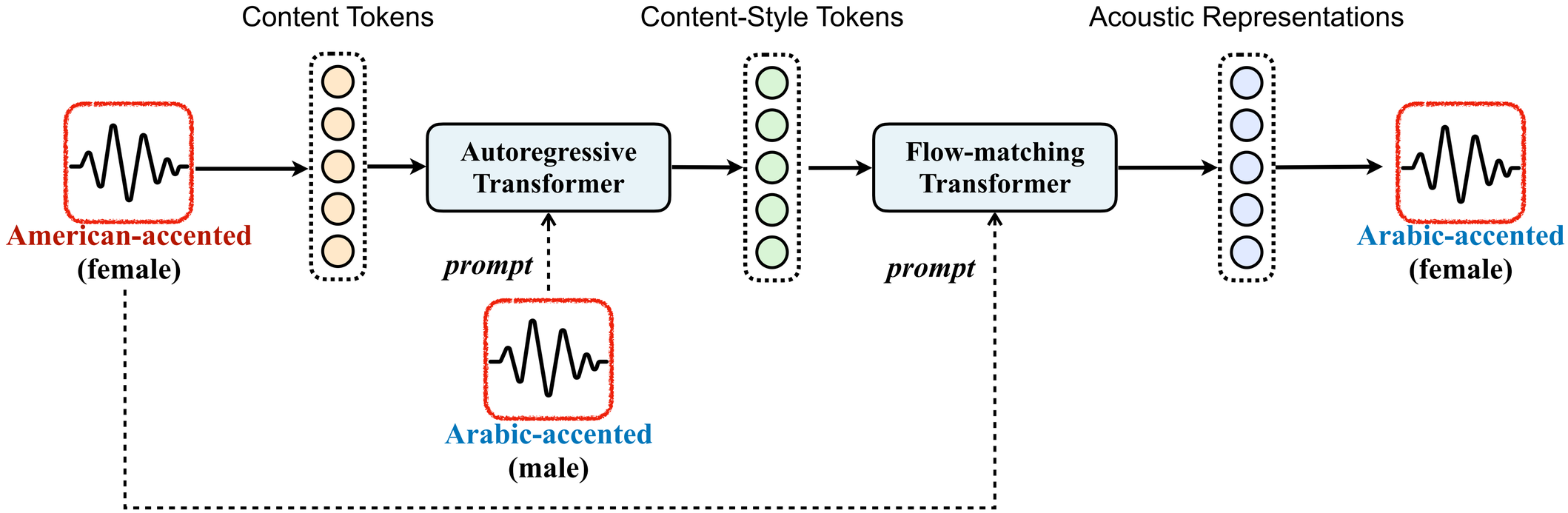
Auto-regressive Transformer (AR model)



Flow-matching Transformer (FM model)



Inference Case: Zero-Shot Accent Conversion



- ★ No parallel corpus
- ★ No accent labels
- ★ No transcriptions

→ Fully self-supervised learning on speech waveforms → Yet enables accent conversion

Experimental Setup

- **Training Data**
  - Libri-light (60K hours of English audiobook speech data)
- **Evaluation Tasks**
  - **Zero-Shot Voice Conversion**
    - Preserve content, Convert speaker
  - **Zero-Shot Style Conversion** (e.g., Accent Conversion, Emotion Conversion)
    - Preserve content, Preserve speaker, Convert Style
  - **Zero-Shot Text to Speech**
    - Follow text, imitate speaker

Results: Zero-Shot Voice Conversion

Style-preserved Zero-Shot VC

Model	AR?	Training Data (ContRep / Model)	WER (↓)	S-SIM (to r) (↑)	FPC (to i) (↑)	N-MOS (↑)	SS-MOS (to r) (↑)	PS-MOS (to i) (↑)
HierSpeech++ [113]	✗	500K / 2.8K	4.233	0.385	<u>0.634</u>	3.05 $\pm$0.23	3.24 $\pm$0.25	3.08 $\pm$0.26
LM-VC [220]	✓	1K / 60K	8.623	0.310	0.524	2.90 $\pm$0.11	2.98 $\pm$0.18	2.16 $\pm$0.26
UniAudio [235]	✓	1K / 100K	7.241	0.264	0.575	3.04 $\pm$0.15	2.47 $\pm$0.20	2.51 $\pm$0.25
FACodec [90]	✗	60K / 60K	<u>3.682</u>	0.327	0.611	2.50 $\pm$0.20	3.10 $\pm$0.24	<u>3.10</u> $\pm$0.23
Vevo-Voice	✓	60K / 60K	7.694	**0.458**	0.485	<u>3.09</u> $\pm$0.13	**3.51** $\pm$0.24	2.60 $\pm$0.23
Vevo-Timbre	✗	60K / 60K	**2.968**	<u>0.420</u>	**0.686**	**3.35** $\pm$0.09	<u>3.36</u> $\pm$0.16	**3.45** $\pm$0.17

Style-converted Zero-Shot VC

Model	WER (↓)	S-SIM (to r) (↑)	A-SIM (to r) (↑)	E-SIM (to r) (↑)	N-MOS (↑)	SS-MOS (to r) (↑)	AS-MOS (to r) (↑)	ES-MOS (to r) (↑)
Ground Truth	10.917	0.762	0.763	0.965	-	-	-	-
HierSpeech++ [113]	12.921	0.466	0.526	0.658	3.04 $\pm$0.14	3.15 $\pm$0.23	3.13 $\pm$0.22	2.55 $\pm$0.19
LM-VC [220]	20.353	0.312	0.426	0.649	2.40 $\pm$0.10	2.56 $\pm$0.15	3.02 $\pm$0.19	2.46 $\pm$0.17
UniAudio [235]	15.751	0.311	0.486	0.611	2.95 $\pm$0.11	2.39 $\pm$0.17	2.42 $\pm$0.15	2.41 $\pm$0.26
FACodec [90]	<u>12.731</u>	0.434	0.514	0.688	2.36 $\pm$0.18	3.19 $\pm$0.22	3.01 $\pm$0.16	2.30 $\pm$0.22
Vevo-Timbre	**12.351**	<u>0.486</u>	<u>0.567</u>	<u>0.816</u>	**3.43** $\pm$0.09	<u>3.46</u> $\pm$0.15	<u>3.55</u> $\pm$0.25	<u>2.66</u> $\pm$0.26
Vevo-Voice	15.214	**0.517**	**0.614**	**0.872**	<u>3.24</u> $\pm$0.11	**3.70** $\pm$0.24	**3.90** $\pm$0.19	**3.20** $\pm$0.16

Vevo-Timbre:
- Only using FM model
- Content-preserved, Style-preserved, Timbre-converted

Vevo-Voice:
- Using both AR and FM models
- Content-preserved, Style-converted, Timbre-converted

Key Findings

① **Style-Preserved VC:** Vevo-Timbre dominates common VC metrics vs. existing baselines

② **Style-Converted VC:** Vevo-Voice significantly outperforms existing baselines on accent/emotion imitation

③ **Vevo-Timbre vs. Vevo-Voice:**
- **Vevo-Timbre**: excels at preserving source style
- **Vevo-Voice**: excels at style imitation + higher speaker similarity
- **Trade-off**: Vevo-Voice has higher WER due to the auto-regressive design

We will address this issue via intelligibility alignment (see *Part III*)

Results: Zero-Shot Style Conversion

Accent Conversion

Model	Zero-shot	Supervision			WER (↓)	A- / E-ACC (↑)	A- / E-SIM (↑)	N-COMS (↑)	A- / E-CMOS (↑)
		PC	SL	Text					
ASR-AC [89]	✗	✗	✓	✓	4.775	0.633	-	0.00	0.00
Vevo-Style (ASR)	✓	✗	✗	✓	**1.550**	**0.723**	**0.570**	**0.32** ± 0.11	**0.49** ± 0.14
Vevo-Style	✓	✗	✗	✗	3.083	0.663	0.562	0.30 ± 0.13	0.35 ± 0.21
VoiceShop [6]	✗	✓	✓	✓	5.547	0.642	-	0.00	0.00
Vevo-Style (ASR)	✓	✗	✗	✓	**3.553**	**0.735**	**0.585**	**0.26** ± 0.16	**0.18** ± 0.20
Vevo-Style	✓	✗	✗	✗	5.464	0.673	0.554	0.12 ± 0.10	0.13 ± 0.08
Conv-Speak [230]	✗	✓	✓	✗	9.950	0.571	-	0.00	0.00
Vevo-Style (ASR)	✓	✗	✗	✓	**2.778**	0.864	0.574	0.10 ± 0.05	0.40 ± 0.12
Vevo-Style	✓	✗	✗	✗	3.889	**0.903**	**0.580**	**0.15** ± 0.12	**0.60** ± 0.16

Emotion Conversion

Model	Zero-shot	Supervision			WER (↓)	A- / E-ACC (↑)	A- / E-SIM (↑)	N-COMS (↑)	A- / E-CMOS (↑)
		PC	SL	Text					
Emovox [271]	✗	✗	✓	✓	15.444	0.750	-	0.00	0.00
Vevo-Style (ASR)	✓	✗	✗	✓	**9.842**	0.692	0.800	1.74 ± 0.20	0.45 ± 0.11
Vevo-Style	✓	✗	✗	✗	10.221	**0.754**	**0.825**	**1.78** ± 0.20	**0.49** ± 0.13

Vevo-Style:
- Using both AR and FM models
- Content-preserved, Style-converted, Timbre-converted

Vevo-Style (ASR):
- We replace the input content tokens as ASR-predicted phones. (i.e., introducing the ***text supervision***)

Key Findings

① **Vevo-Style: Self-Supervised Only, Zero-Shot, Yet Superior**
- Trained solely on unlabeled audiobook speech — no accent/emotion fine-tuning
- Outperforms existing baselines in intelligibility, quality, and accent/emotion imitation

② **Vevo-Style (ASR): Text Supervision Further Boosts Performance**
- Replacing content tokens with ASR-predicted phones further improves WER and accent imitation

We will borrow this idea and introduce text supervision in Vevo2 (see *Part II*)

Results: Zero-Shot TTS

Zero-Shot TTS

Model	AR?	Training Data	WER (↓)	S-SIM (↑)	A-SIM (↑)	E-SIM (↑)	N-CMOS (↑)	SS-MOS (↑)	AS-MOS (↑)	ES-MOS (↑)
Ground Truth	-	-	11.348	0.710	0.633	0.936	0.00	-	-	-
CosyVoice [48]	✓	171K	8.400	0.614	0.640	0.839	-0.18 $\pm$0.19	4.11 $\pm$0.19	3.99 $\pm$0.23	3.66 $\pm$0.19
MaskGCT [218]	✗	100K	9.442	0.659	0.645	0.822	-0.04 $\pm$0.19	4.16 $\pm$0.16	4.38 $\pm$0.14	3.76 $\pm$0.25
VALL-E [212]	✓	45K	13.226	0.400	0.485	0.735	-1.24 $\pm$0.42	2.82 $\pm$0.40	2.77 $\pm$0.45	2.63 $\pm$0.36
Voicebox [111]	✗	60K	**9.414**	0.463	0.575	0.811	-0.35 $\pm$0.21	3.87 $\pm$0.21	3.49 $\pm$0.29	3.61 $\pm$0.19
VoiceCraft [157]	✓	9K	13.057	0.392	0.517	0.788	-0.50 $\pm$0.23	3.47 $\pm$0.32	3.29 $\pm$0.28	3.52 $\pm$0.25
Vevo-TTS	✓	60K	12.066	**0.505**	**0.579**	**0.840**	**-0.14** $\pm$0.18	**4.05** $\pm$0.21	**4.12** $\pm$0.21	**4.03** $\pm$0.19

Vevo-TTS:
- Using both AR and FM models
- For AR model, we input text directly rather than using content tokens
- Text-followed, Style-imitated, Timbre-imitated

Key Findings

① **Vevo-TTS vs. Voicebox:**
- Identical training data, slightly higher WER (common AR model weakness), but excels across all other metrics

We will address this issue via intelligibility alignment (see *Part III*)

② **Vevo-TTS vs. SOTA**
- Outperforms CosyVoice & MaskGCT in emotion imitation, despite training on far less diverse data (audiobook vs. large-scale in-the-wild data)

We will follow this data design in Vevo2 (see *Part II*)

- Verifies our content-style tokens effectively capture style info and **are directly usable by downstream models (e.g., TTS) without extra adaptation**

Impact and Recognition in the Field

VEVO: CONTROLLABLE ZERO-SHOT VOICE IMITATION WITH SELF-SUPERVISED DISENTANGLEMENT

Xueyao Zhang[1]* **Xiaohui Zhang**[2] **Kainan Peng**[2] **Zhenyu Tang**[2] **Vimal Manohar**[2], **Yingru Liu**[2] **Jeff Hwang**[2] **Dangna Li**[2] **Yuhao Wang**[2] **Julian Chan**[2] **Yuan Huang**[2] **Zhizheng Wu**[1]† **Mingbo Ma**[2]
[1]The Chinese University of Hong Kong, Shenzhen [2]Meta AI

TITLE	CITED BY	YEAR
Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement X Zhang, X Zhang, K Peng, Z Tang, V Manohar, Y Liu, J Hwang, D Li, ... ICLR 2025	61	2025

Published at ICLR 2025; cited 60+ times

(Cited and used as a baseline by CMU, JHU, PKU, Tencent, Bilibili, Kuaishou, and others)

Towards Controllable Speech Synthesis in the Era of Large Language Models: A Systematic Survey

Tianxin Xie[1], **Yan Rong**[1], **Pengfei Zhang**[1], **Wenwu Wang**[2], **Li Liu**[1*],
[1]The Hong Kong University of Science and Technology (Guangzhou), [2]University of Surrey

HKUST & University of Surrey (EMNLP 2025)

*"In the zero-shot setting, among the six models, Vevo performs best in both naturalness (4.43±0.55) and expressiveness (4.32±0.75), indicating **strong general quality without explicit guidance**."*

GenVC: Self-Supervised Zero-Shot Voice Conversion

Zexin Cai, Henry Li Xinyuan, Ashi Garg, Leibny Paola García-Perera, Kevin Duh, Sanjeev Khudanpur, Matthew Wiesner, Nicholas Andrews
Human Language Technology Center of Excellence
Johns Hopkins University

Johns Hopkins University (ASRU 2025)

*"Vevo addresses similar limitations by **introducing a controllable framework for timbre and style conversion**. The system comprises two stages: an autoregressive transformer followed by a flow-matching transformer. **Both stages are trained with self-supervised, in-context learning, making the framework scalable.**"*

Contents

Motivation: Why Unify Speech and Singing Voice Generation?

Key Intuition: Speech and singing generation can mutually benefit from unified modeling.

Vevo2: Three Core Designs for Unified Voice Generation

1 Unified Formulation & Extended Controllability

Content | Style | Timbre | **Prosody** | **Melody**

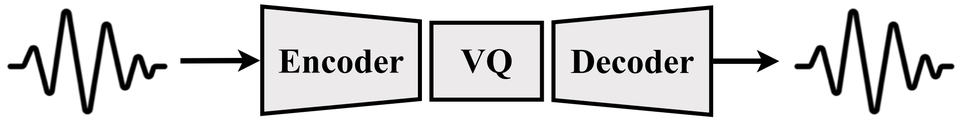
From Vevo to Vevo2: Additional controllability on **prosody** and **melody**

Tasks	From Vevo to Vevo2
Synthesis Tasks	○ Text to Speech ○ **Singing Voice Synthesis** ○ **Humming to Singing** ○ **Instrument to Singing**
Conversion Tasks	○ Voice Conversion ○ Accent Conversion ○ Emotion Conversion ○ **Singing Voice Conversion** ○ **Singing Style Conversion**
Editing Tasks	○ **Speech Editing** ○ **Singing Lyric Editing**

Red: New tasks enabled by Vevo2

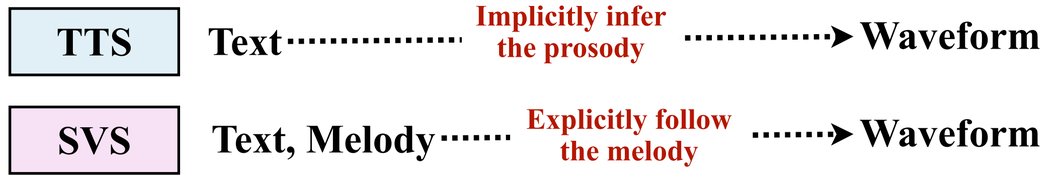
Design a unified and controllable framework that supports both speech and singing voice generation.

2 Unified Voice Representations



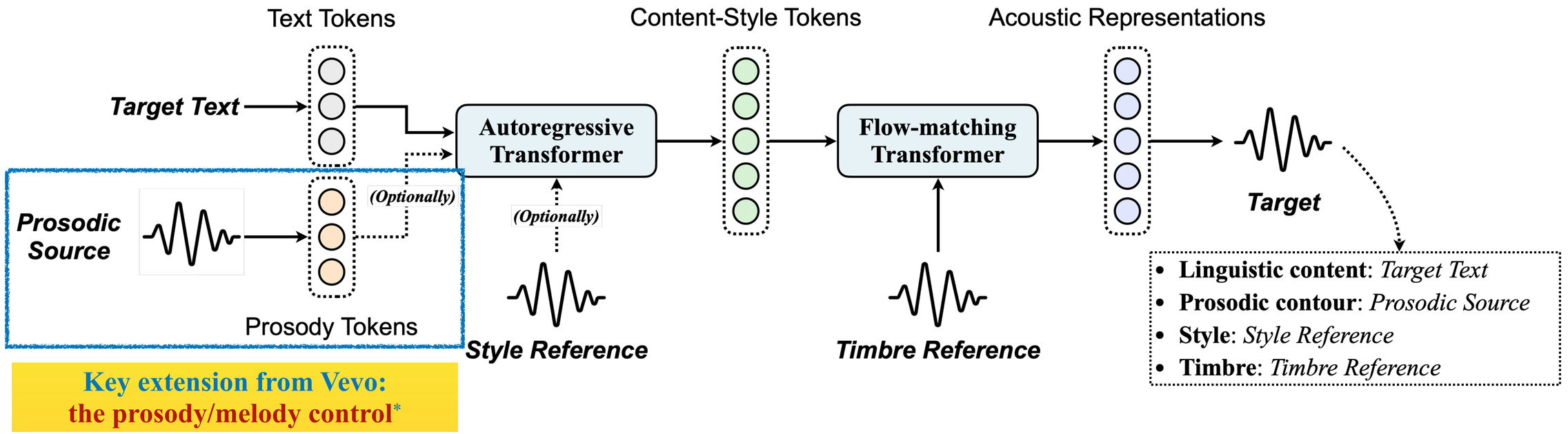
Propose unified tokenizers that capture speech-singing shared patterns and singing-specific attributes.

3 Speech-Singing Joint Training



Jointly train one model to support both implicit prosody inference and explicit melody following.

1. A Unified and Controllable Framework for Voice Generation

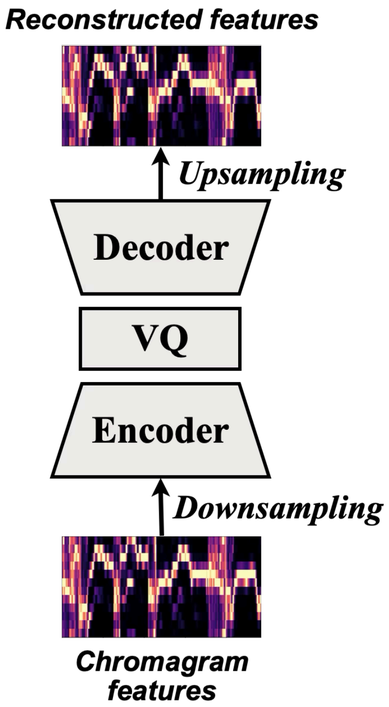


Vevo2 decomposes unified voice generation into controllable attributes through multi-prompt conditioning.

* **Prosody:** suprasegmental cues such as pitch, duration, rhythm, stress, and loudness. **Melody:** A pitch-duration contour constrained by musical notes, treated as a musically structured prosody in Vevo2.

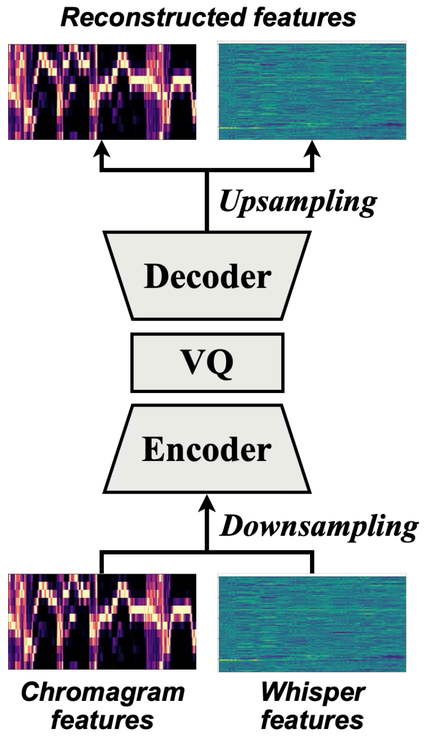
2. Unified Speech-Singing Tokenizers

Unified Prosody Tokenizer



- **Unified prosody and melody encoding**
  - Chromagram features capture both speech prosody and singing melody signals
- **Octave- and notation-free**
  - Friendly for speech–singing unification and scalable training
- **Robust across audio domains**
  - Applicable to speech, singing, and even instrumental music

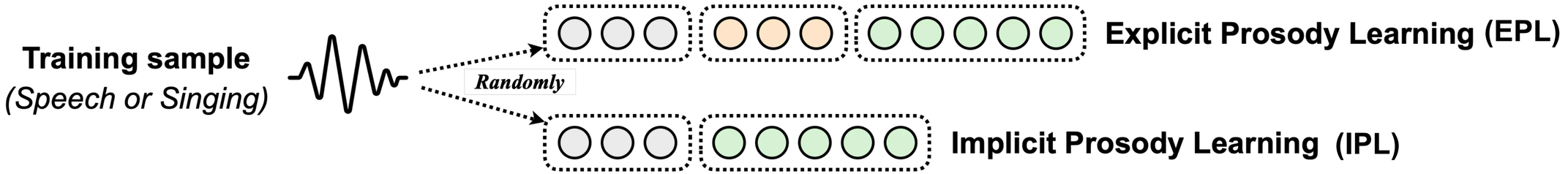
Unified Content-Style Tokenizer



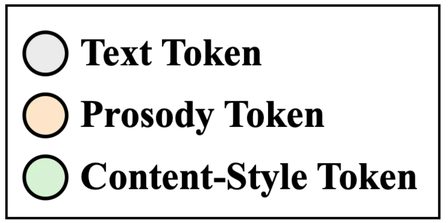
- **Weak text supervision**
  - Whisper features provide ASR-pretrained linguistic cues
- **Prosody / melody preservation**
  - Chromagram features help retain prosodic and melodic information
- **AR-friendly token sequence**
  - Low frame rate (12.5 Hz) reduces sequence length and alleviates the AR modeling burden

3. Speech-Singing Joint Training

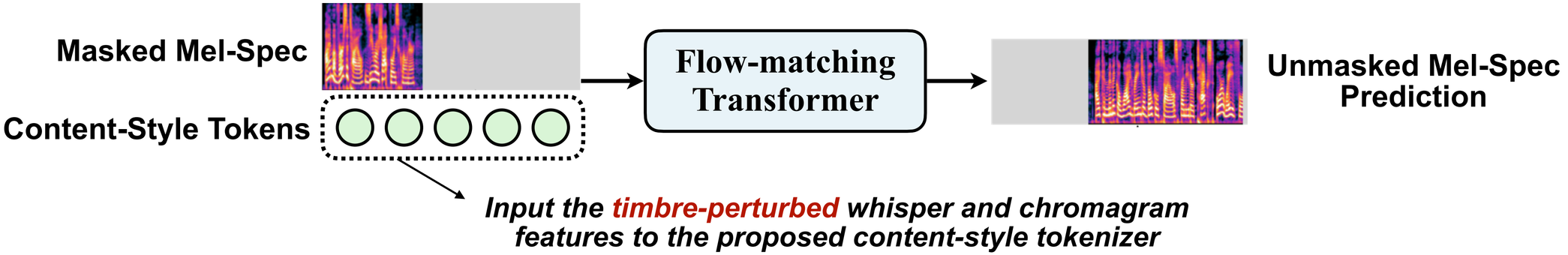
(1) Unified Content-Style Modeling: Explicit & Implicit Prosody Learning



- EPL for singing-like control: prosody / melody is explicitly provided.
- IPL for speech-like generation: prosody is implicitly inferred from text and context.



(2) Unified Acoustic Modeling: Timbre-disentangled Flow-matching



Experimental Setup

- **Training Data**
  - Emilia (101K hours of in-the-wild speech data) + SingNet (7K hours of source separated singing voice data)
- **Evaluation Tasks**
  - **Synthesis Tasks:**
    - Zero-Shot TTS, Zero-Shot SVS, Humming to Singing, Instrument to Singing, etc.
  - **Conversion Tasks**:
    - Zero-Shot VC, Zero-Shot SVC, Accent & Emotion Conversion, Singing Style Conversion, etc.
  - **Editing Tasks:**
    - Zero-Shot Speech Editing, Zero-Shot Singing Lyric Editing

Results: Synthesis Tasks on Speech and Singing Voice

Zero-Shot Text to Speech & Text to Singing

Model	#Hz	Data (#hours)	PT?	Expressive Speech				Singing Voice			
				WER	SIM	N-CMOS	SS-CMOS	WER	SIM	N-CMOS	SS-CMOS
Ground Truth	-	-	-	10.91	0.687	1.47 $\pm$0.24	0.87 $\pm$0.19	12.65	0.658	1.27 $\pm$0.14	0.55 $\pm$0.16
F5-TTS [26]	-	101K	✗	11.77	0.695	-1.05 $\pm$0.18	-0.06 $\pm$0.17	16.12	0.597	-1.89 $\pm$0.07	-1.29 $\pm$0.10
MaskGCT [218]	50	101K	✗	13.42	**0.736**	-1.14 $\pm$0.21	**0.07** $\pm$0.31	<u>11.71</u>	**0.753**	-1.74 $\pm$0.14	-0.97 $\pm$0.13
CosyVoice 2 [49]	25	167K	✓	**11.20**	<u>0.706</u>	**0.10** $\pm$0.29	-0.16 $\pm$0.13	16.18	0.659	-1.68 $\pm$0.15	-1.12 $\pm$0.16
Vevo2-base	12.5	101K	✗	15.52	0.677	-0.84 $\pm$0.15	-0.06 $\pm$0.27	19.39	0.658	-1.02 $\pm$0.16	-0.48 $\pm$0.18
		7K	✗	37.57	0.611	-1.25 $\pm$0.26	-0.57 $\pm$0.17	25.38	0.692	-0.69 $\pm$0.26	-0.36 $\pm$0.17
		101K, 7K	✗	14.32	0.681	-0.49 $\pm$0.24	-0.03 $\pm$0.19	15.78	0.708	<u>-0.28</u> $\pm$0.14	<u>-0.09</u> $\pm$0.21
Vevo2	12.5	101K, 7K	✓	<u>11.48</u>	0.689	<u>0.00</u>	<u>0.00</u>	**7.66**	<u>0.725</u>	**0.00**	**0.00**

Vevo2-base:
- The pre-trained model of Vevo2

Vevo2:
- Using the multi-objective alignment based on Vevo2-base (see ***Part III***)

Key Findings

① **Mutual benefit from unified speech-singing pre-training**
- **Speech data improves singing intelligibility; singing data enhances expressiveness, prosody, and similarity for speech generation.**
- Compared with speech-only or singing-only training, **the unified model achieves better overall quality**.

② **Competitive with existing models**
- Vevo2 is competitive on **expressive speech** and **significantly improves subjective singing voice quality** over zero-shot TTS baselines.
- The multi-objective alignment further improves the full system; details are left to ***Part III***.

Results: Conversion Tasks on Speech and Singing Voice

Zero-Shot Voice Conversion

Model	English			Chinese		
	WER	SIM	UTMOS	WER	SIM	UTMOS
Ground Truth	2.15	-	3.52	1.25	-	2.79
FACodec [90]	4.88	0.355	2.79	8.20	0.495	2.10
Vevo-FM	3.01	0.536	3.51	4.06	0.695	2.86
CosyVoice2-FM [49]	4.66	0.530	**3.86**	2.80	0.728	**3.08**
NeuCoSVC 2 [174]	3.57	0.227	3.15	2.90	0.468	2.51
SeedVC (VC) [134]	**2.97**	0.565	3.31	**2.45**	0.737	2.69
Vevo2-FM	9.35	0.645	3.59	6.88	0.725	2.83
Vevo2	3.53	**0.692**	3.81	3.01	**0.755**	3.00

Vevo2-FM:
- Only using FM model
- Content-preserved, Melody-preserved, Style-preserved, Timbre-converted

Vevo2:
- Using both AR and FM models
- Content-preserved, Melody-preserved, Style-converted, Timbre-converted

Zero-Shot Singing Voice Conversion

Model	English					Chinese				
	WER	SIM	N-CMOS	Style-CMOS	Melody-MOS	WER	SIM	N-CMOS	Style-CMOS	Melody-MOS
Ground Truth	16.48	-	-	-	-	12.82	-	-	-	-
FACodec [90]	32.81	0.434	-	-	-	36.97	0.473	-	-	-
Vevo-FM	24.05	0.567	-0.64 $\pm$0.07	-0.45 $\pm$0.16	0.71 $\pm$0.21	22.85	0.610	-0.55 $\pm$0.19	-0.44 $\pm$0.08	0.88 $\pm$0.27
CosyVoice2-FM [49]	23.59	0.553	-0.64 $\pm$0.11	-0.63 $\pm$0.18	2.25 $\pm$0.15	20.14	0.589	-0.63 $\pm$0.15	-0.54 $\pm$0.07	1.65 $\pm$0.23
NeuCoSVC 2 [174]	30.61	0.481	-	-	-	32.26	0.519	-	-	-
SeedVC (SVC) [134]	22.86	0.508	-0.05 $\pm$0.13	-0.24 $\pm$0.15	2.89 $\pm$0.09	15.65	0.550	-0.33 $\pm$0.21	-0.54 $\pm$0.14	**2.93** $\pm$0.14
Vevo2-FM	29.82	0.587	-0.12 $\pm$0.16	-0.11 $\pm$0.12	**2.91** $\pm$0.12	22.54	0.611	-0.03 $\pm$0.30	-0.32 $\pm$0.17	2.90 $\pm$0.08
Vevo2	**11.64**	**0.601**	**0.00**	**0.00**	2.24 $\pm$0.18	**14.53**	**0.623**	**0.00**	**0.00**	2.38 $\pm$0.25

Key Findings

① **Timbre Conversion**
- Vevo2-FM achieves strong speaker / singer similarity in both VC and SVC. This validates **the effectiveness of the FM model for timbre conversion** while preserving content, style, and melody.

② **Melody Following**
- Vevo2-FM obtains competitive Melody-MOS in SVC. This shows that **chromagram-enhanced content-style tokens effectively capture melody information** for singing voice conversion.

③ **Vevo2-FM vs. Vevo2**
- Vevo2-FM excels at preserving source style and melody. Vevo2 achieves better intelligibility and style imitation by introducing AR-based content-style modeling.

Results: Editing Tasks on Speech and Singing Voice

Zero-Shot Speech Editing & Singing Lyric Editing

Model	Expressive Speech					Singing Voice				
	WER	SIM	FPC	N-CMOS	PS-CMOS	WER	SIM	FPC	N-CMOS	MS-CMOS
SSR-Speech [213]	29.56	0.588	0.721	-1.36 ± 0.09	-0.89 ± 0.14	51.92	0.710	0.804	-1.43 ± 0.17	-1.10 ± 0.08
F5-TTS [26]	21.35	0.733	0.730	-0.19 ± 0.12	-0.21 ± 0.16	29.78	0.784	0.821	-1.15 ± 0.23	-0.97 ± 0.19
Vevo2-base	23.54	0.795	0.782	-	-	29.69	0.841	0.872	-	-
Vevo2	**16.83**	**0.799**	**0.792**	**0.00**	**0.00**	**17.98**	**0.848**	**0.877**	**0.00**	**0.00**

Key Findings

① **Unified Editing across Speech and Singing**
- Vevo2 supports both speech editing and singing lyric editing **within the same framework**.
- It modifies the target text / lyrics while preserving the original voice characteristics.

② **Prosody-Preserved Editing**
- Vevo2 preserves the original speech prosody and singing melody, which verifies **the effectiveness of the unified prosody tokenizer for explicit prosody/melody control**.

③ **Better Intelligibility and Voice Quality**
- Compared with SSR-Speech and F5-TTS, Vevo2 achieves lower WER and higher SIM.
- The post-trained Vevo2 further improves intelligibility, naturalness, and prosody following; details are left to ***Part III***.

Impact and Recognition in the Field

Vevo2: A Unified and Controllable Framework for Speech and Singing Voice Generation

Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, Zhizheng Wu

Published in IEEE TASLP, 2026

(Cited and used as a baseline by CMU, NUS, Meta, Kuaishou, and others)

THE SINGING VOICE CONVERSION CHALLENGE 2025: FROM SINGER IDENTITY CONVERSION TO SINGING STYLE CONVERSION

Lester Phillip Violeta[1], Xueyao Zhang[2], Jiatong Shi[3], Yusuke Yasuda[4], Wen-Chin Huang[1], Zhizheng Wu[2], Tomoki Toda[1]

[1]Nagoya University, Japan, [2]The Chinese University of Hong Kong, Shenzhen, China, [3]Carnegie Mellon University, USA, [4]National Institute of Informatics, Japan

ICASSP 2026

Vevo2 serves as a primary baseline in Singing Voice Conversion Challenge 2025

INSTRUCTAUDIO: UNIFIED SPEECH AND MUSIC GENERATION WITH NATURAL LANGUAGE INSTRUCTION

Chunyu Qiang[1,2], Kang Yin[2], Xiaopeng Wang[2], Yuzhe Liang[2], Jiahui Zhao[1], Ruibo Fu[3], Tianrui Wang[1], Cheng Gong[1], Chen Zhang[2], Longbiao Wang[1†], Jianwu Dang[1]

[1] Tianjin University, Tianjin, China
[2] Kuaishou Technology, Beijing, China
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China

Kuaishou (ICASSP 2026)

*"Vevo2 introduced **the first unified speech and singing generation framework**, demonstrating that joint modeling leverages **rich speech data to improve singing quality while utilizing singing's expressive characteristics to enhance TTS.**"*

CARTOONSING: UNIFYING HUMAN AND NONHUMAN TIMBRES IN SINGING GENERATION

Jionghao Han[1], Jiatong Shi[1], Zhuoyan Tao[2], Yuxun Tang[3], Yiwen Zhao[1], Gus Xia[4], Shinji Watanabe[1]

[1] Carnegie Mellon University, [2] University of Southern California,
[3] Renmin University of China, [4] Mohamed bin Zayed University of Artificial Intelligence

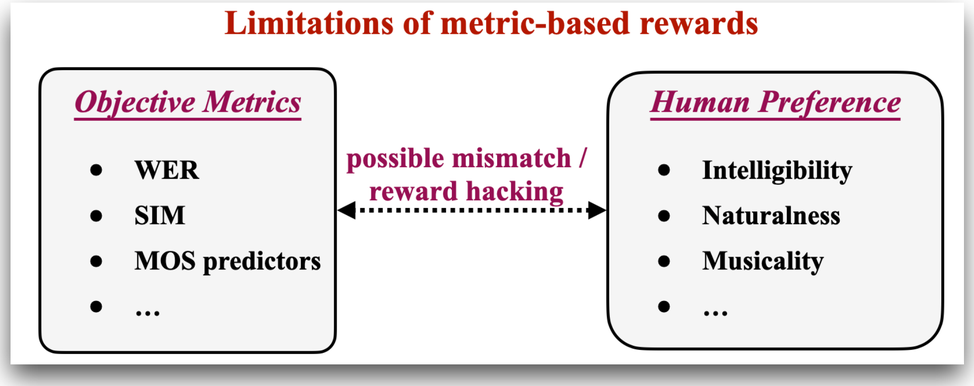
CMU

*"Vevo2 investigates humming-to-singing and instrument-to-singing, **with non-vocal inputs purely as melodic or prosodic guidance** rather than timbral conditioning."*

Contents

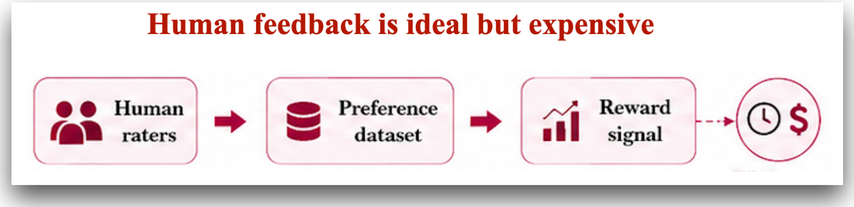
From Unified Voice Generation to Human-Aligned Voice Generation

1 RQ1: Preference Dataset Construction



How to build reliable human preference datasets?

2 RQ2: Reward Model Development



Can we build reliable reward models that approximate human perception?

3 RQ3: Preference-based Post-Training

Challenges in Algorithms *(i.e., How to Align?)*

- **Diverse generative architectures**
  - Auto-regressive based
  - Flow-Matching based
  - Masked Generative Model based
- **Multiple preference objectives**
  - Speech emphasizes intelligibility and naturalness.
  - Singing additionally requires melody quality and pitch accuracy.
  - A key challenge is ***multi-objective alignment*** without harmful trade-offs.

How to effectively post-train the voice generators base on the established preference data and reward models?

RQ1: Preference Dataset — From Automatic Labels to Human Judgements

INTP

Automatically constructed intelligibility preferences

- ✦ **250K** synthetic speech pairs across diverse domains
- ✦ Preference labels come from a **novel automatic pipeline**
- ✦ Focus: **intelligibility**

Scalable automatic preference labels

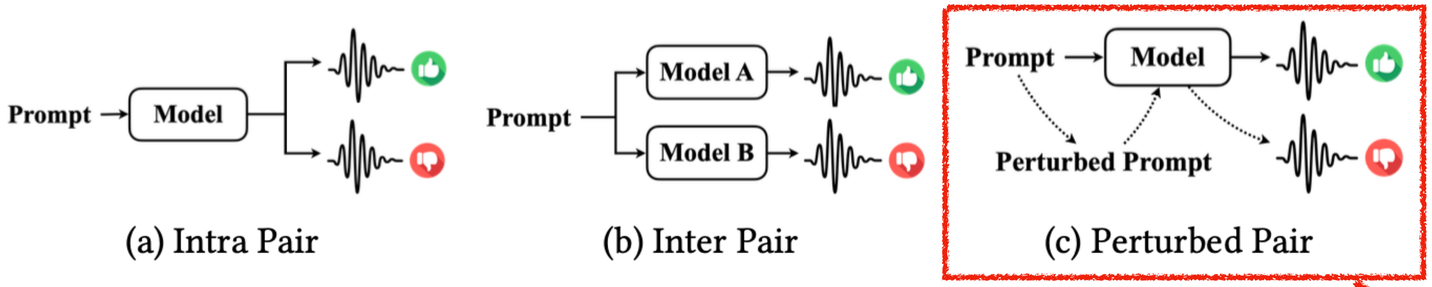
SpeechJudge-Data

Human-annotated naturalness preferences

- ✦ **247K** human preference choices on **99K** speech pairs
- ✦ Preference labels come from **real human listeners**
- ✦ Focus: **naturalness**

Reliable human preference labels

INTP: Intelligibility Preference Speech Dataset



(a) Intra Pair (b) Inter Pair (c) Perturbed Pair

Three Preference-Pair Construction Strategies

Text Type	Example
Regular	*A panda eats shoots and leaves.*
Repeated	*A panda panda eats shoots and leaves and leaves and leaves.*
Code-Switching	熊猫吃 *shoots* 和 *leaves*。
Pronunciation-perturbed	*A pan duh eights shots n leafs.*
Punctuation-perturbed	*A panda eats, shoots, and leaves.*

Diverse Challenging Text Types

DeepSeek-V3 are prompted to generate intelligibility-challenging text perturbations

System Prompt:
假设你是一个 Text To Speech (TTS) 领域的专家，现在，让我们对一个 TTS 系统进行攻击。具体地：我输入一个文本，请你修改这条文本里面的若干词语，从而使 TTS 系统更容易出错。例如：你可以修改为把某些字修改为容易读错的形近字、把多音字做替换，等等，但你不要增加和删除原有的文本。注意：你只需要返回给我转换后的结果，不需要任何解释。

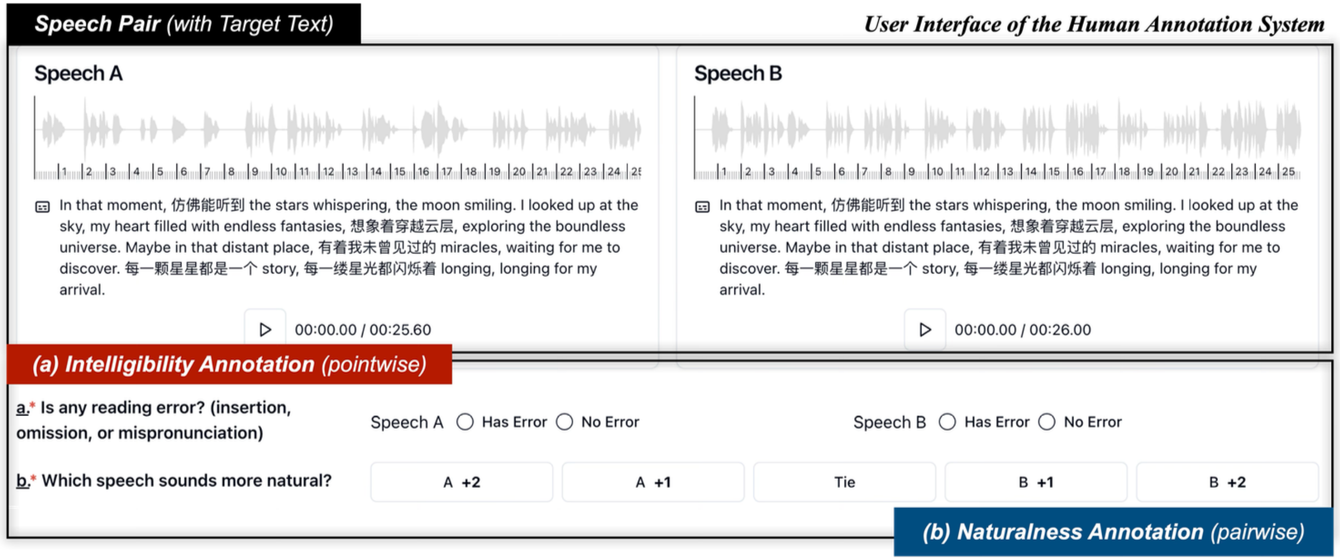
例子1:
【我的输入】我今天很高兴
【你的输出】窝锦添狠搞醒

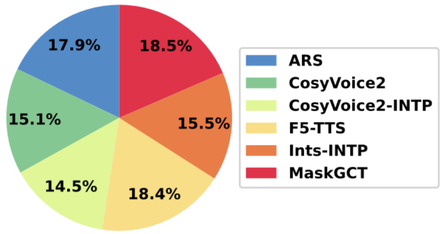
...

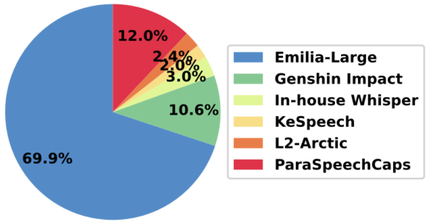
例子3:
【我的输入】And the idea of standing all by himself in a crowded market, to be pushed and hired by some big, strange farmer, was very disagreeable. Why not sing that high note and grow potatoes?
【你的输出】And the eye dear of standing awl bye himself in a crowd dead market, two bee pushed and high red buy sum big, strange far mer, was vary dis agreeable. Y knot sing that hi note and grow poe eight toes?

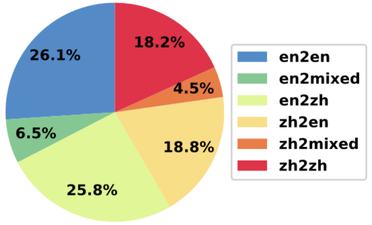
Key Findings: Human-guided perturbations enable scalable negative sample construction.

SpeechJudge-Data: Naturalness Preference Speech Dataset









Diverse Data Coverage

- Six TTS Models with diverse architectures
- Both regular and expressive speech cases
- Chinese, English, and code-switching settings

A Large-Scale Human Preference Dataset for Speech Naturalness

(99K pairs, with 2.49 annotations per pair on average)

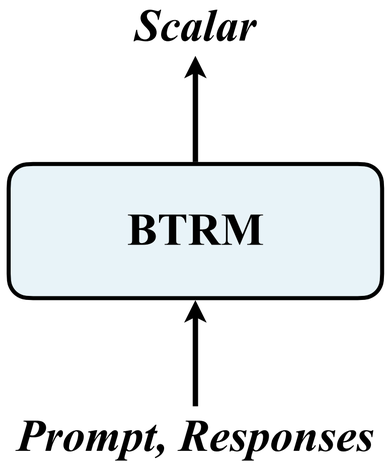
Dataset	Preference	Regular		Expressive			Total
		en	*zh*	*en*	*zh*	*mixed*	
SpeechJudge-Data (pref)	Binary	61.7 $\pm$0.3%	74.4 $\pm$0.2%	59.9 $\pm$0.8%	73.0 $\pm$0.6%	62.5 $\pm$0.5%	69.0 $\pm$0.2%
ImageReward [229]		–	–	–	–	–	65.3 $\pm$5.6%
InstructGPT [153]		–	–	–	–	–	72.6 $\pm$1.5%

High Annotation Quality

69.0% inter-annotator agreement, comparable to established human-preference datasets in text[1] and vision[2]

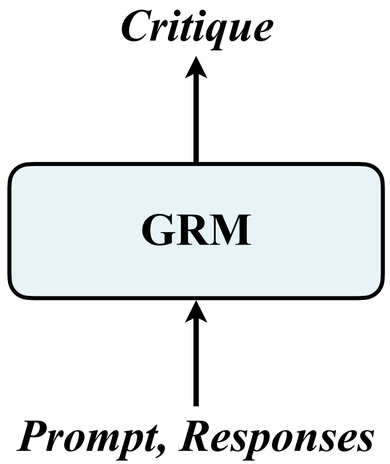
[1] Long Ouyang, et al.Training language models to follow instructions with human feedback. arXiv 2022. **OpenAI.** (**25K+ citations**)
[2] Jiazheng Xu, et al. ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation. NeurIPS 2023. (**1.4K+ citations**)

RQ2: Reward Model Development — From BTRM to GRM



Bradley-Terry Reward Model (BTRM)

- ⭐ Simpler pipeline, Easier training
- ⭐ Efficient inference
- ⭐ Deterministic scoring



Generative Reward Model (GRM) [1][2]

- ⭐ Native judgment, AudioLLM-ready
- ⭐ More compute, Better Performance
- ⭐ Clear critiques, Better explainability
- ⭐ Richer reasoning, Better generalization

[1] Lunjun Zhang, et al. Generative verifiers: Reward modeling as next-token prediction. ICLR 2025. **Google DeepMind.**
[2] Zijun Liu, et al. Inference-time scaling for generalist reward modeling. arXiv 2025. **DeepSeek-AI.**

Key Idea: Human Judgement as a Verifiable Reward

Chain-of-Thought Prompt

- We are comparing the naturalness of two text-to-speech models' outputs. The models need to speak the target text accurately and naturally.
- Target text: $\{\mathcal{T}\}$. Output A: $\{\mathcal{A}_1\}$, Output B: $\{\mathcal{A}_2\}$. Analyze the two output above, and **score them with number from 1 to 10**.
- Note:
  - (1) Please evaluate the naturalness of both audio outputs based on the following criteria: ***Prosody and Intonation***, ***Pacing and Rhythm***, ***Articulation and Clarity***, and ***Overall Naturalness***.
  - (2) After conducting a detailed analysis of each criterion, using the following output template to highlight your conclusion: Output A: X, Output B: X.

SpeechJudge-GRM

Rationale-based Output

Here's an analysis of the naturalness of Output A and Output B:

Output A:
- **Prosody and Intonation:** The intonation generally follows the sentence structure well, …
- **Pacing and Rhythm:** The pacing is consistent and relatively fast, but …
- **Articulation and Clarity:** Articulation is excellent…
- **Overall Naturalness:** Output A is clear and intelligible, demonstrating …

Output B:
- **Prosody and Intonation:** The prosody and intonation are highly unnatural…
- **Pacing and Rhythm:** The pacing is extremely inconsistent and erratic …
- **Articulation and Clarity:** While some individual sounds are correctly pronounced, the severe fragmentation of words…
- **Overall Naturalness:** Output B sounds highly artificial…

Conclusion: **Output A: 7.5, Output B: 2**

Preference Parse

Audio A
Audio B

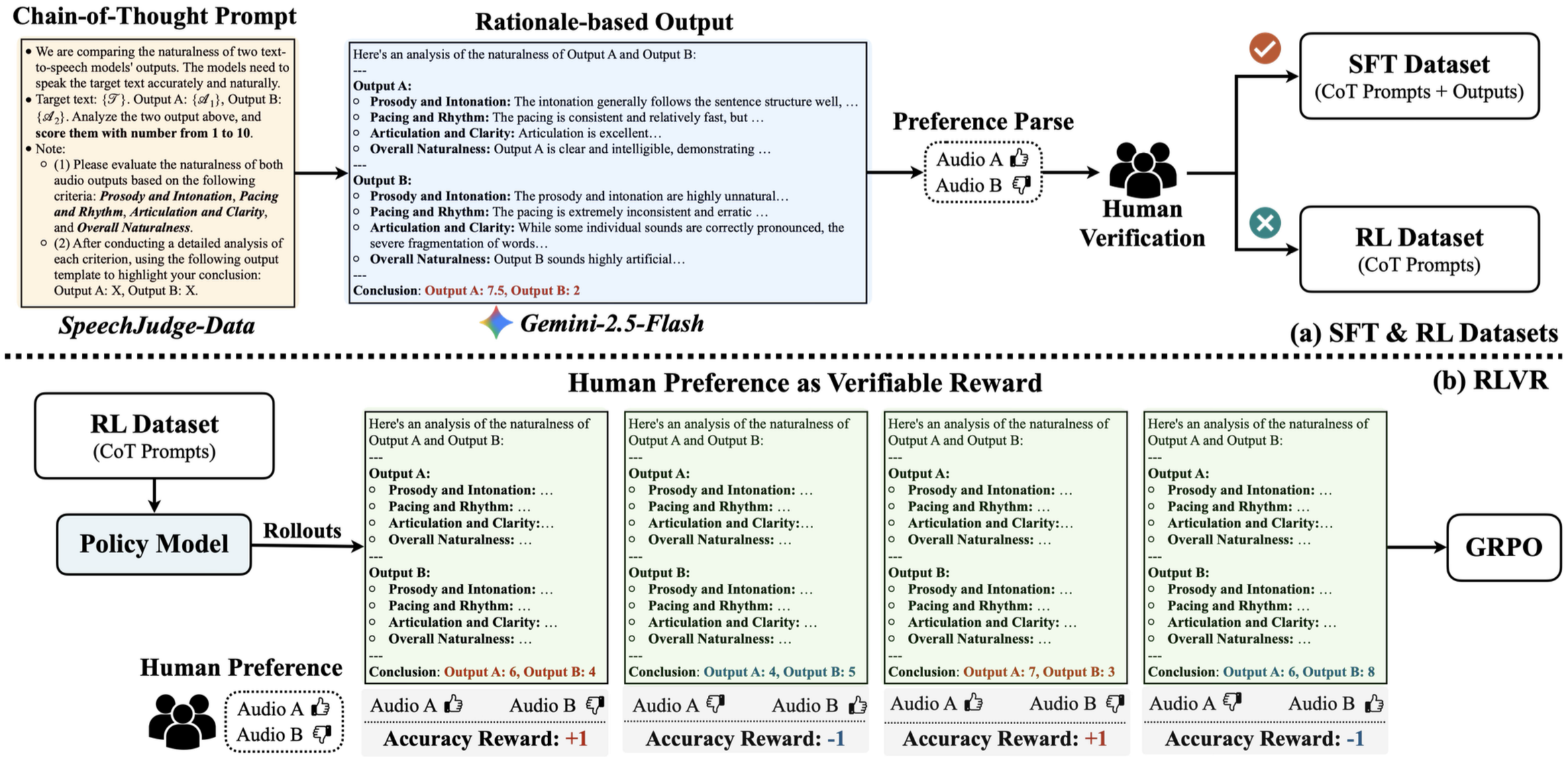
Human Verification

RLVR

Reinforcement Learning with Verifiable Rewards (RLVR) [1][2]

[1] Zhihong Shao, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv 2024. **DeepSeek-AI.**
[2] Daya Guo, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.. arXiv 2025. **DeepSeek-AI.**

SpeechJudge-GRM: Supervised Fine-Tuning (SFT) + RLVR



RQ3: Alignment Algorithms — From AR models to Others

Contribution 1: We propose the DPO extensions for FM-based and MGM-based models.

AR-based

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\left(\log\frac{p_\theta(y_w|x)}{p_{\text{ref}}(y_w|x)} - \log\frac{p_\theta(y_l|x)}{p_{\text{ref}}(y_l|x)}\right)\right)\right]$$

Vanilla DPO algorithm

FM-based

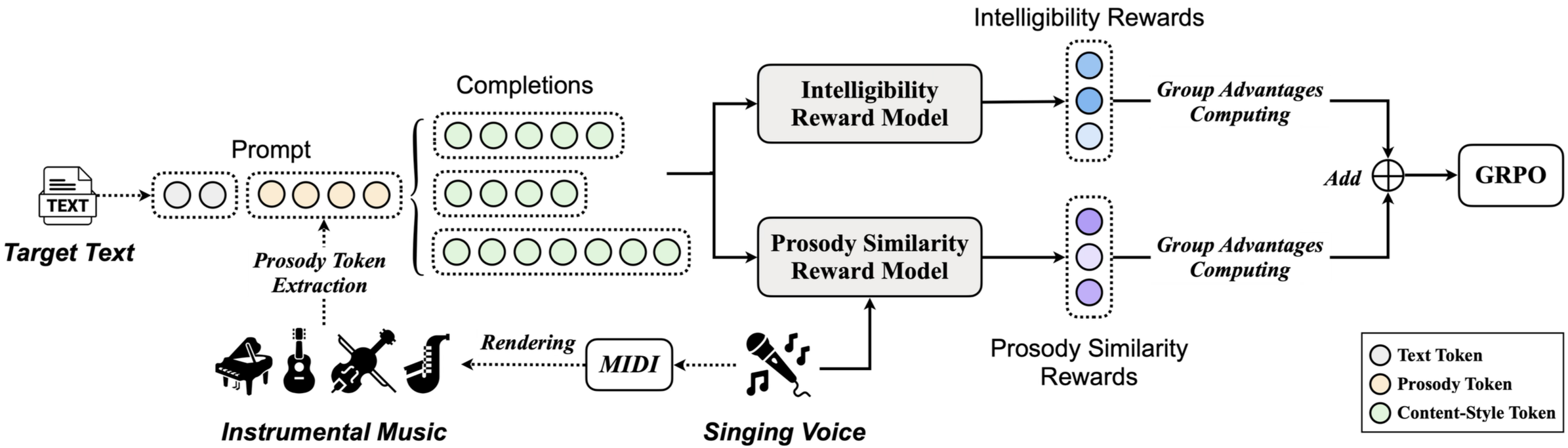
$$\mathcal{L}_{\text{DPO-FM}} = -\mathbb{E}_{(y_1^w, y_1^l, x)\sim\mathcal{D}, t}$$
$$\log\sigma\left(\beta\left(\log\frac{p_\theta(y_1^w|y_t^w, t, x)}{p_{\text{ref}}(y_1^w|y_t^w, t, x)} - \log\frac{p_\theta(y_1^l|y_t^l, t, x)}{p_{\text{ref}}(y_1^l|y_t^l, t, x)}\right)\right)$$

MGM-based

$$\mathcal{L}_{\text{DPO-MGM}} = -\mathbb{E}_{(y^w, y^l, x)\sim\mathcal{D}, t}$$
$$\log\sigma\left(\beta\left(\log\frac{p_\theta(y_0^w|y_t^w, x)}{p_{\text{ref}}(y_0^w|y_t^w, x)} - \log\frac{p_\theta(y_0^l|y_t^l, x)}{p_{\text{ref}}(y_0^l|y_t^l, x)}\right)\right)$$

RQ3: Alignment Algorithms — From Single-Objective to Multi-Objective

Contribution 2: We propose the multi-objective alignment algorithm based on GRPO.



Multi-objective alignment for both intelligibility and prosody similarity based on Vevo2

Results: Effectiveness of INTP

Zero-Shot TTS models of Different Architectures (before and after the INTP alignment)

Model	Regular cases			Articulatory cases			Code-switching cases			Cross-lingual cases			Avg		
	WER	SIM	N-CMOS	WER	SIM	N-CMOS	WER	SIM	N-CMOS	WER	SIM	N-CMOS	WER	SIM	N-CMOS
ARS	3.96	0.717	-	20.03	0.693	-	54.15	0.693	-	19.76	0.630	-	24.47	0.683	-
w/ INTP	2.32	0.727	0.47 ± 0.22	12.83	0.713	0.64 ± 0.31	36.91	0.698	0.63 ± 0.34	9.57	0.632	0.82 ± 0.28	15.41	0.692	0.64 ± 0.12
F5-TTS	3.44	0.670	-	16.84	0.635	-	33.99	0.609	-	16.86	0.546	-	17.78	0.615	-
w/ INTP	2.38	0.652	0.38 ± 0.26	12.97	0.628	0.30 ± 0.23	15.98	0.576	0.67 ± 0.36	7.13	0.509	0.47 ± 0.30	9.62	0.591	0.44 ± 0.12
MaskGCT	2.34	0.738	-	12.43	0.714	-	29.06	0.696	-	12.34	0.629	-	14.04	0.694	-
w/ INTP	2.23	0.737	0.23 ± 0.20	9.13	0.722	0.57 ± 0.36	19.70	0.704	0.19 ± 0.16	7.87	0.633	0.29 ± 0.18	9.73	0.699	0.32 ± 0.15
CosyVoice 2	2.09	0.709	-	8.12	0.696	-	33.36	0.672	-	8.78	0.600	-	13.09	0.669	-
w/ INTP	1.65	0.709	0.24 ± 0.25	6.87	0.696	0.20 ± 0.16	28.31	0.671	0.63 ± 0.30	5.39	0.603	0.28 ± 0.31	10.56	0.670	0.33 ± 0.12
Ints	3.14	0.688	-	12.08	0.666	-	22.88	0.646	-	9.78	0.572	-	11.97	0.643	-
w/ INTP	2.36	0.686	0.20 ± 0.36	9.38	0.664	0.11 ± 0.22	13.80	0.642	0.20 ± 0.38	6.28	0.571	0.18 ± 0.23	7.96	0.641	0.17 ± 0.15

- **AR-based:** ARS, **CosyVoice 2**, **Ints**, **FM-based:** F5-TTS, **MGM-based:** MaskGCT
- **CosyVoice2** and **Ints** have not participated in the INTP construction.

Key Findings

① **INTP alignment improves zero-shot TTS across different architectures:**
- After INTP alignment, all evaluated models show clear improvements across diverse cases. The gains are consistent across AR-based, FM-based, and MGM-based TTS models, demonstrating **the broad applicability of INTP**.

② **Weak-to-strong generalization of INTP:**
- INTP remains effective even for stronger and more intelligible models, such as **CosyVoice 2** and **Ints**. This suggests that INTP provides **transferable preference signals** rather than only overfitting to the models used to build the dataset.

Results: Effectiveness of SpeechJudge-Data and SpeechJudge-GRM

Agreement (%) of Different Reward Models with Human Preferences

Model	Regular	Expressive	Total
Qwen2.5-Omni-7B	62.0	59.7	60.6
Gemini-2.5-Flash	73.5	66.2	69.1
SpeechJudge-BTRM	77.5	69.5	72.7
SpeechJudge-GRM (SFT)	77.8	73.7	75.3
w/ Voting@10	77.4	77.6	77.6
SpeechJudge-GRM (SFT+RL)	79.0	76.0	77.2
w/ Voting@10	80.5	78.7	79.4

- **w/ Voting@10:** For each prompt, the GRM generates 10 outputs, and the final decision is obtained by **majority voting**.

Subjective Evaluation of Vevo2 Before and After SpeechJudge Alignment

Model	T-ACC	N-CMOS
Vevo2-base	84.0%	0.00
w/ INTP	87.0%	0.18 $\pm$0.07
w/ SpeechJudge-Data	91.0%	0.16 $\pm$0.08
w/ SpeechJudge-GRM (offline)	91.0%	0.21 $\pm$0.12
w/ SpeechJudge-GRM (online)	90.0%	0.25 $\pm$0.09

- **w/ SpeechJudge-GRM (offline):** We use the SpeechJudge-GRM as an **offline data annotator**
- **w/ SpeechJudge-GRM (online):** We use the SpeechJudge-GRM as an **online reward scorer**

T-ACC: Subjective intelligibility metric (Text Accuracy)
N-CMOS: Subjective naturalness metric (Naturalness CMOS)

Key Findings

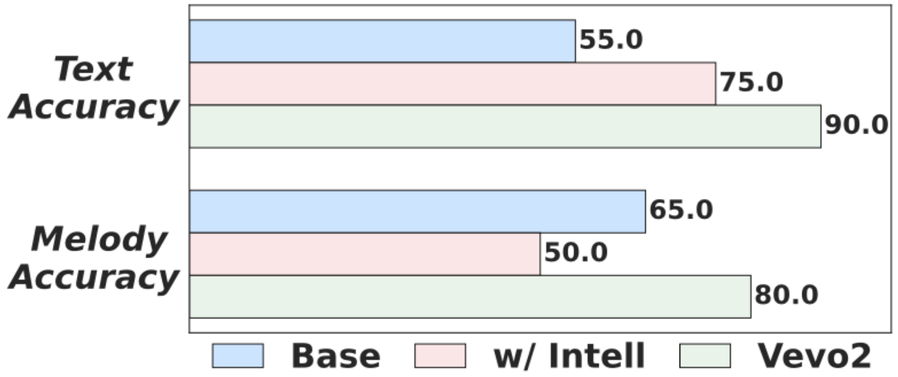
① **SpeechJudge-GRM better aligns with human preferences**
- SpeechJudge-BTRM achieves higher agreement with human judgments than existing AudioLLMs.
- SpeechJudge-GRM further improves agreement **through CoT-based SFT and RLVR**. Besides, with **inference-time scaling**, Voting@10 boosts the final agreement to 79.4%.

② **SpeechJudge-based alignment improves Vevo2 generation quality**
- Alignment with INTP, SpeechJudge-Data, and SpeechJudge-GRM consistently improves Vevo2's intelligibility and naturalness.
- **SpeechJudge-GRM online alignment** achieves the best naturalness improvement.

Results: Effectiveness of Multi-Objective Alignment for Vevo2

Model	*Humming-to-Singing*			*Instrument-to-Singing*		
	WER	SIM	FPC	WER	SIM	FPC
Base	32.86	0.534	0.769	40.38	0.503	0.716
w/ Intell	17.49	0.581	0.774	20.03	0.502	0.731
w/ Prosody	17.27	0.585	0.784	17.94	0.525	0.745

* The first and last model represent Vevo2-base and Vevo2, respectively. *w/ Intell*: Only intelligibility reward is used for post-training.



Subjective Results

Key Findings

① **Single-objective alignment improves intelligibility but causes trade-offs**
- Using only the intelligibility reward significantly improves WER and Text Accuracy. However, it can **hurt melody-following ability**, with Melody Accuracy dropping below the pre-trained model.

② **Benefits of prosody modeling for intelligibility**
- Jointly optimizing intelligibility and prosody rewards improves both Text Accuracy and Melody Accuracy. This suggests that **stronger prosody modeling can also benefit pronunciation and intelligibility**.

③ **Vevo2 benefits from multi-objective post-training**
- Compared with single-objective alignment, **multi-objective alignment better matches the requirements of unified speech and singing generation**. It improves controllability without sacrificing one perceptual objective for another.

Impact and Recognition in the Field

Advancing Zero-shot Text-to-Speech Intelligibility across Diverse Domains via Preference Alignment

Xueyao Zhang[*,1], **Yuancheng Wang**[*,1], **Chaoren Wang**[1], **Ziniu Li**[1], **Zhuo Chen**[2], **Zhizheng Wu**[1]
[1]The Chinese University of Hong Kong, Shenzhen
[2]ByteDance Seed

SPEECHJUDGE: TOWARDS HUMAN-LEVEL JUDGMENT FOR SPEECH NATURALNESS

Xueyao Zhang[1*] **Chaoren Wang**[1*] **Huan Liao**[1] **Ziniu Li**[1] **Yuancheng Wang**[1]
Li Wang[1] **Dongya Jia**[2] **Yuanzhe Chen**[2] **Xiulin Li**[3] **Zhuo Chen**[2] **Zhizheng Wu**[1,4,5,6†]
[1]The Chinese University of Hong Kong, Shenzhen [2]ByteDance Seed [3]DataBaker Technology
[4]Shenzhen Loop Area Institute [5]City University of Macau [6]Amphion Technology Co., Ltd

Published at ACL 2025 and ICLR 2026

(Cited by Meta, MIT, THU, Tencent, Xiaomi, and others)

MEASURING PROSODY DIVERSITY IN ZERO-SHOT TTS: A NEW METRIC, BENCHMARK, AND EXPLORATION

Yifan Yang[1*], *Bing Han*[1*], *Hui Wang*[3], *Long Zhou*[2†], *Wei Wang*[1], *Mingyu Cui*[2], *Xu Tan*[2], *Xie Chen*[1,4‡]

[1]X-LANCE Lab, MoE Key Lab of Artificial Intelligence, Jiangsu Key Lab of Language Computing, Shanghai Jiao Tong University [2]Tencent Hunyuan [3]Nankai University [4]Shanghai Innovation Institute

Tencent & SJTU

*"We evaluate the prosody diversity of zero-shot TTS systems from [32], which are aligned via **Direct Preference Optimization (DPO) on the INTP dataset for intelligibility**, using vanilla DPO for AR and **extended DPO for NAR MGM**."*

GSRM: Generative Speech Reward Model for Speech RLHF

Maohao Shen[1,2,*,†], **Tejas Jayashankar**[1,*], **Osama Hanna**[1,*], **Naoyuki Kanda**[1,*], **Yancheng Wang**[1,3,†], **Kateřina Žmolíková**[1], **Ruiming Xie**[1], **Niko Moritz**[1], **Anfeng Xu**[1,4,†], **Yashesh Gaur**[1], **Gregory Wornell**[2], **Qing He**[1], **Jilong Wu**[1]

[1]Meta Superintelligence Labs, [2]Massachusetts Institute of Technology, [3]Arizona State University, [4]University of Southern California

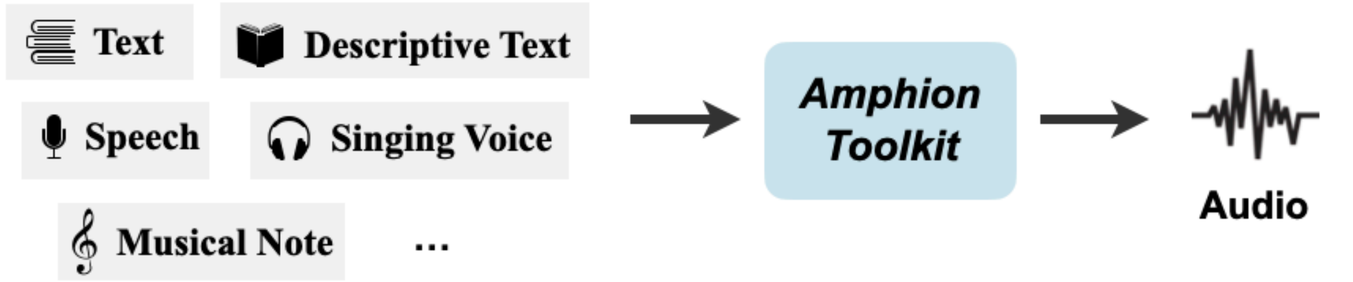
Meta & MIT

*"SpeechJudge **extends generative reward modeling to judge speech naturalness**. ... SpeechJudge synthesizes CoT using a teacher speech LLM directly conditioned on the raw audio signal, **resulting in reasoning that is mediated by implicit acoustic representations the speech LLM can capture**."*

Contents

Amphion: An Open-Source Audio, Music, and Speech Generation Toolkit

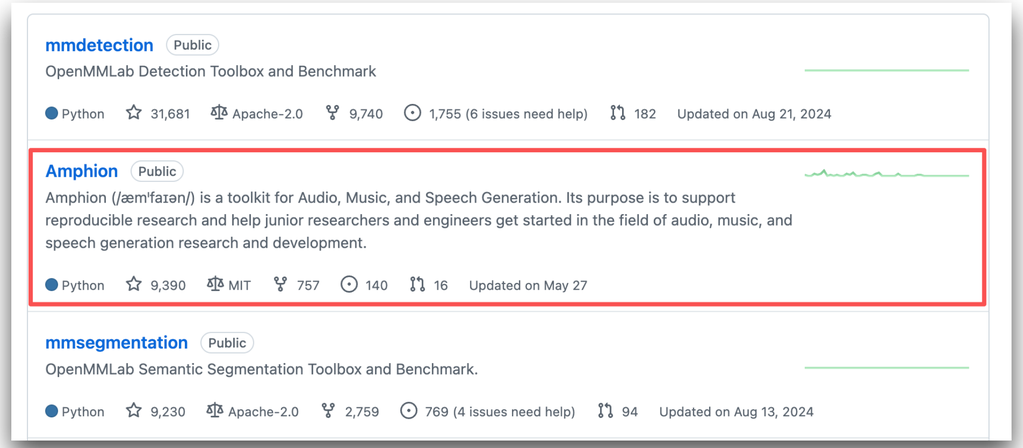
- Support **reproducible research** and help **junior researchers and engineers** get started in the field of audio, music, and speech generation research and development.



Our North-Star Objective:

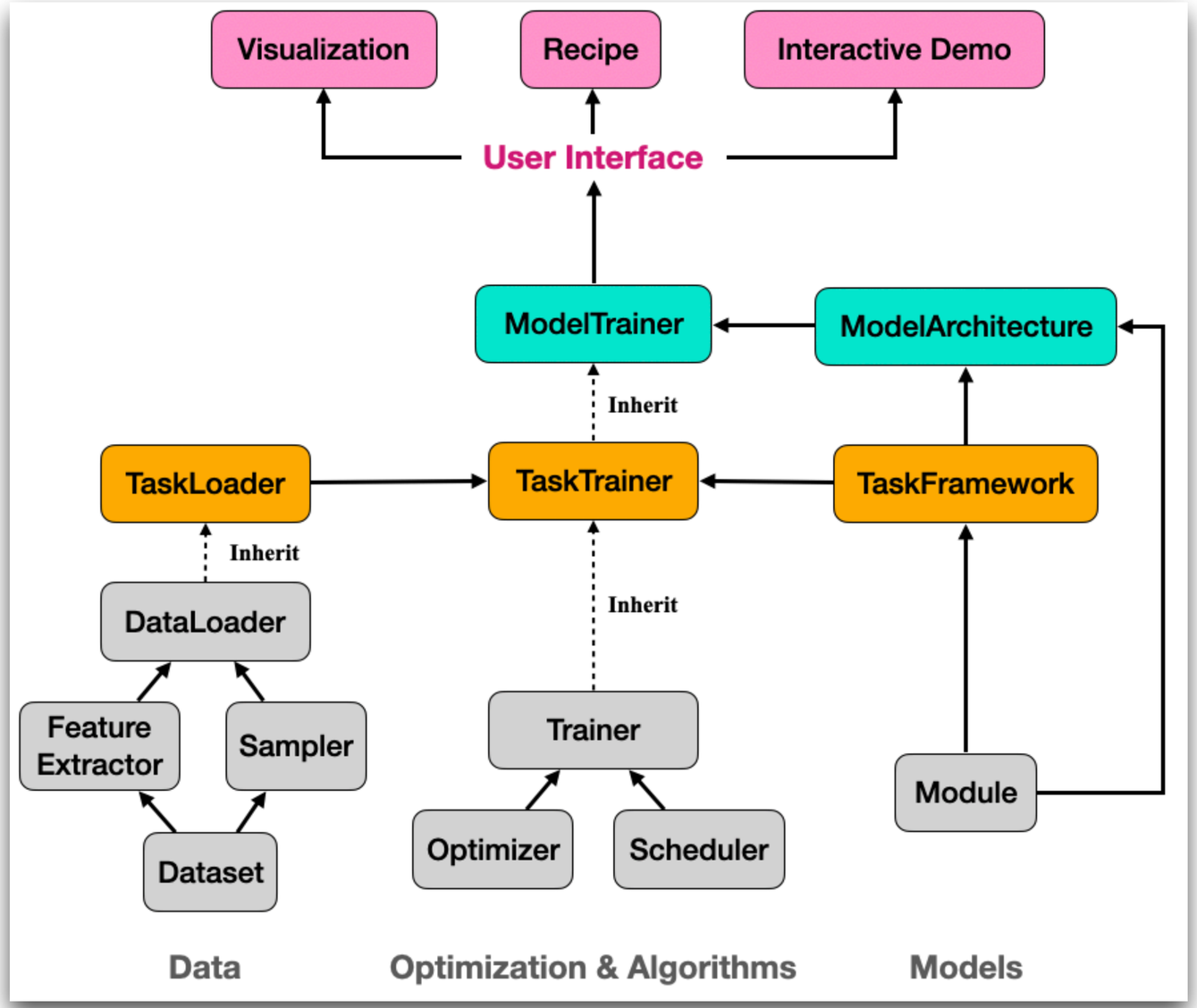
Any to Audio

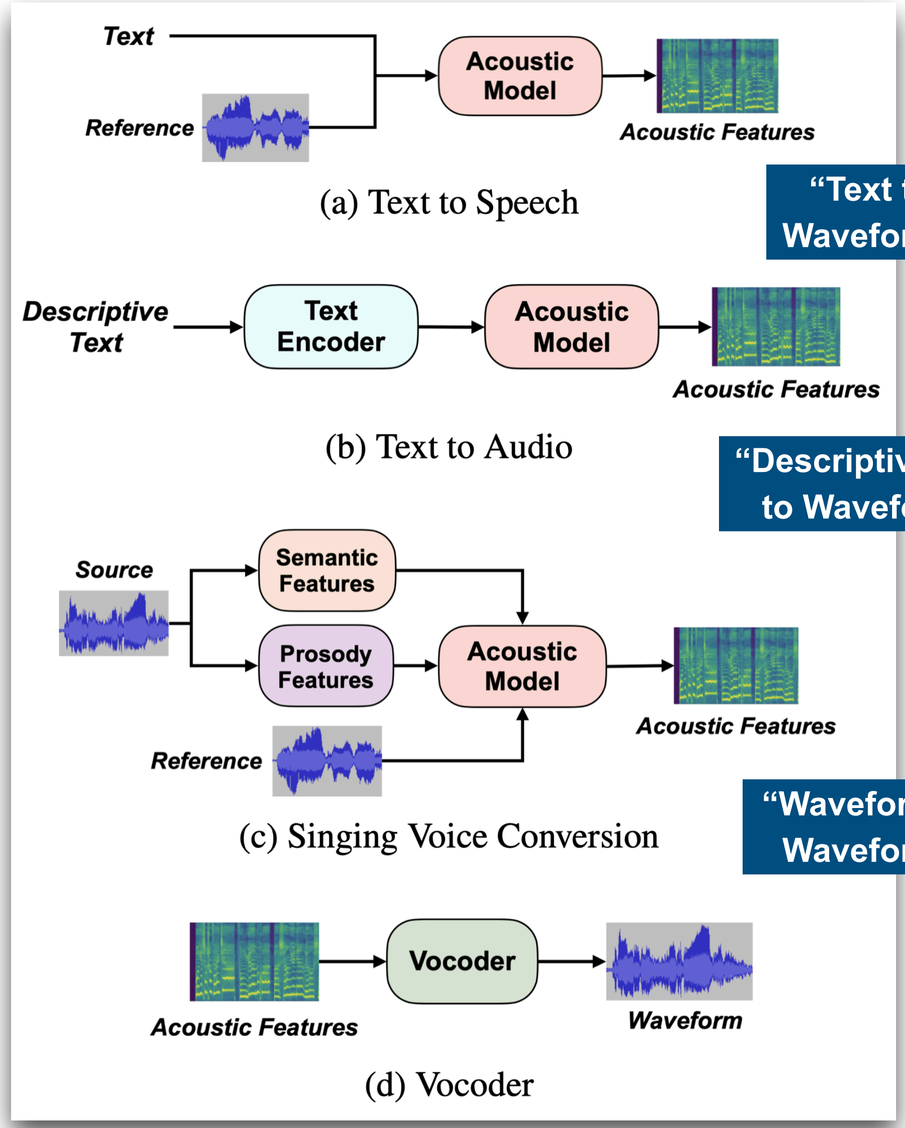
- **TTS**: Text to Speech (⛳ supported)
- **SVS**: Singing Voice Synthesis (⛳ supported)
- **VC**: Voice Conversion (⛳ supported)
- **AC**: Accent Conversion (⛳ supported)
- **SVC**: Singing Voice Conversion (⛳ supported)
- **TTA**: Text to Audio (⛳ supported)
- **TTM**: Text to Music (👷 developing)
- more...



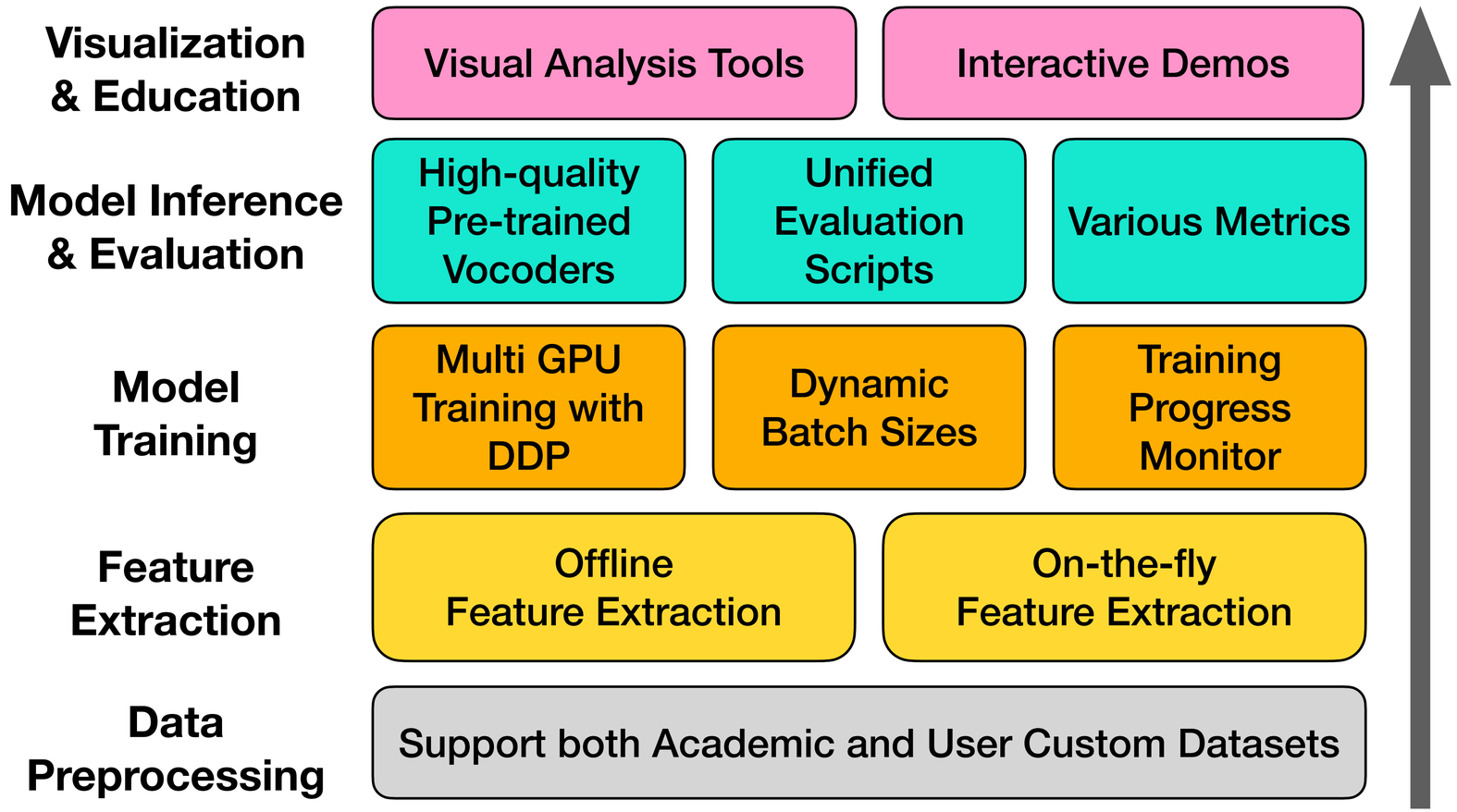
>9.8K stars at GitHub, Top 2 of OpenMMLab

Strength 1: Unified Audio Generation Framework



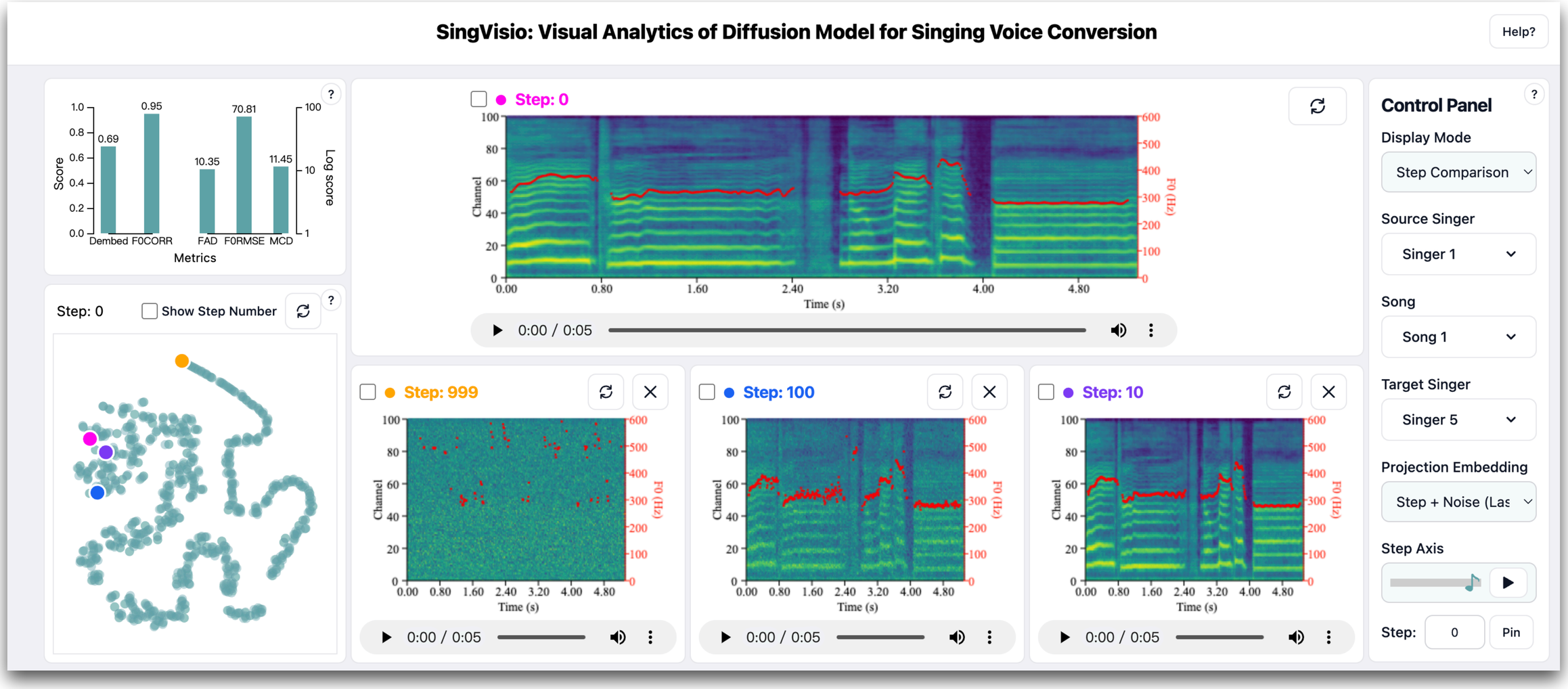


Strength 2: Beginner-friendly End-to-End Workflow



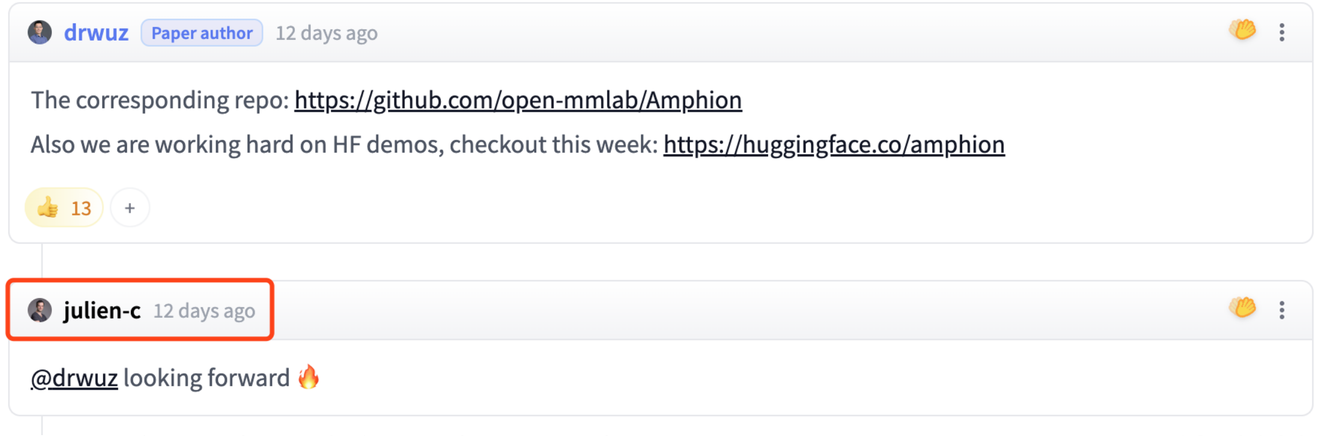
One-stop research platform suitable for both novices and experienced researchers

Strength 3: Visualization and Interactivity



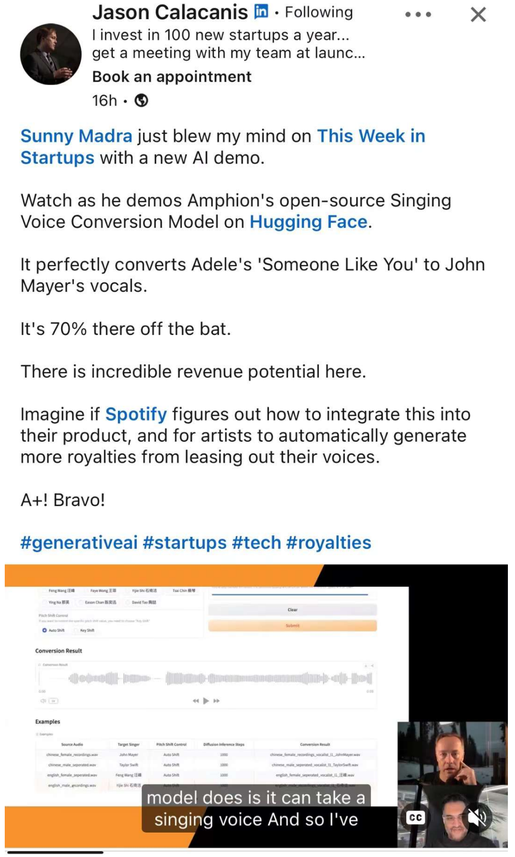
Liumeng Xue*, Chaoren Wang*, Mingxuan Wang, Xueyao Zhang, Jun Han, Zhizheng Wu. *SingVisio: Visual Analytics of Diffusion Model for Singing Voice Conversion.* Computers & Graphics.

Recognition in the Field



- **Amphion v0.1 and its technical report were released on Dec. 18, 2023. The report ranked No. 1 on Hugging Face Daily Papers and attracted attention from Hugging Face co-founder and CTO Julien Chaumond.**

- **AK, a major AI community influencer on Twitter/X, created a dedicated Hugging Face Slack channel to support Amphion.**



- **On Dec. 19, 2023, Silicon Valley investor Jason Calacanis featured and highly praised Amphion's singing voice conversion demo on his show *This Week in Startups*.**

Recognition in the Field

霉霉演唱《稻香》，国内团队的Amphion音频生成火了

机器之心 2023-12-20 14:48 Posted on 北京

机器之心专栏

机器之心编辑部

> 香港中文大学（深圳）数据科学学院武执政副教授团队联合上海人工智能实验室 OpenMMLab 团队开源了综合音频生成项目 Amphion（安菲翁）。该系统旨在打造一个集语音合成转换、歌声合成转换、音效音乐生成等多功能为一体的开源平台。截至目前，Amphion 已经多次进入 GitHub Trending Repositories 榜单。

2022 年被称为 AIGC 元年，ChatGPT、Stable Diffusion、MidJourney 为代表的文字、图像应用带火了 AI 领域。2023 年，AI 孙燕姿、AI 郭德纲、音效生成、音乐生成也在社交媒体上火了一把。

让泰勒·斯威夫特唱周杰伦的歌？来自深圳龙岗港中大团队的Amphion音频生成火了！

深圳特区报记者 罗实宜 文/图

12-21 22:11

深圳特区报

深圳市委机关报，深圳经济特区权威媒体和第一大报

近日，香港中文大学（深圳）数据科学学院武执政副教授团队联合上海人工智能实验室 OpenMMLab 团队开源了综合音频生成项目 Amphion（安菲翁）。该系统旨在打造一个面向科研群体及刚进入或想要进入该领域的工程师的，集语音合成及转换、歌声合成及转换、音效及音乐生成等多功能为一体的开源平台。目前，该研究已经在海外社交平台上引发了极大的关注。

Amphion: An Open-Source Audio, Music and Speech Generation Toolkit

MARKTECHPOST

ML News | LLMs | Other AI News | AI Tools | Free AI Courses | About Us

Home › Technology › AI Shorts › Meet Amphion: An Open-Source Audio, Music and Speech Generation AI Toolkit

Meet Amphion: An Open-Source Audio, Music and Speech Generation AI Toolkit

By **Adnan Hassan** - December 21, 2023

Reddit | Y | | in | | 0 SHARES

In the dynamic landscape of artificial intelligence, audio, music, and speech generation has undergone transformational strides. As open-source communities thrive, numerous toolkits emerge, each contributing to the expanding repository of algorithms and techniques. Among these, one standout, Amphion, by researchers from The Chinese University of Hong Kong, Shenzhen, Shanghai AI Lab, and Shenzhen Research Institute of Big Data, takes center stage with its unique features and commitment to fostering reproducible research.

香港中文大学深圳

Yesterday 19:54

港中大（深圳）音乐学院钢琴与键盘学部 2023-2024 学年度秋季...

香港中文大学（深圳）国家级一流本科课程遴选推荐的公示

大湾区新闻

2023-12-22 来自 微博视频号

香港中文大学（深圳）数据科学学院武执政副教授团队联合上海人工智能实验室 OpenMMLab 团队开源了综合音频生成项目 Amphion（安菲翁）。该系统旨在打造一个集语音合成转换、歌声合成转换、音效音乐生成等多功能为一体的开源平台。目前，该研究已经在海外社交平台上引发了极大的关注。 大湾区新闻的微博视频

Recognition in the Field

- Invited by leading Chinese AI research communities and media platforms, including SpeechHome, BAII Talk, and JiangMen, to introduce Amphion.
- Invited talks at SLT 2024, HKUST(GZ), and Nankai University.

Contents

Key Contributions of This Thesis

1 Disentangled Representation Learning for Controllable Speech Generation

Contribution 1: Self-supervised disentanglement enables controllable and zero-shot speech generation.

- ✦ *Xueyao Zhang, et al. Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement.* ***ICLR 2025****.*

2 Unified Voice Modeling across Speech and Singing Voice

Contribution 2: A unified and controllable framework for speech and singing voice generation.

- ✦ *Xueyao Zhang, et al. Vevo2: A Unified and Controllable Framework for Speech and Singing Voice Generation.* ***TASLP 2026****.*

3 Human-aligned Voice Generation through Preference Alignment

Contribution 3: Alignment improves intelligibility, naturalness, and melody quality for voice generation.

- ✦ *Xueyao Zhang*, et al. Advancing Zero-shot Text-to-Speech Intelligibility across Diverse Domains via Preference Alignment.* ***ACL 2025****.*
- ✦ *Xueyao Zhang*, et al. SpeechJudge: Towards Human-Level Judgment for Speech Naturalness.* ***ICLR 2026****.*

4 Open-Source Ecosystem for Broader Impact

Contribution 4: Amphion releases models, datasets, tools, and demos for reproducible and extensible research.

- ✦ *Xueyao Zhang*, et al. Amphion: An Open-Source Audio, Music and Speech Generation Toolkit.* ***SLT 2024****.*
- ✦ *Xueyao Zhang, et al. Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion.* ***SLT 2024****.*

**: Co-first author*

Acknowledgement

Many thanks to all my labmates for their support and encouragement!

Thanks for listening!

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

SCHOOL OF
DATA SCIENCE
數據科學學院