Contact: xueyaozhang@link.cuhk.edu.cn Website: https://versavoice.github.io/

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

Meta

Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, Yuan Huang, Zhizheng Wu, Mingbo Ma

The Chinese University of Hong Kong, Shenzhen Meta AI

Motivation: To design a unified, controllable, and zero-shot voice imitation system

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

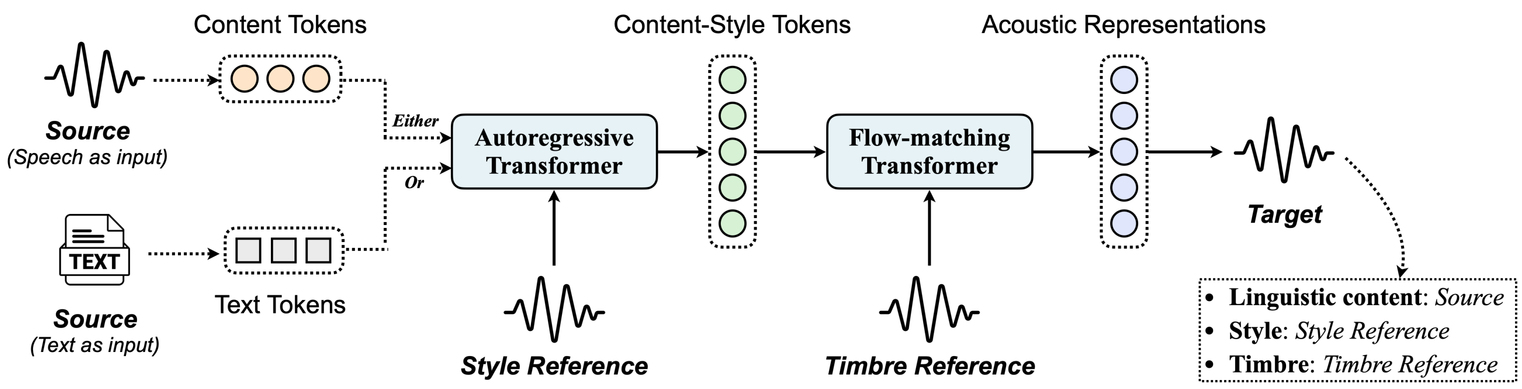
Content: What to speak? Style: How to speak? Timbre: Who speaks?

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?

Vevo: A Versatile Zero-Shot Voice Imitation Framework with Controllable Timbre and Style



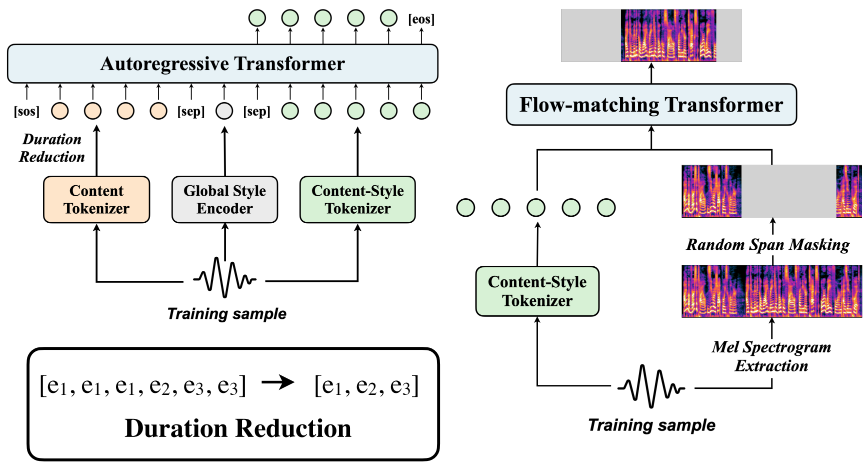
Given a source speech (U_i) or text (T_i), and a reference speech (U_r):

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

KEY POINT: 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 (Starting point of information filtering)	-	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 ↓
VQ-VAE tokens (Content-style Tokens)	4096	6.908 ↑	0.403	0.236 ↓	0.797 ↓	
VQ-VAE tokens	1024	6.967 ↑	0.418	0.249	0.764 ↓	
VQ-VAE tokens (Content Tokens)	32	9.731 ↑	0.426	0.161 ↓	0.706 ↓	
VQ-VAE tokens	16	13.169 ↑	0.441	0.146 ↓	0.672 ↓	
VQ-VAE tokens	8	21.813 ↑	0.392	0.109 ↓	0.675	

Content / Content-Style Tokenizer: VQ-VAE tokenizer (32 / 4096) on HuBERT, Fully Self-Supervised Learning (SSL)



60K hours of audiobook speech data

Large-scale In-Context Learning (ICL)

Experiments and Results

Model	AR?	Training Data (ContRep / Model)	Naturalness	Prosody Similarity	Speaker Similarity	Accent Similarity	Emotion Similarity
HierSpeech++	✗	500K / 2.8K	3.04	3.08	3.15	3.13	2.55
LM-VC	✓	1K / 60K	2.40	2.16	2.56	3.02	2.46
UniAudio	✓	1K / 100K	2.95	2.51	2.39	2.42	2.41
FACodec	✗	60K / 60K	2.36	*3.10*	3.19	3.01	2.30
Vevo-Timbre	✗	60K / 60K	**3.43**	**3.45**	*3.46*	*3.55*	*2.66*
Vevo-Voice	✓	60K / 60K	*3.24*	2.60	**3.70**	**3.90**	**3.20**

Model	Zero-Shot	Supervision: Parallel Corpus	Supervision: Style Labels	Supervision: Text	CMOS: Naturalness	CMOS: Accentedness	CMOS: Emotiveness
VoiceShop	✗	✓	✓	✓	0.00	0.00	-
Vevo-Style	✓	✗	✗	✗	**0.12**	**0.13**	-
Emovox	✗	✗	✓	✓	0.00	0.00	-
Vevo-Style	✓	✗	✗	✗	**1.78**	-	**0.49**

Model	AR?	Training Data	Naturalness	Speaker Similarity	Accent Similarity	Emotion Similarity
CosyVoice	✓	171K	-0.18	4.11	3.99	3.66
MaskGCT	✗	100K	-0.04	4.16	4.38	3.76
VALL-E	✓	45K	-1.24	2.82	2.77	2.63
Voicebox	✗	60K	*-0.35*	*3.87*	*3.49*	*3.61*
VoiceCraft	✓	9K	-0.50	3.47	3.29	3.52
Vevo-TTS	✓	60K	**-0.14**	**4.05**	**4.12**	**4.03**

TAKE HOME MESSAGE

Large-scale Data

HuBERT *(SSL)*

VQ-VAE *(SSL)*

Autoregressive Transformer *(SSL, ICL)* **Flow-matching Transformer**

No Style-Specific Annotations → **Self-Supervised Learning at Scale** → **Versatile and Controllable Imitation**

More Information

Paper and Demo

Code and Checkpoint

We will release code and checkpoint at Amphion: https://github.com/open-mmlab/Amphion