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

Controllable and Unified Speech and Singing Voice Generation

Xueyao Zhang

The Chinese University of Hong Kong, Shenzhen

2025/04

About me

Xueyao Zhang (张雪遥)

✦ **Third-year PhD student**, Supervised by Prof Zhizheng Wu
School of Data Science, CUHK-Shenzhen
Homepage: https://www.zhangxueyao.com/

✦ **Amphion co-founder**
Project: https://github.com/open-mmlab/Amphion **(9k stars)**

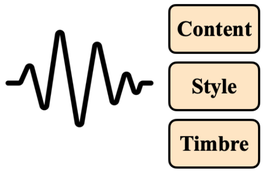
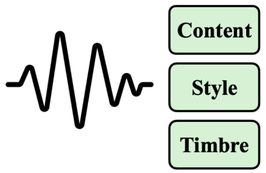
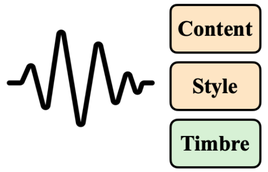
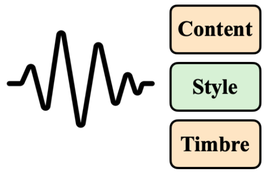
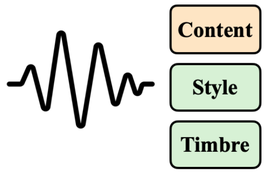
✦ **Research interest**
- Speech Generation
- AI Music

📎 **Vevo *(ICLR 2025)*: Controllable and Unified Speech Generation** ***[Paper]*** ***[Code]*** ***[Model]***

👷 **Vevo2 (*Ongoing*): Extend Vevo to Singing Voice Generation** ***[Blog]*** ***[Code]*** ***[Model]***

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

Vevo: Controllable Zero-Shot Voice Imitation with Self-Supervised Disentanglement

Task	Source	Reference	Target	Related Areas
Zero-Shot Timbre Imitation				Voice Conversion
Zero-Shot Style Imitation				Accent Conversion, Emotion Conversion
Zero-Shot Voice Imitation				Voice Conversion Text to Speech

Content — What to speak?

Style — How to speak?

Timbre — Who speaks?

🤔 Question 1

How can we accomplish various zero-shot imitation tasks using a unified framework?

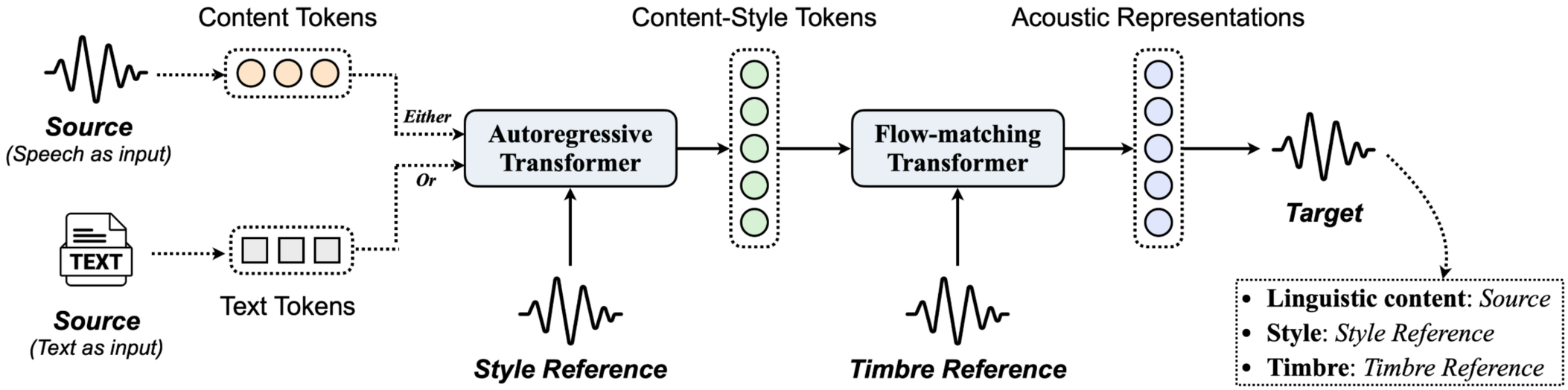
🤔 Question 2

How can we minimize dependency on annotated data to maximize the benefits of large-scale self-supervised learning?

🤔 Question 3

How can we effectively decouple timbre, style, and content to achieve controllable generation?

Inference Pipeline: Single forward pass, Versatile tasks



Model	Source	Style Reference	Timbre Reference
Vevo-Timbre	U_i	/	U_r
Vevo-Style	U_i	U_r	U_i
Vevo-Voice	U_i	U_r	U_r
Vevo-TTS	T_i	U_r	U_r

[Input] Speech (U_i) or Text (T_i)
[Prompt] A reference speech (U_r)

★ **Unify TTS and VC**

Credit to **content-style tokens** that *decouple timbre*

★ **Identity-preserving zero-shot style conversion**

Credit to **content tokens** that can be considered as *pseudo texts*.

Key idea: Codebook size as the disentanglement bottleneck

Representations	#Vocab	WER (↓)	S-SIM (to ref) (↑)	S-SIM (to src) (↓)	FPC (to src) (↑)	Analysis
Ground Truth	-	5.526	0.762	0.087	1.000	-
24th layer features	-	5.706	0.266	0.400	0.768	***Pros***: Intelligibility, Style consistency ***Cons***: Timbre imitation
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	***Pros***: Intelligibility, Timbre imitation ***Cons***: Style consistency
ASR tokens	29	7.836	0.463	0.125	0.698	
K-means tokens	1024	11.493	0.398	0.150	0.734	Worse than VQ-VAE tokens (1024)
VQ-VAE tokens	16384	6.807	0.398	0.306	0.826	As the vocabulary size decreases, ***Pros***: Timbre imitation ↑ ***Cons***: Intelligibility ↓ Style consistency ↓
	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 ↓	
	16	13.169 ↑	0.441	0.146 ↓	0.672 ↓	
	8	21.813 ↑	0.392	0.109 ↓	0.675	

Starting point of information filtering

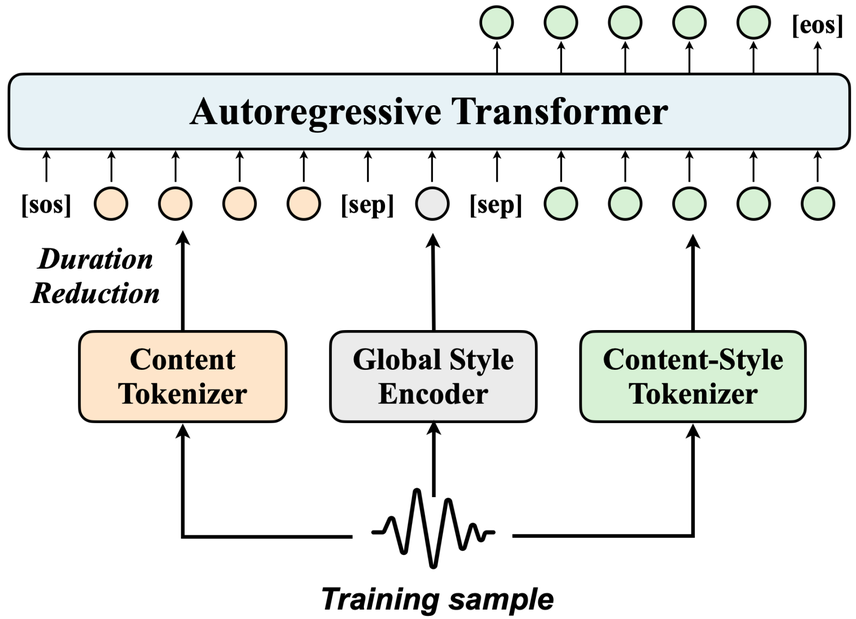
Content-style Tokens

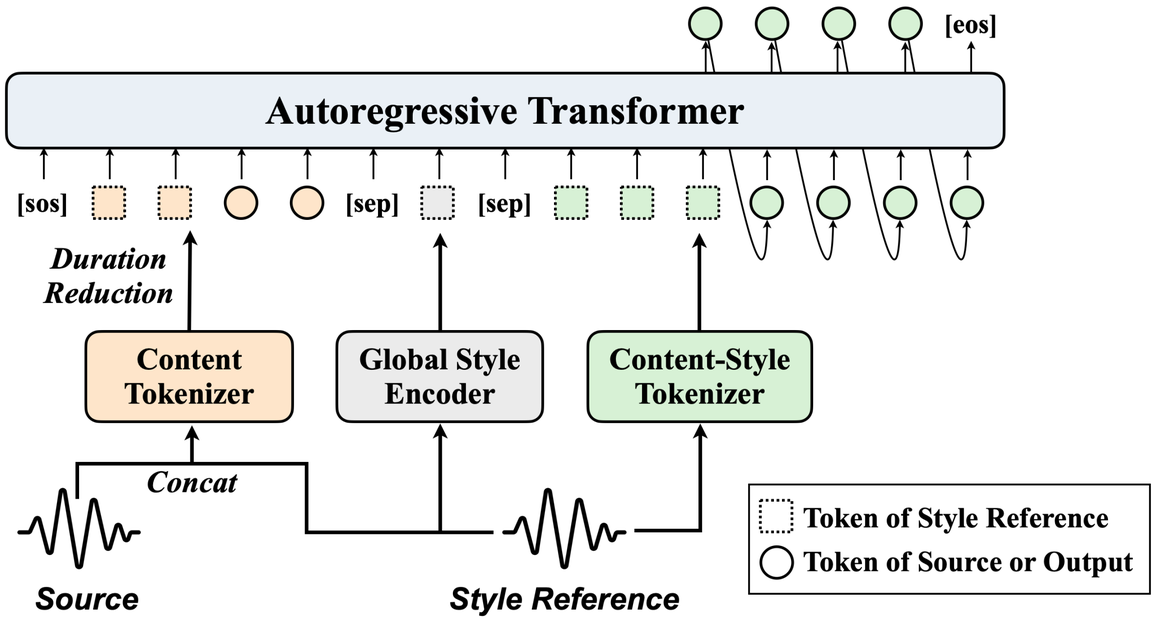
Content Tokens

① From **hidden features** to **content-style tokens**, we can see that **the timbre leakage issue is significantly mitigated**. (S-SIM to ref/src: 0.250/0.505 -> 0.403/0.236). Meanwhile, **style information is largely preserved** (FPC: 0.824 -> 0.797).

② From **content-style tokens** to **content tokens**, we observe **substantial removal of style information** (FPC: 0.797 -> 0.706).

Content-Style Modeling *(Content to Content-Style)*





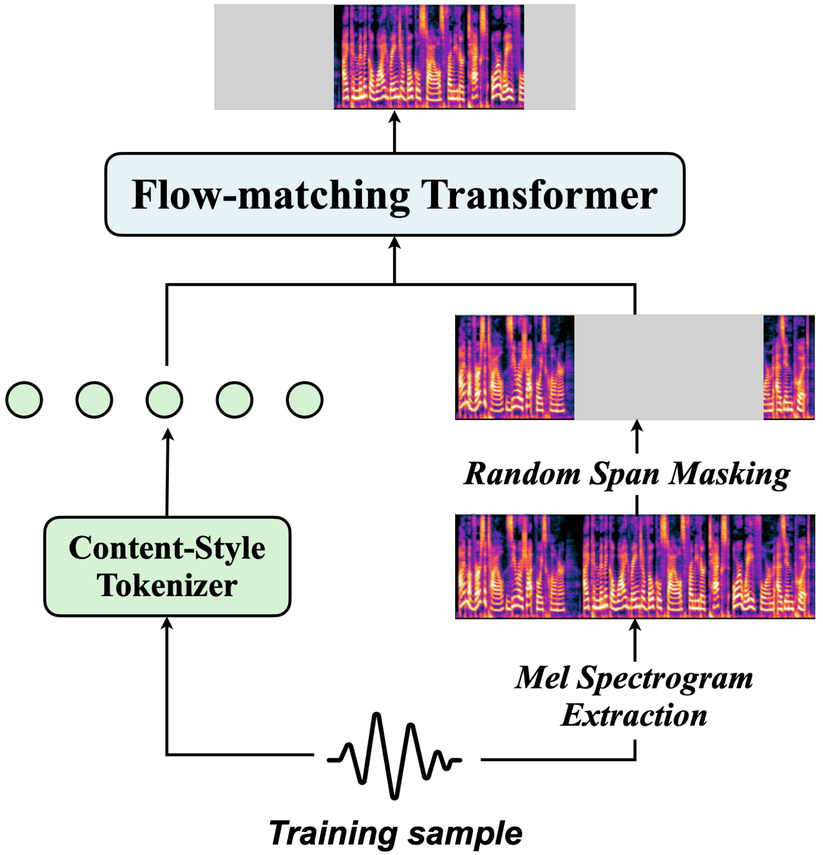
Training ***Inference***

$[e_1, e_1, e_1, e_2, e_3, e_3] \rightarrow [e_1, e_2, e_3]$

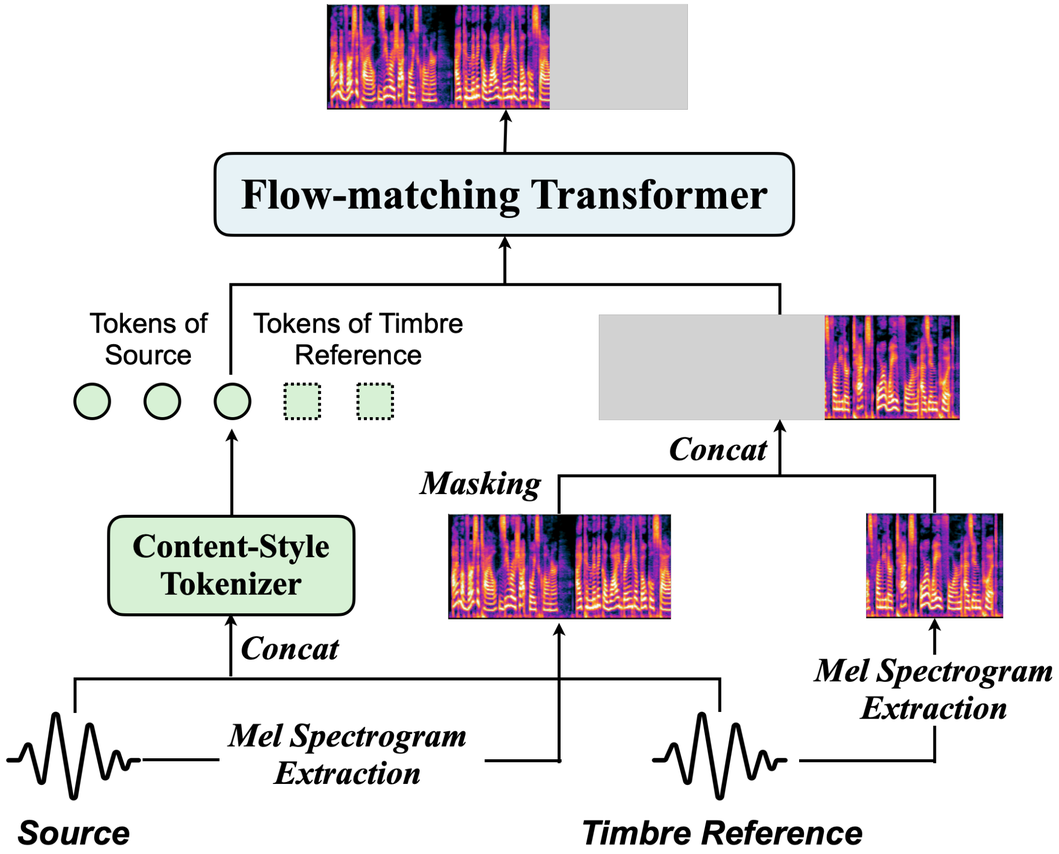
Duration Reduction

★ ***Remove the style information like unit-level duration***

Acoustic Modeling *(Content to Acoustic)*



Training

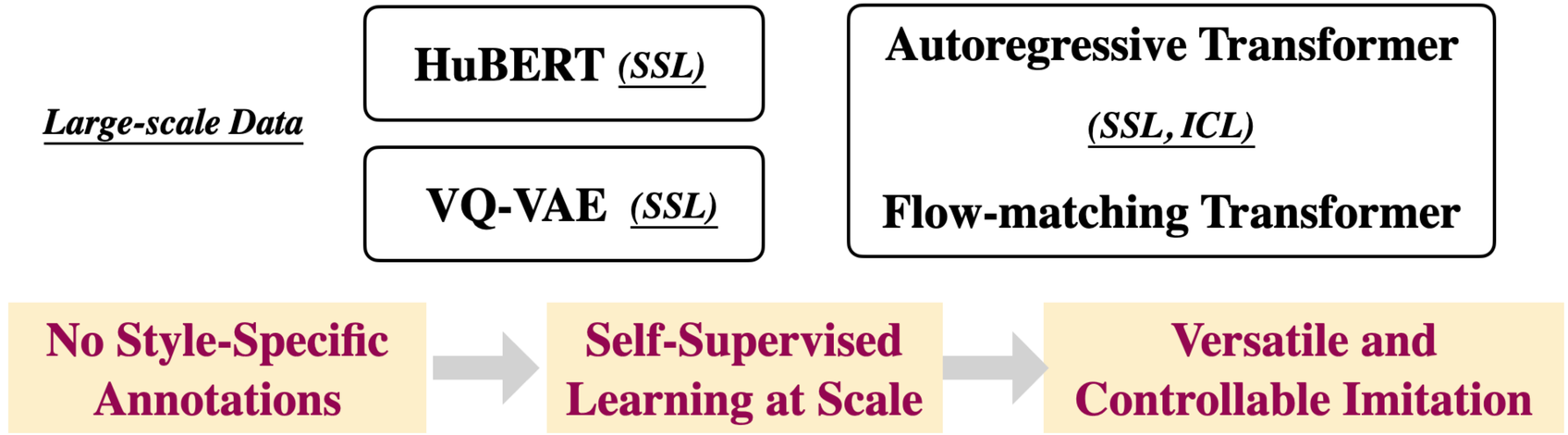


Inference

What can Vevo do?

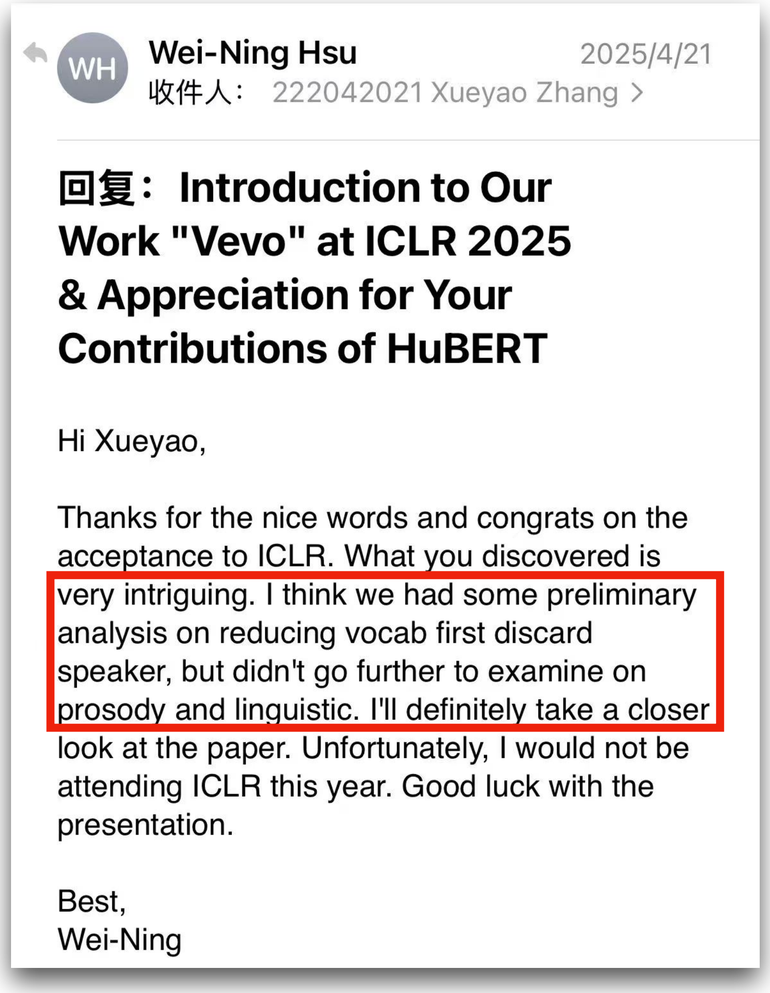
Input	Style Reference	Timbre Reference	Results	
I don't really care what you call me. I've been a silent spectator, watching species evolve, empires rise and fall. But always remember, I am mighty and enduring. Respect me and I'll nurture you; ignore me and you shall face the consequences.	Arabic-accented, Male	Female	Arabic-accented, Female	**Style and Timbre Controllable TTS**
American-accented, Female	N/A	Arabic-accented, Male	American-accented, Male	**Style and Timbre Controllable VC**
	Arabic-accented, Male		Arabic-accented, Male	
		Input	Arabic-accented, Female	**Zero-Shot Accent Conversion**

Conclusions



1. Style and timbre controllable zero-shot TTS and VC.
2. Zero-shot style conversion (including accent and emotion).
3. Fully self-supervised learning, easy to scale up.

Recognition in the Field



Wei-Ning Hsu 2025/4/21
收件人： 222042021 Xueyao Zhang >

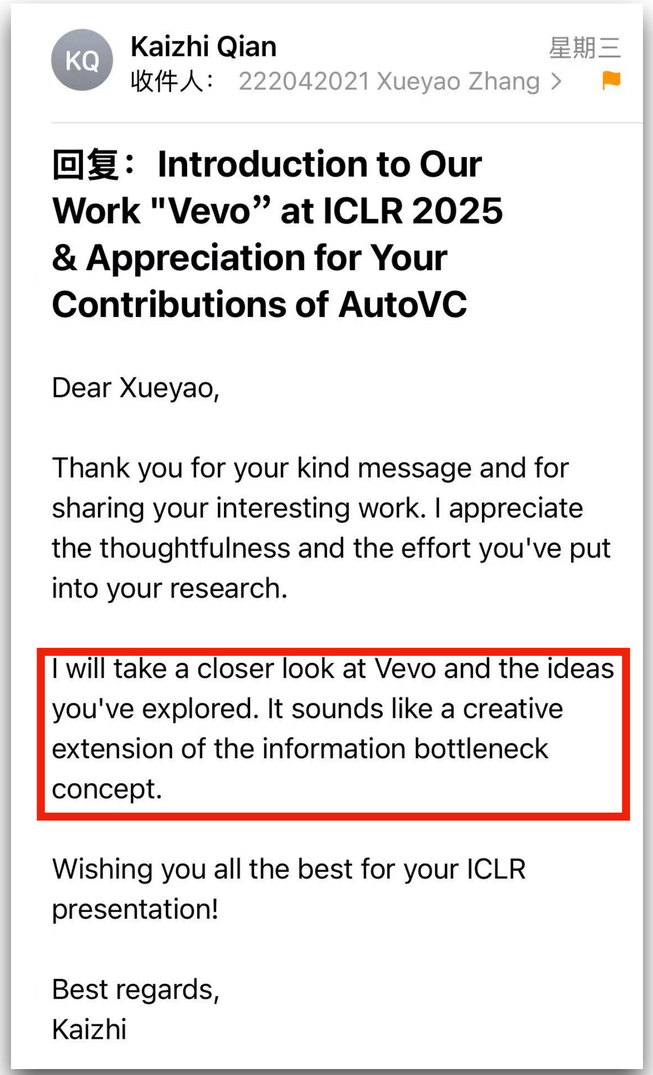
回复： Introduction to Our Work "Vevo" at ICLR 2025 & Appreciation for Your Contributions of HuBERT

Hi Xueyao,

Thanks for the nice words and congrats on the acceptance to ICLR. What you discovered is very intriguing. I think we had some preliminary analysis on reducing vocab first discard speaker, but didn't go further to examine on prosody and linguistic. I'll definitely take a closer look at the paper. Unfortunately, I would not be attending ICLR this year. Good luck with the presentation.

Best,
Wei-Ning

— **Wei-Ning Hsu @ Meta** (HuBERT, VoiceBox, Textless S2ST, …)



Kaizhi Qian 星期三
收件人： 222042021 Xueyao Zhang >

回复： Introduction to Our Work "Vevo" at ICLR 2025 & Appreciation for Your Contributions of AutoVC

Dear Xueyao,

Thank you for your kind message and for sharing your interesting work. I appreciate the thoughtfulness and the effort you've put into your research.

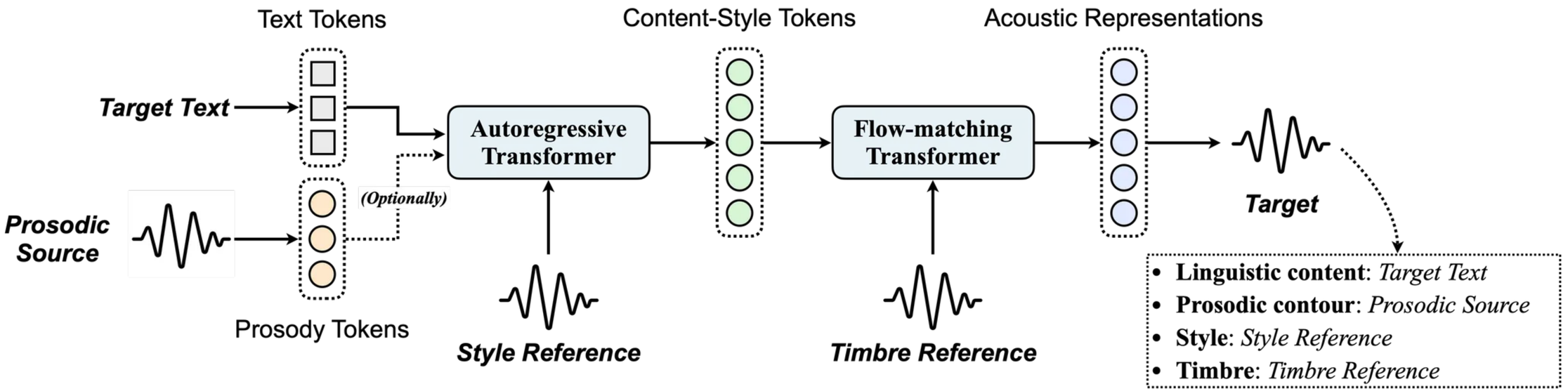
I will take a closer look at Vevo and the ideas you've explored. It sounds like a creative extension of the information bottleneck concept.

Wishing you all the best for your ICLR presentation!

Best regards,
Kaizhi

— **Kaizhi Qian @ MIT-IBM** (AutoVC, SpeechSplit, ContentVec, …)

Vevo2: Unified Generation for Both Speech and Singing Voice



- **Unique Capabilities of Vevo2**
  - Text to Singing
  - Humming to Singing
  - Lyric Editing
  - …