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

A Comprehensive Guide to Amphion's Singing Voice Conversion

Xueyao Zhang

The Chinese University of Hong Kong, Shenzhen

2024/12

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/

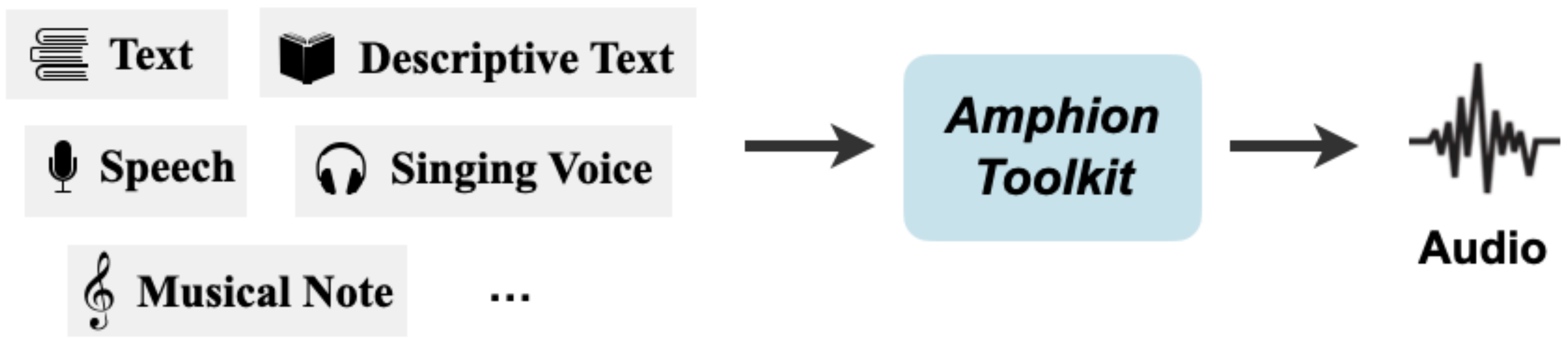
- **Amphion v0.1's co-founder**
  Project: https://github.com/open-mmlab/Amphion **(7.8k stars)**

- **Research interest: "AI + Music"**, especially on:
  - Singing Voice Processing
  - Music Generation

Amphion Technical Report: https://arxiv.org/abs/2312.09911
Amphion GitHub: https://github.com/open-mmlab/Amphion
Amphion Demos/Models/Datasets: https://huggingface.co/amphion

About Amphion

- 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

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

Xueyao Zhang[*,1], Liumeng Xue[*,1], Yicheng Gu[*,1], Yuancheng Wang[*,1], Haorui He[3], Chaoren Wang[1], Xi Chen[1], Zihao Fang[1], Haopeng Chen[1], Junan Zhang[2], Tze Ying Tang[1], Lexiao Zou[3], Mingxuan Wang[1], Jun Han[1], Kai Chen[2], Haizhou Li[1], Zhizheng Wu[†,1,2,3]

[1]School of Data Science, The Chinese University of Hong Kong, Shenzhen
[2]Shanghai AI Lab
[3]Shenzhen Research Institute of Big Data

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

Roadmap

- **Singing Voice Conversion**
  - Definition, Classic Works, and Modern Pipeline
- **Singing Voice Conversion in Amphion**
  - Supported Model Architectures
  - Our research: *Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion*
- **Amphion's Philosophy**
  - Unique strengths, Supported Features, and Visualization
- **Singing Voice Conversion: Next Steps**

What is Singing Voice Conversion (SVC)?

Professional Singer1 → **Professional Singer2**

Inter-singer Conversion

Speaker → **Singer**

Cross-domain Conversion

Amateur Singer → **Professional Singer**

Intra-singer Conversion

Parallel Singing Voice Conversion

Professional Singer1

→

Professional Singer2

X

(Song1, Singer1)
(Song2, Singer1)
…
(SongN, Singer1)

→ $f(X)$ →

(Song1, Singer2)
(Song2, Singer2)
…
(SongN, Singer2)

Y

Limited parallel data

Limited flexibility

Non-Parallel Singing Voice Conversion

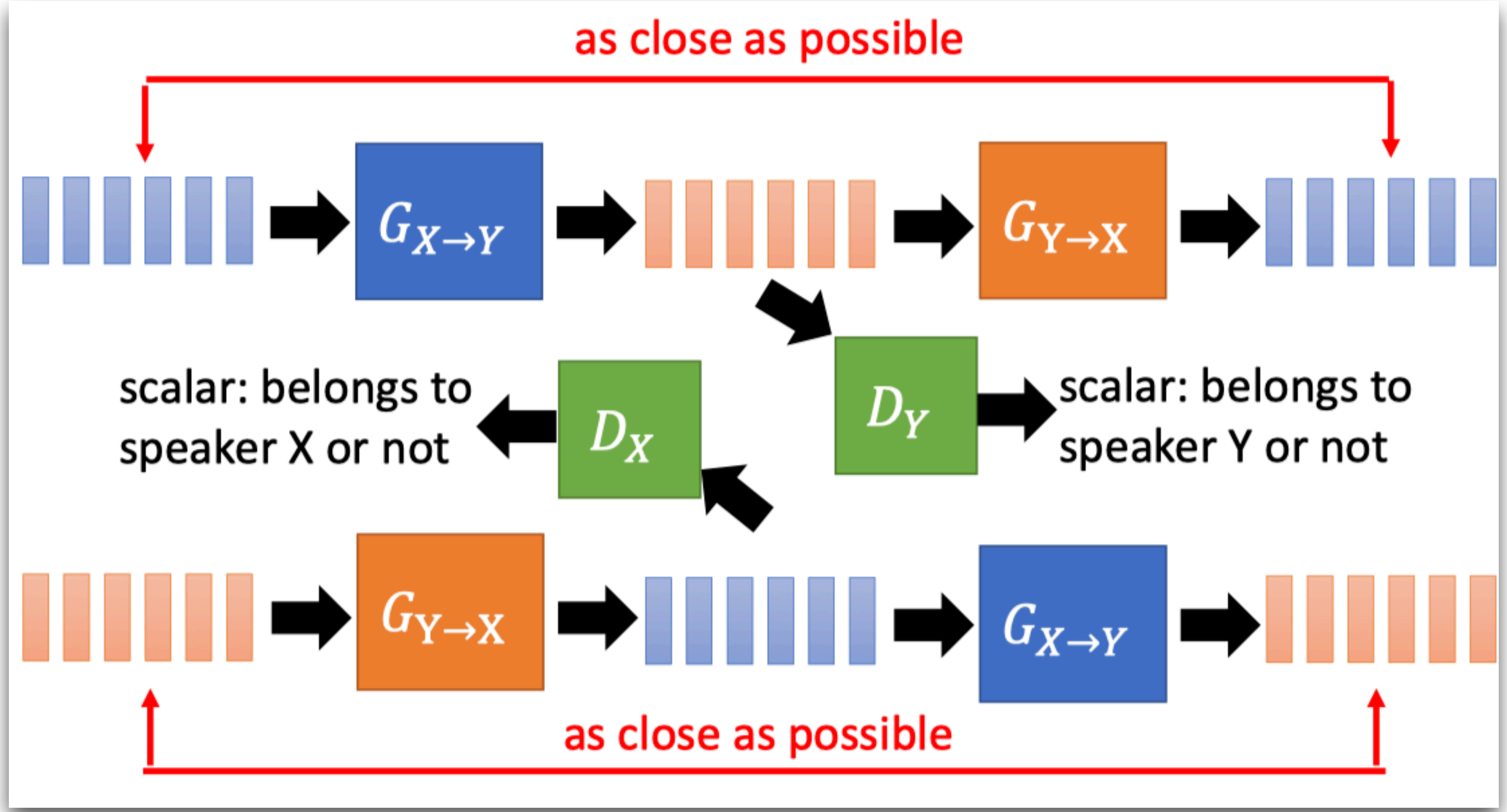
Professional Singer1

Professional Singer2

X Singer1's Songs Singer2's Songs

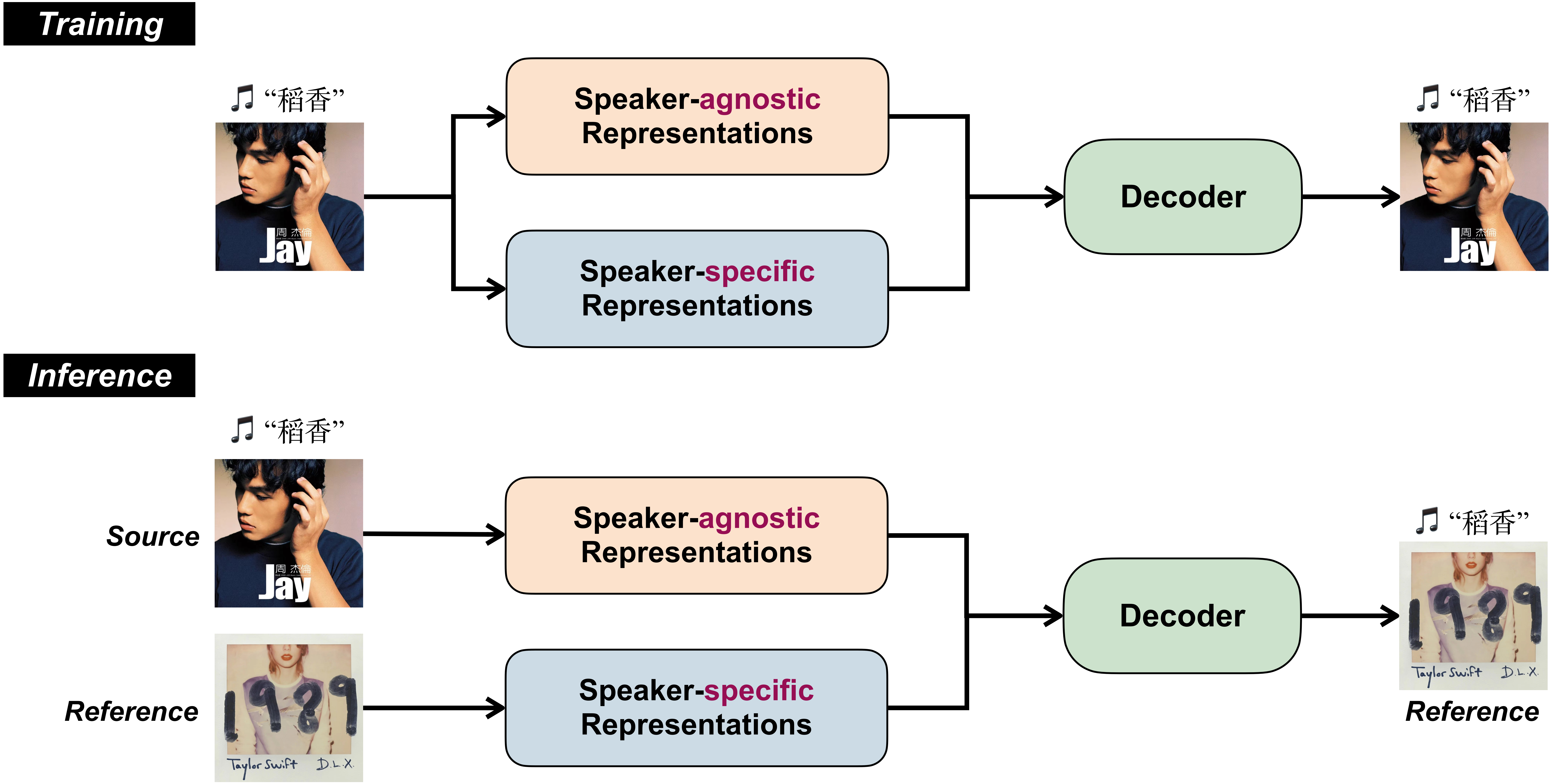
How to decouple the singer identity?

Non-Parallel SVC: GAN School

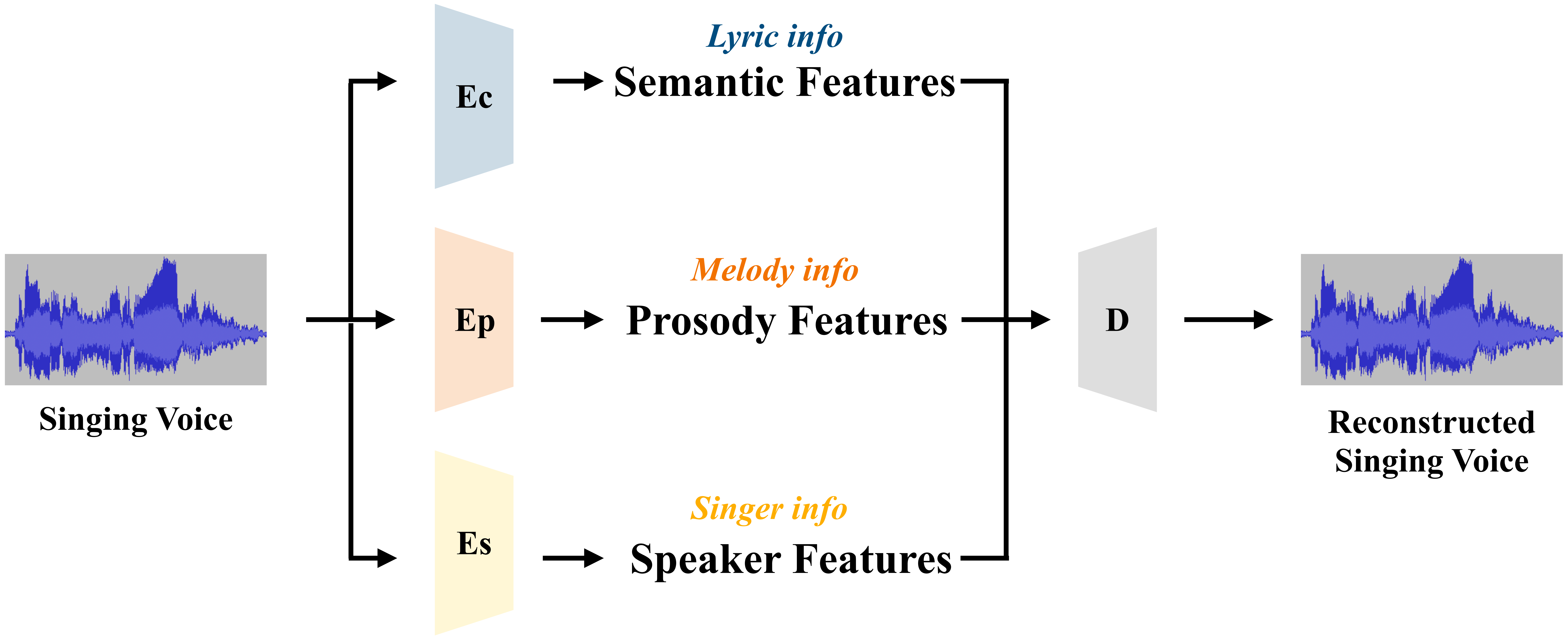


Credit: Voice Conversion, Hung-yi Lee.

Non-Parallel SVC: Auto-Encoder School



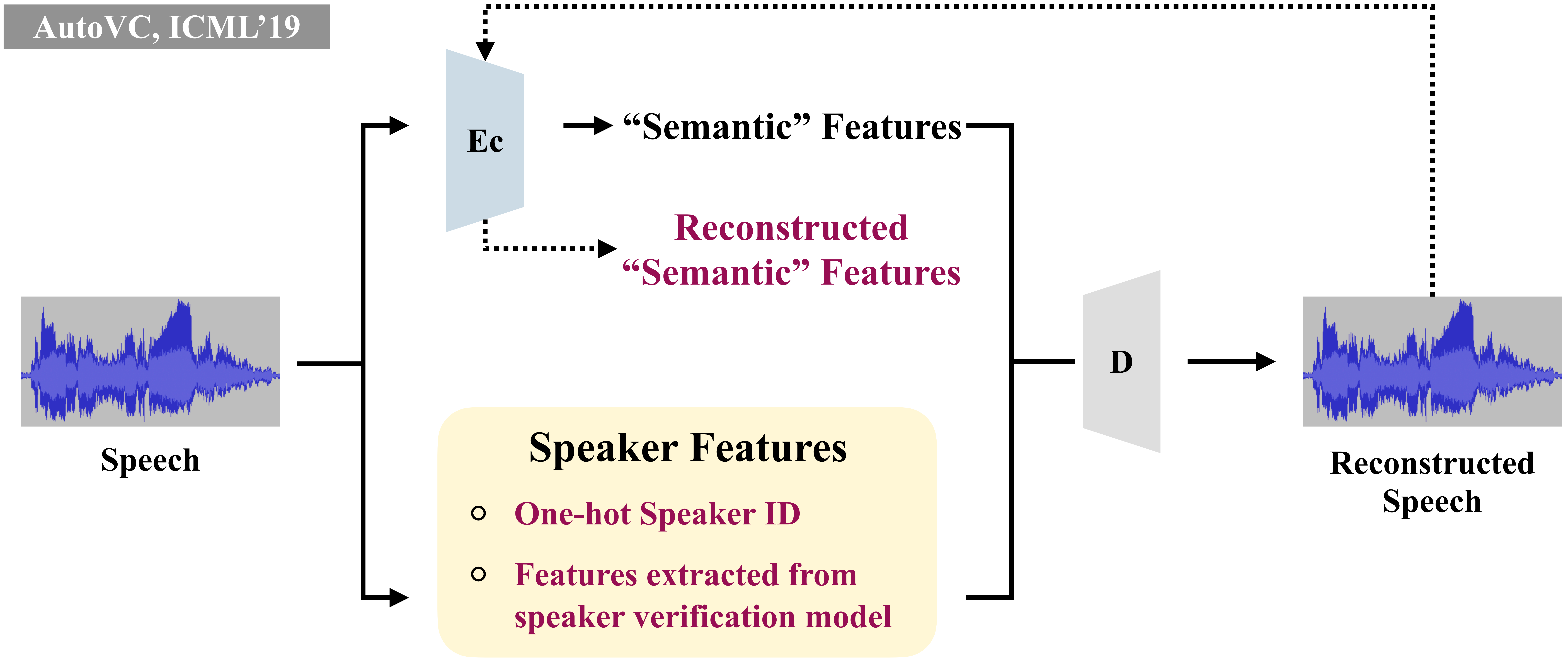
Non-Parallel SVC: Auto-Encoder School



- How to ensure the disentanglement of different features?
- How to ensure there is enough information of each features?

Auto-Encoder VC: The Early Researches

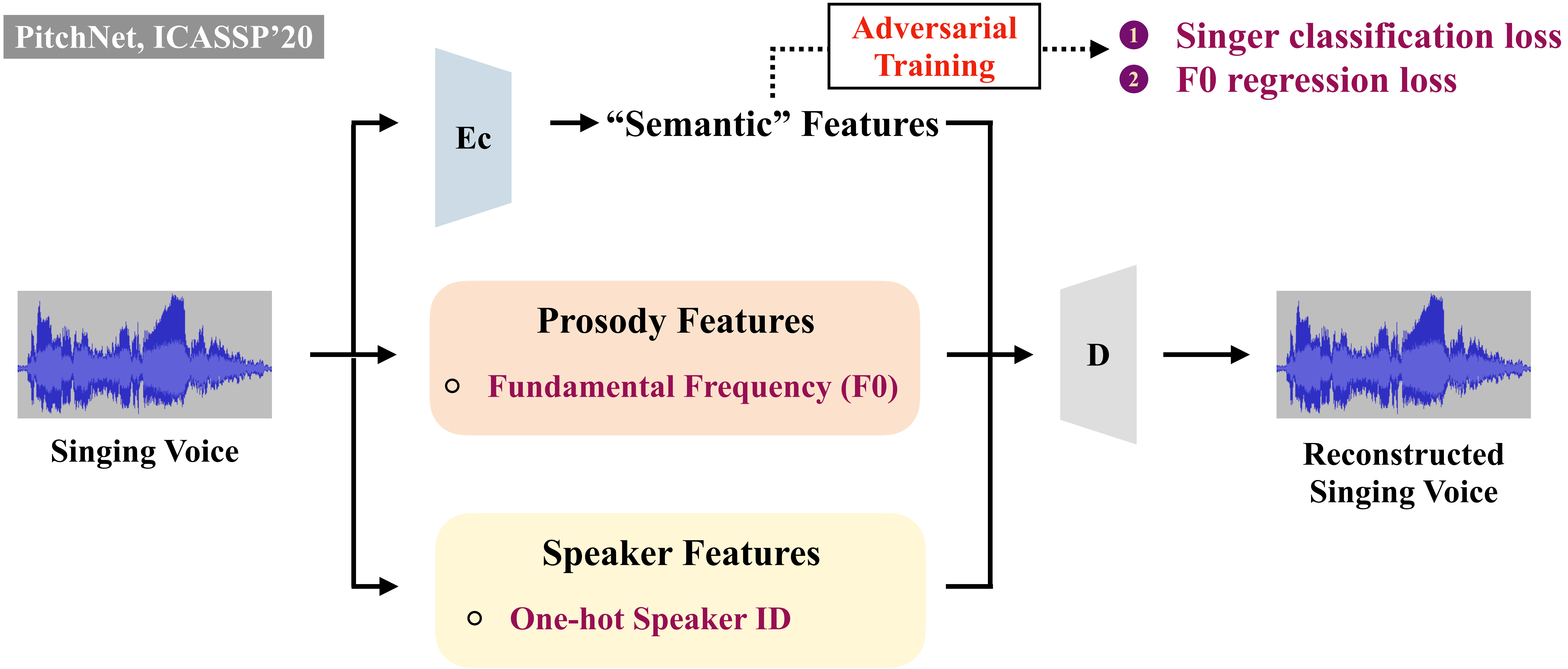
AutoVC, ICML’19



AutoVC: “To carefully design the dimension of the *semantic* features"

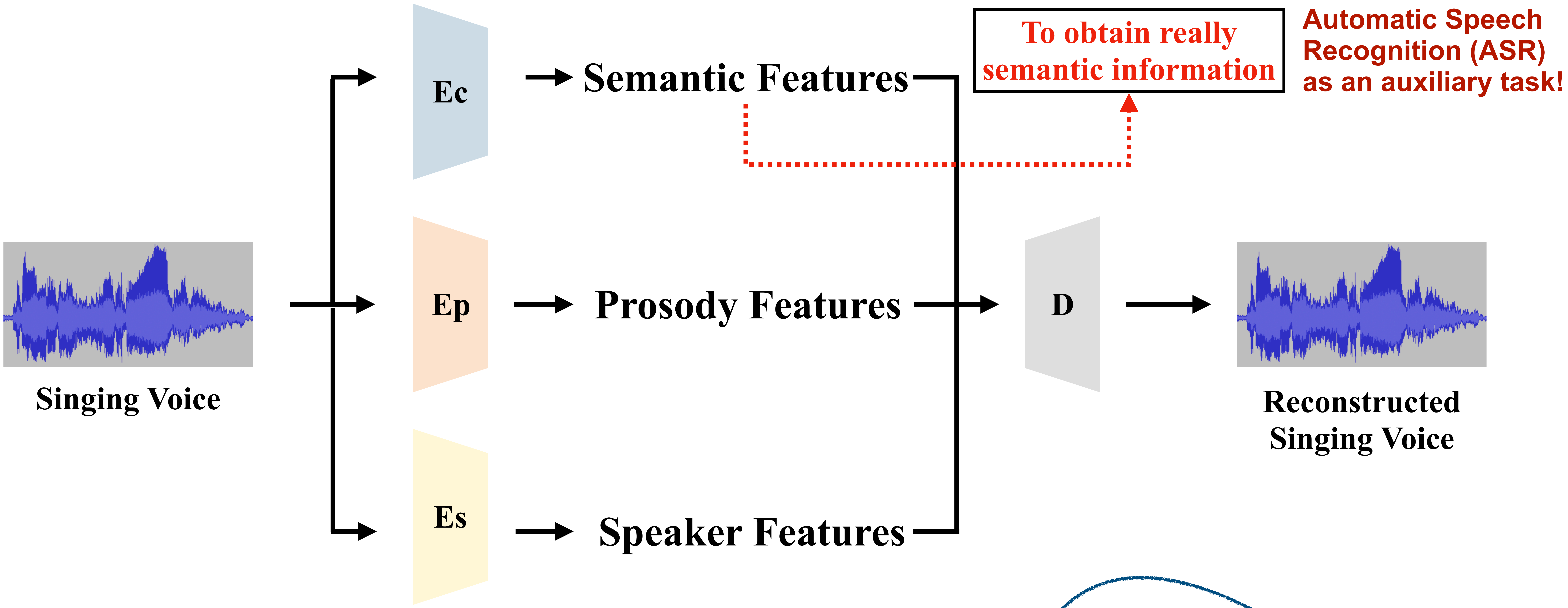
Auto-Encoder SVC: The Early Researches

PitchNet, ICASSP'20



PitchNet: "Adopt adversarial training to disentangle better"

(Review) Non-Parallel SVC: Auto-Encoder School

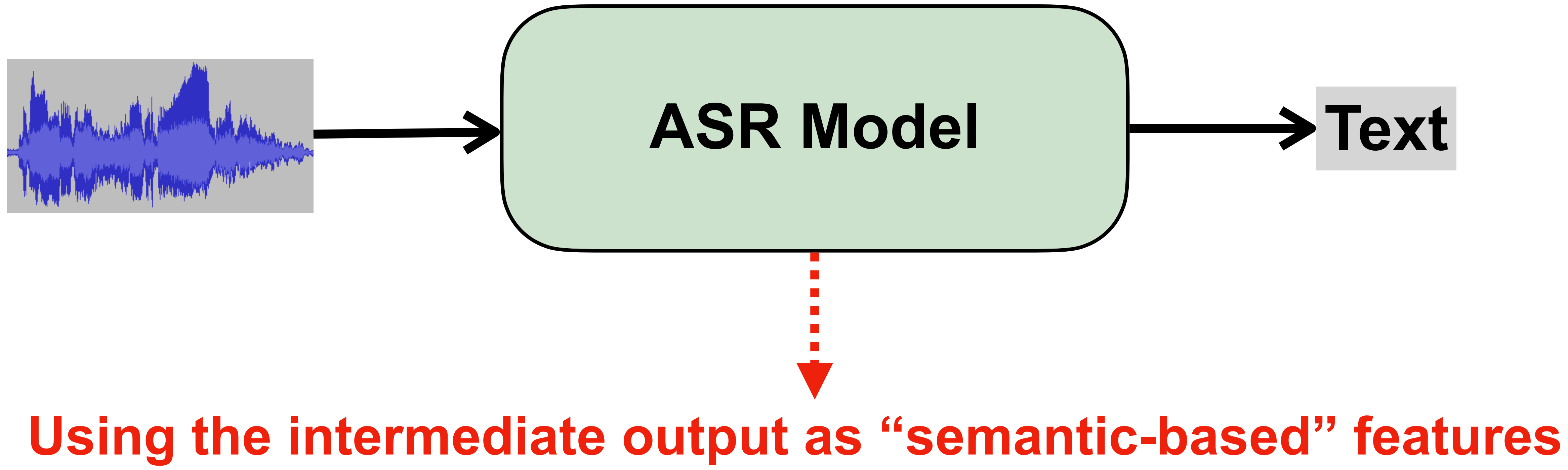


- How to ensure the disentanglement of different features?
- How to ensure there is enough information of each features?

🎯 Solved to some extent

🤔 How to address?

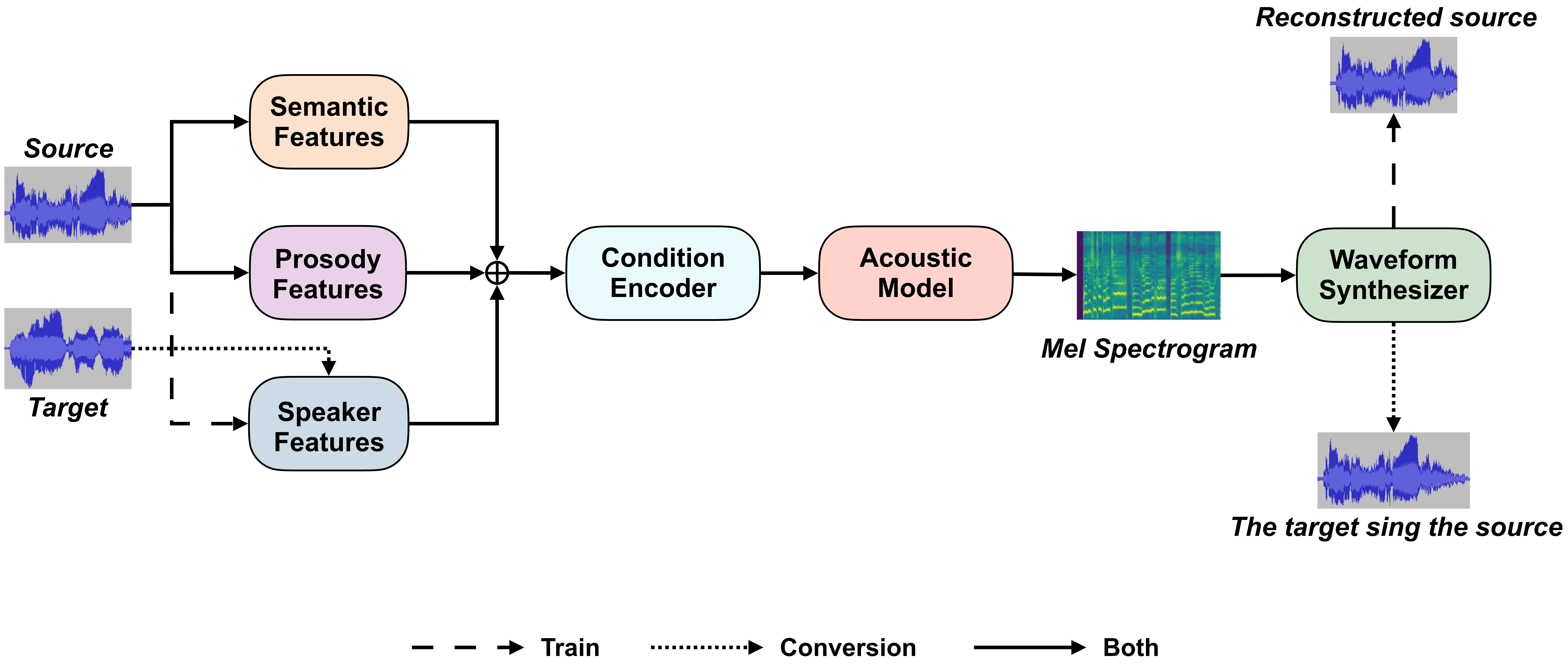
Non-Parallel VC/SVC — a.k.a Recognition & Synthesis VC/SVC



🤔 Why do we use the *continuous semantic features* instead of the *symbolic text*?

1. There are errors for the recognized symbolic text.
2. It takes more time to obtain the symbolic text than just extracting dense features.
3. There are more acoustic information (such as pronunciation) in the dense features, which is better for improving the intelligibility of the synthesized voice.

Modern Singing Voice Conversion Pipeline



Roadmap

- **Singing Voice Conversion**
  - Definition, Classic Works, and Modern Pipeline
- **Singing Voice Conversion in Amphion**
  - Supported Model Architectures
  - Our research: *Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion*
- **Amphion's Philosophy**
  - Unique strengths, Supported Features, and Visualization
- **Singing Voice Conversion: Next Steps**

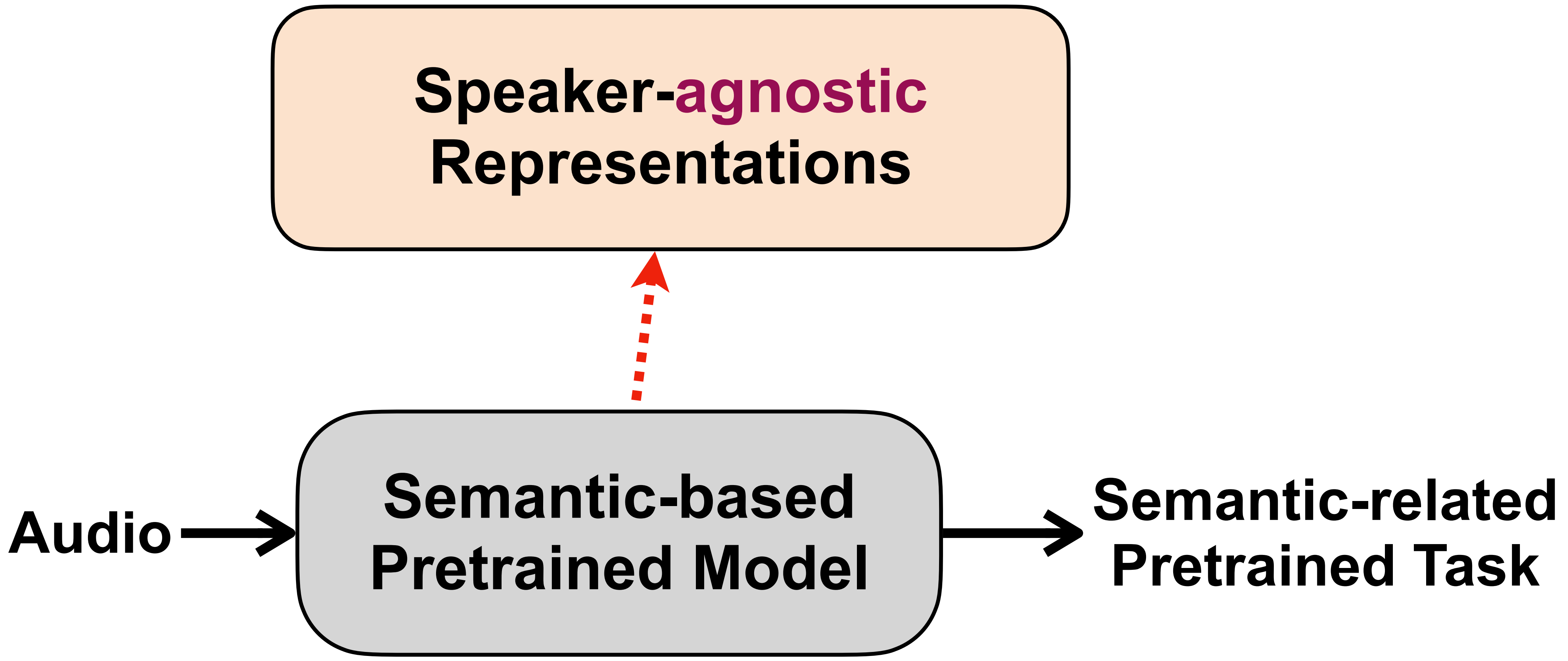
Amphion SVC: Supported Model Architectures

- **Semantic Features Extractor**
  - WeNet, Whisper, ContentVec, HuBERT
  - Joint Usage of Diverse Semantic Features Extractors
- **Prosody Features**
  - F0 and energy
- **Speaker Features**
  - One-hot Speaker ID
  - Features of Pretrained SV model
- **Acoustic Model**
  - Diffusion-based
  - Transformer-based
  - VAE- and Flow-based
- **Waveform Synthesizer**
  - GAN-based
  - Diffusion-based

The Importance of Semantic-based Pretrained Models

Can be:

- One-hot speaker ID
- Embeddings from speaker verification models
- Mel spectrogram



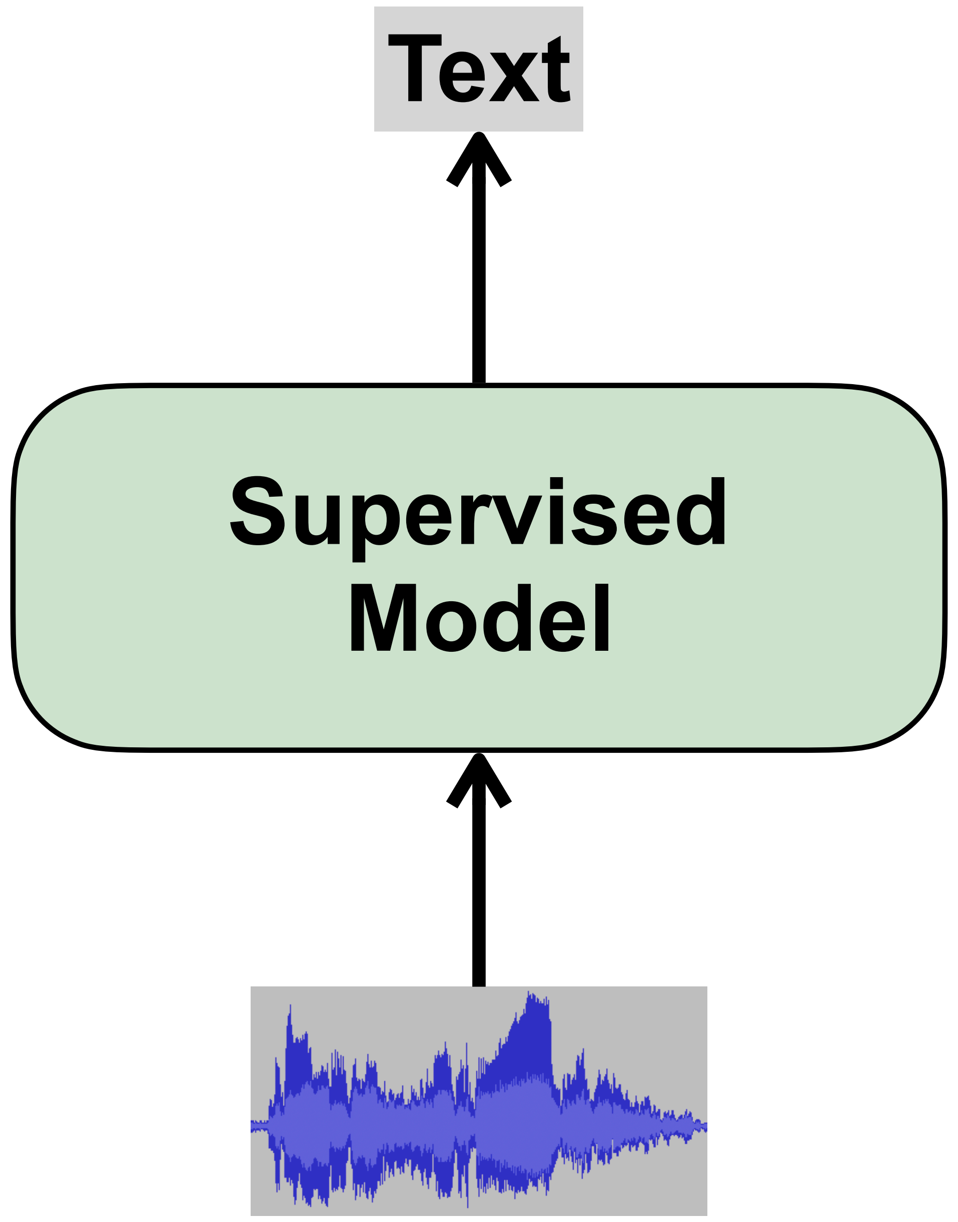
Semantic-related pretrained tasks:

- Automatic Speech Recognition (ASR)
- Semantic-guided self-supervised learning (eg: HuBERT)

Requirements for Speaker-agnostic Representations

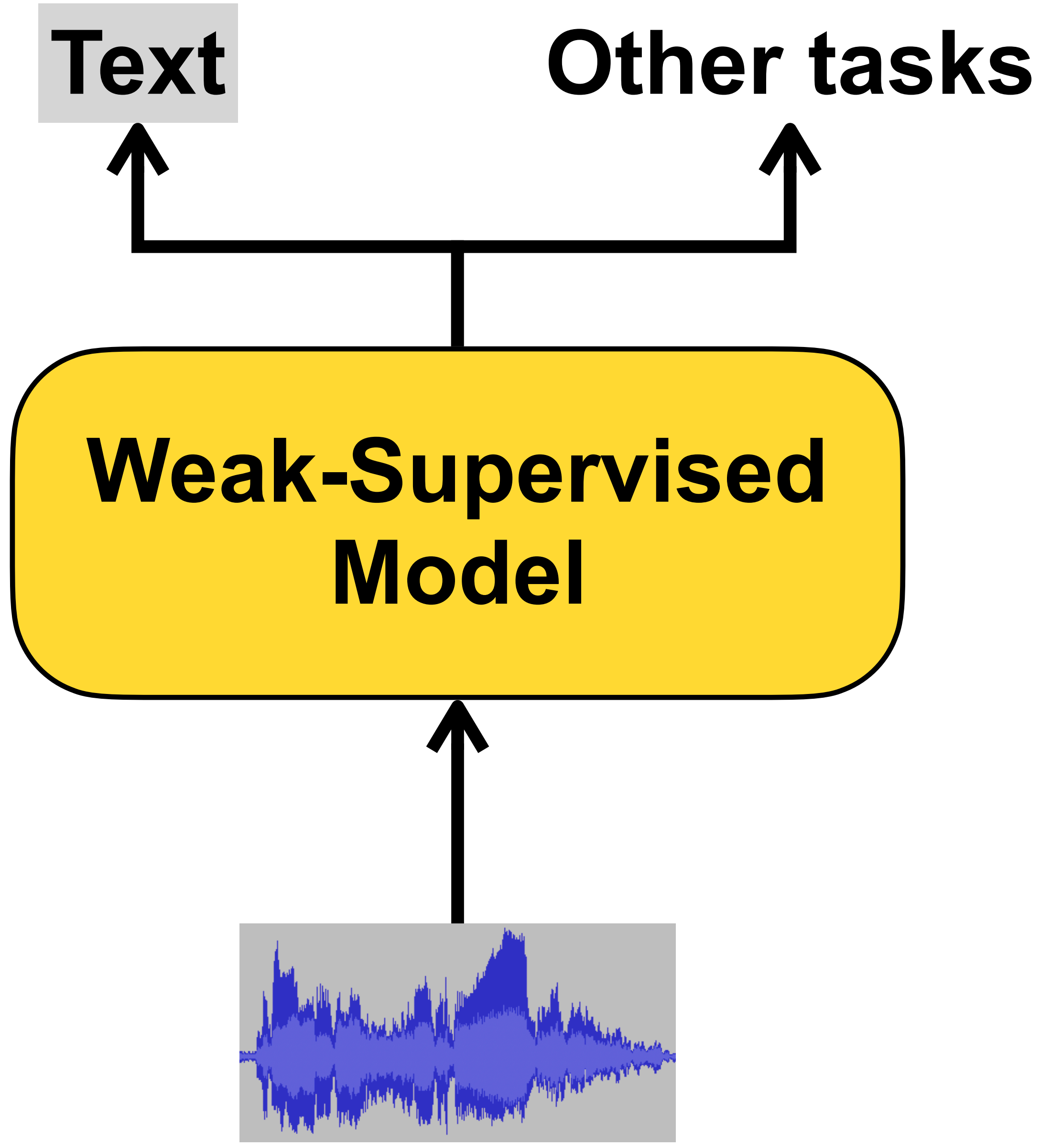
Requirements of SVC	**Capability of the Semantic-based Features**
To model melody	**Whether could or not** remains unknown
To model lyrics	Could. But **exactly how much** remains unknown
To model auxiliary acoustic information	Could. But **whether the information is speaker-agnostic or not** remains unknown
To be robust for in-the-wild acoustic environment	**Whether is robust or not** remains unknown

Analysis: Three Schools of Semantic-based Pretraining



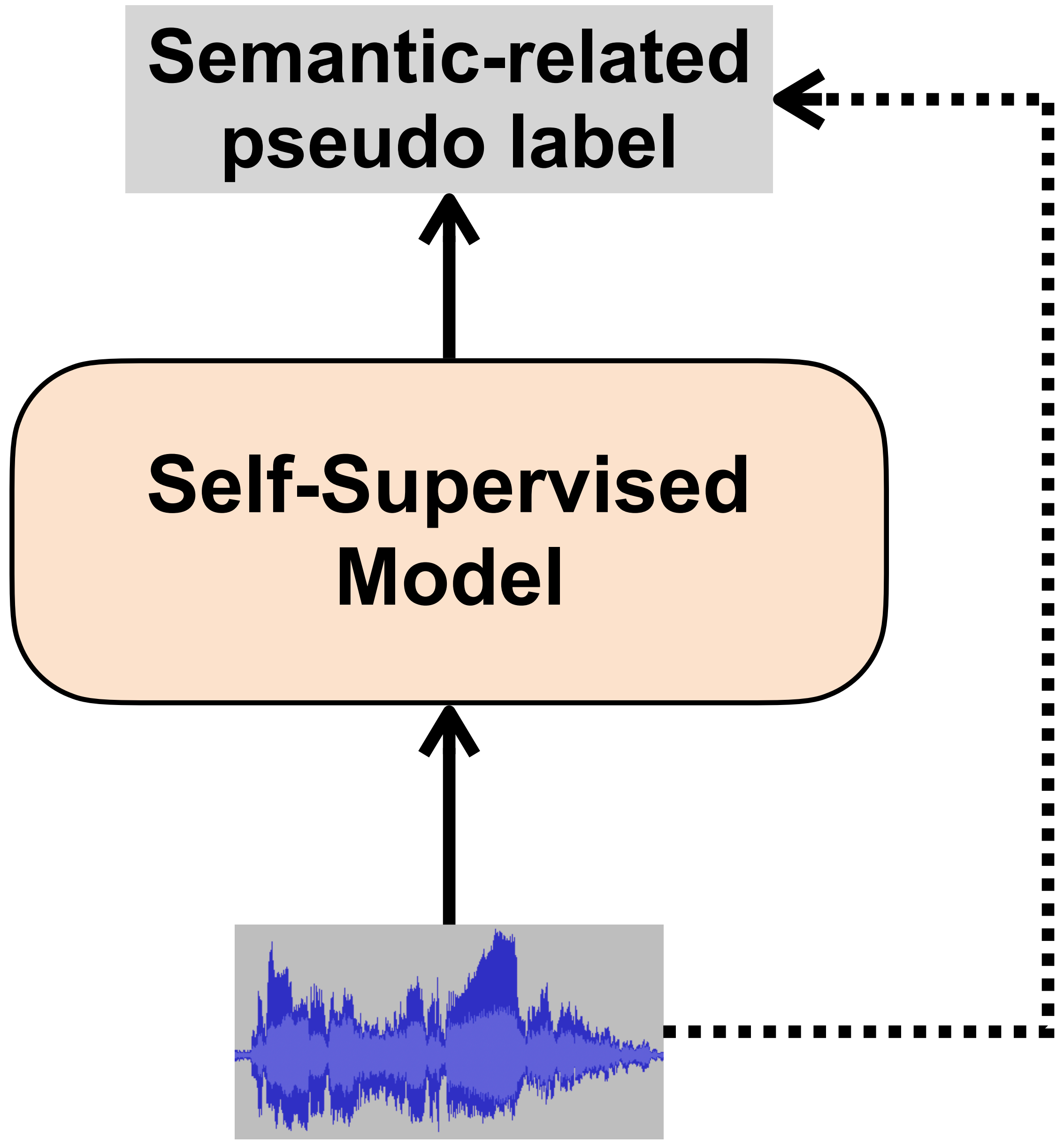
Supervised Model
(eg: WeNet)

10k hours of speech,
English or Chinese



Weak-Supervised Model
(eg: Whisper)

680k hours,
multilingual and multi-task

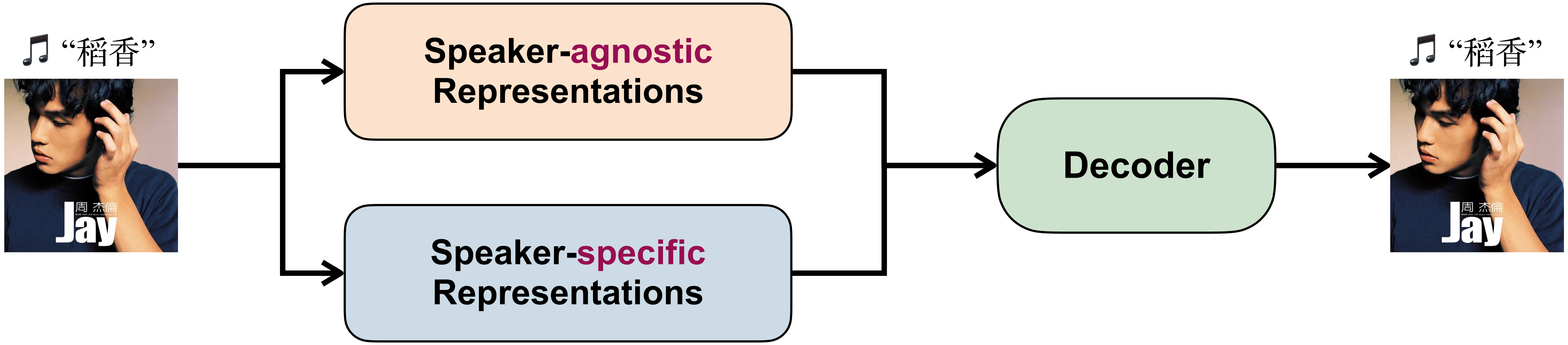


Self-Supervised Model
(eg: HuBERT / ContentVec)

1k hours of speech,
English

Experiments: Using Only Semantic-based Features for SVC

Training



- **Speaker-agnostic representations**
  - WeNet / Whisper / ContentVec
  - Output of the top layer of encoder
- **Speaker-specific representations**
  - One-hot speaker ID
  - Mel-spectrogram
- **Decoder**
  - Diffusion (WaveNet), DDPM (1000 steps)
- **Training Data (Decoder)**
  - **Studio Recording**: 83.1 hours of speech, 128.3 hours of singing voice
  - **In the wild**: 6.4 hours of source separated singing voice

Results: Using Only Semantic-based Features for SVC

Semantic-based Features	MCD (↓)	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)
Ground Truth	0.000	1.000	0.0	12.9%	1.000
WeNet	10.324	0.203	423.4	38.2%	0.912
Whisper	*8.229*	0.524	297.3	18.9%	*0.914*
ContentVec	8.972	0.491	361.0	22.1%	**0.918**

① ② ③

On studio recording eval-set

① **To model melody:**

Whisper > ContentVec > WeNet, but all of them are not enough

② **To model lyrics:**

Whisper > ContentVec > WeNet

③ **To be speaker-agnostic:**

When using speaker ID, all of the three are good.

- ✻ **Compared with the classic supervised model, weak-supervised and self-supervised models is more robust for singing voice**
- ✻ **Large-scale pretraining corpus is necessary**

Results: Complementary roles of Diverse Semantic-based Features

Semantic-based Features	MCD (↓)	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)
Ground Truth	0.000	1.000	0.0	12.9%	1.000
WeNet	10.324	0.203	423.4	38.2%	0.912
Whisper	*8.229*	0.524	297.3	18.9%	*0.914*
ContentVec	8.972	0.491	361.0	22.1%	**0.918**
WeNet + Whisper	8.345	0.540	284.2	*16.8%*	0.911
WeNet + ContentVec	8.870	0.525	329.5	19.9%	0.912
Whisper + ContentVec	**8.201**	*0.548*	*279.6*	16.9%	0.912
WeNet + Whisper + ContentVec	8.249	**0.572**	**278.5**	**16.1%**	0.913

①

②

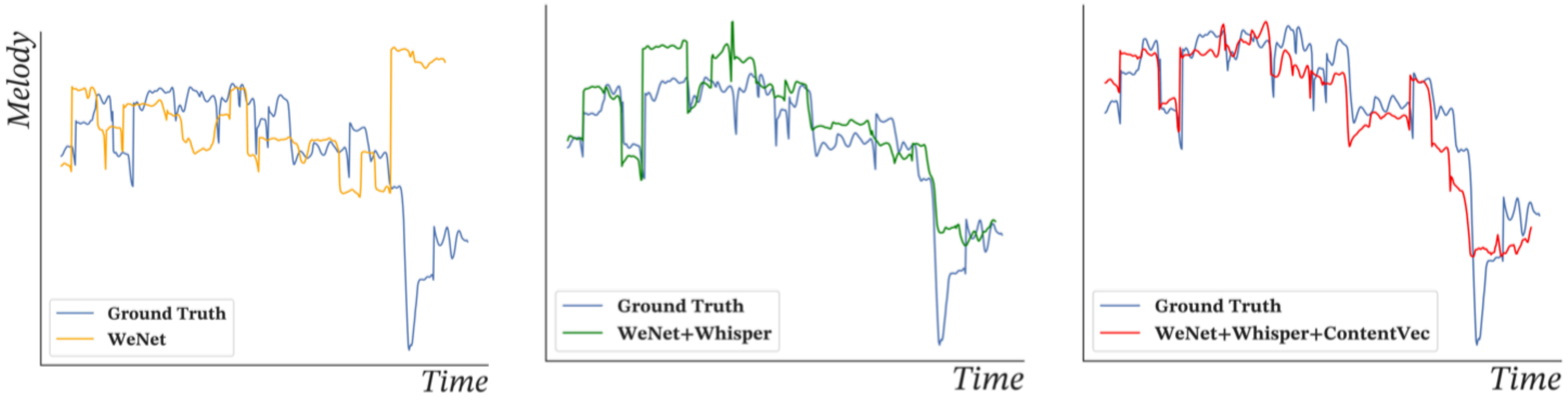
On studio recording eval-set

After Introducing F0

Using diverse semantic-based features:

① Most results are promoted stage by stage

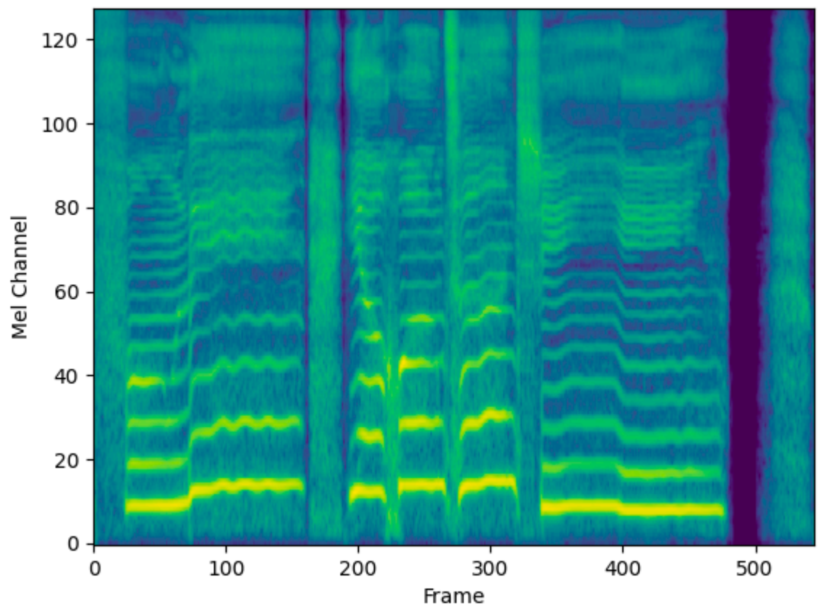
② Introducing explicit melody modeling for SVC remains necessary



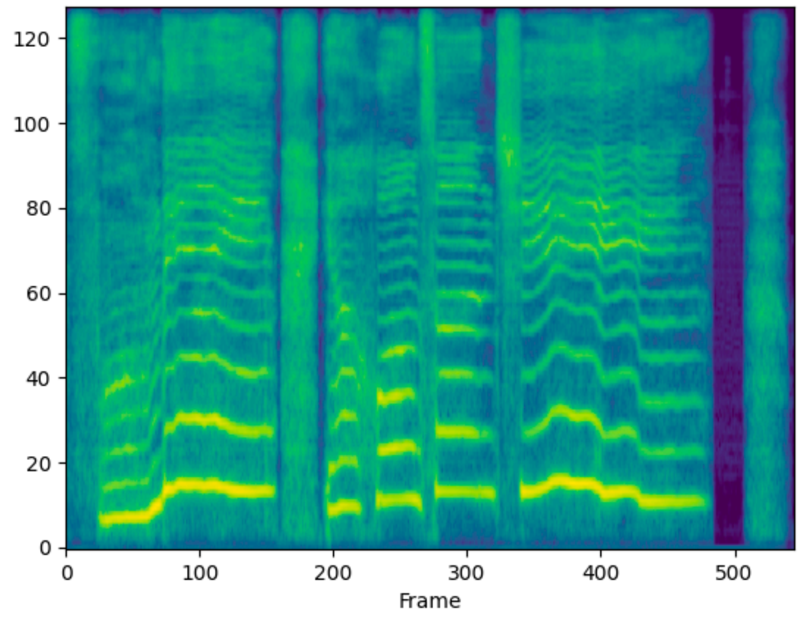
Reference **Source** **WeNet** **WeNet + Whisper** **WeNet + Whisper + ContentVec**

Results: Complementary roles of Diverse Semantic-based Features

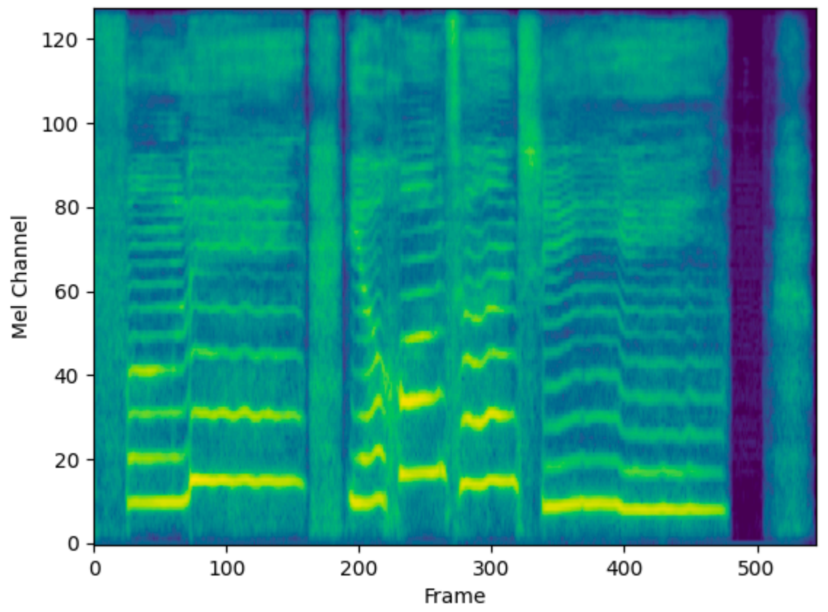
Ground Truth	WeNet	WeNet + Whisper	WeNet + Whisper + ContentVec



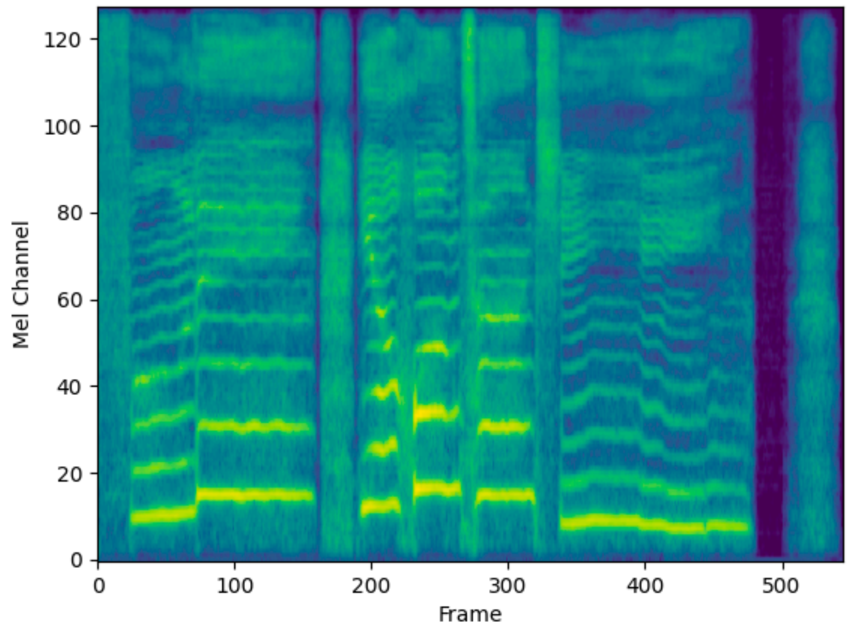
Ground Truth



MCD: 12.56



MCD: 8.00



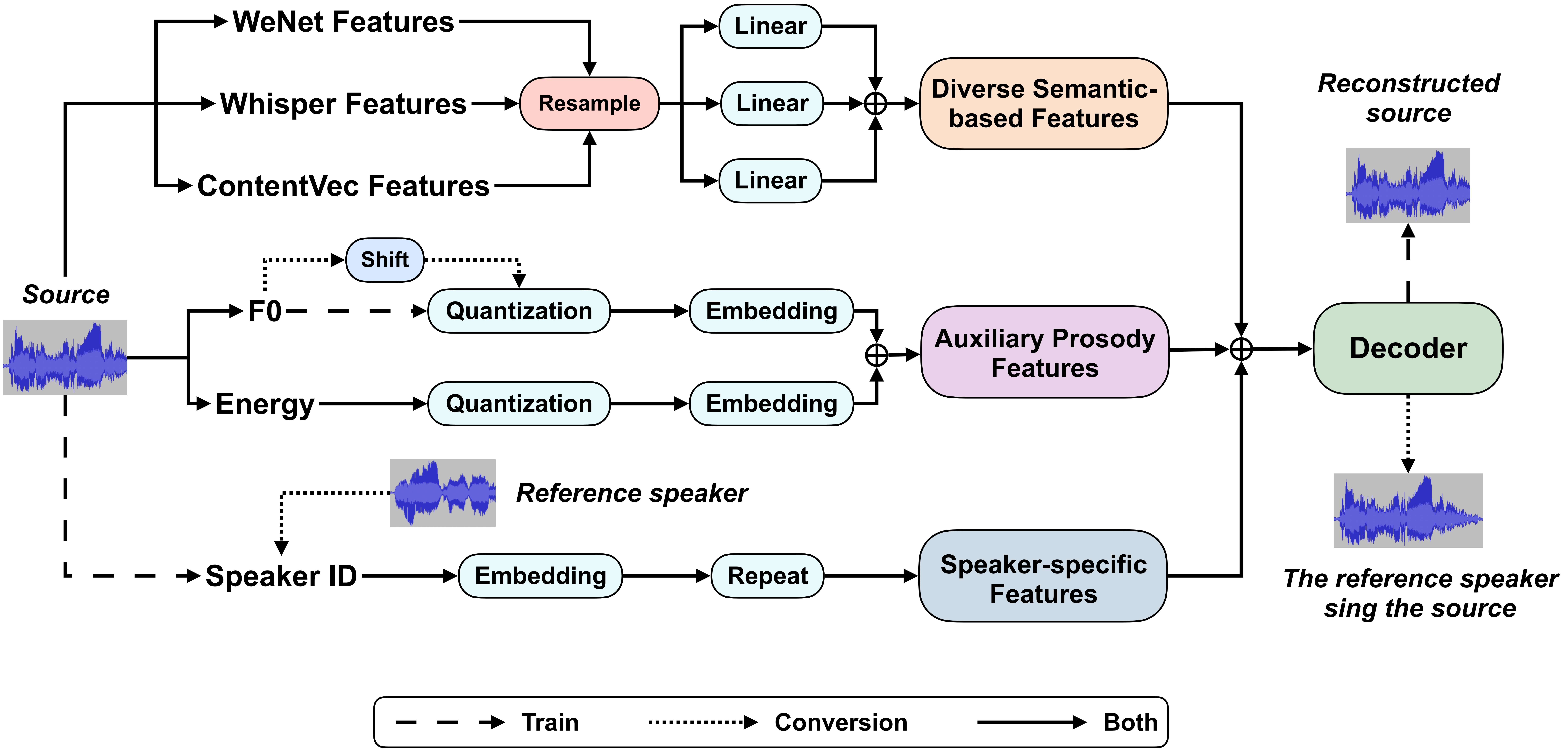
MCD: 5.02

Spectrogram Reconstruction

Ground Truth	WeNet	WeNet + Whisper	WeNet + Whisper + ContentVec
这只是刚刚入门接下来你还会会弹琴会写歌会双截棍	这只是**当童话吗地下安利啊狗货看清单**写**报上节归**	这只是刚刚入门接下来你还会弹琴会写歌会双**戒柜**	**就**只是刚刚入门接下来你还会弹琴会写歌会双截棍
zhè zhǐ shì gāng gāng rù mén jiē xià lái nǐ hái huì huì tán qín huì xiě gē huì shuāng jié gùn	zhè zhǐ shì **dāng tóng huà ma dì** xià **ān lì a gǒu huò kàn qīng dān** xiě **bào shàng jié guī**	zhè zhǐ shì gāng gāng rù mén jiē xià lái nǐ hái huì tán qín huì xiě gē huì shuāng **jiè guì**	**jiù** zhǐ shì gāng gāng rù mén jiē xià lái nǐ hái huì tán qín huì xiě gē huì shuāng jié gùn
Ground Truth	CER: 81.8%	CER: 13.0%	CER: 8.7%

Intelligibility

SVC Framework based on Diverse Semantic-based Features Fusion



Results: Recording studio data v.s. In-the-wild data

Subjective Evaluation

Semantic-based Features	Recording Studio Setting		In-the-Wild Setting	
	Naturalness (↑)	Similarity (↑)	Naturalness (↑)	Similarity (↑)
WeNet	2.72 ±0.22	2.64 ±0.21	2.85 ±0.21	2.34 ±0.20
WeNet + Whisper	4.02 ±0.18	3.13 ±0.17	3.70 ±0.18	2.86 ±0.23
WeNet + Whisper + ContentVec	4.14 ±0.19	3.25 ±0.18	3.71 ±0.18	2.82 ±0.23

The full scores of Naturalness and Similarity are 5 and 4

① **Robustness:**

- Compared with the recording studio setting, all the models get worse in the more challenging for in-the-wild evaluation set.
- Leveraging diverse semantic-based features are effective both on the two settings.

Results: The effect of introducing F0 and Energy

Recording Studio Setting, Using only semantic-based features

Semantic-based Features	MCD (↓)	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)
Ground Truth	0.000	1.000	0.0	12.9%	1.000
WeNet	10.324	0.203	423.4	38.2%	0.912
Whisper	*8.229*	0.524	297.3	18.9%	*0.914*
ContentVec	8.972	0.491	361.0	22.1%	**0.918**

Semantic-based Features	Recording Studio Setting				In-the-Wild Setting			
	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)
WeNet	0.936	55.5	15.8%	0.875	0.901	87.8	60.8%	0.855
WeNet + Whisper	0.943	49.5	15.2%	0.884	0.921	73.6	21.1%	0.865
WeNet + Whisper + ContentVec	0.940	55.2	15.7%	0.884	0.919	79.9	23.3%	0.867

Using both semantic-based and prosody (F0 and Energy) features

① **To model melody**: Introducing F0 and Energy improves a lot

② **To model lyrics**: Introducing F0 and Energy also helps for CER

③ **To be speaker-agnostic**: However, introducing F0 and Energy harms the speaker similarity.

Results: For more generative models

Base Model	Semantic-based Features	Recording Studio Setting				In-the-Wild Setting			
		F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)	F0CORR (↑)	F0RMSE (↓)	CER (↓)	SIM (↑)
TransformerSVC	WeNet	0.849	149.3	15.6%	0.878	0.871	210.0	40.0%	0.865
	WeNet + Whisper	0.924	77.2	14.9%	0.881	0.848	183.8	② 18.7%	0.867
	WeNet + Whisper + ContentVec	0.931	75.5	16.2%	0.883	0.857	186.7	23.3%	0.868
VitsSVC	WeNet	0.937	175.3	19.1%	0.890	0.919	91.3	57.7%	0.869
	WeNet + Whisper	0.945	144.4	17.8%	0.890	0.920	86.9	35.2%	0.869
	WeNet + Whisper + ContentVec	0.946	112.9	17.7%	0.886	0.921	79.5	32.3%	0.870
DiffWaveNetSVC	WeNet	0.936	55.5	15.8%	0.875	0.901	87.8	60.8%	0.855
	WeNet + Whisper	0.943	49.5	15.2%	0.884	0.921	73.6	② 21.1%	0.865
	WeNet + Whisper + ContentVec	0.940	55.2	15.7%	0.884	0.919	79.9	23.3%	0.867

① **Generalization:** The idea of diverse semantic-based features fusion work for various base models in both settings.

② **Robustness:** for the more challenging in-the-wild setting, Whisper is more robust than ContentVec. This might be contributed by its size and diversity of the training data.

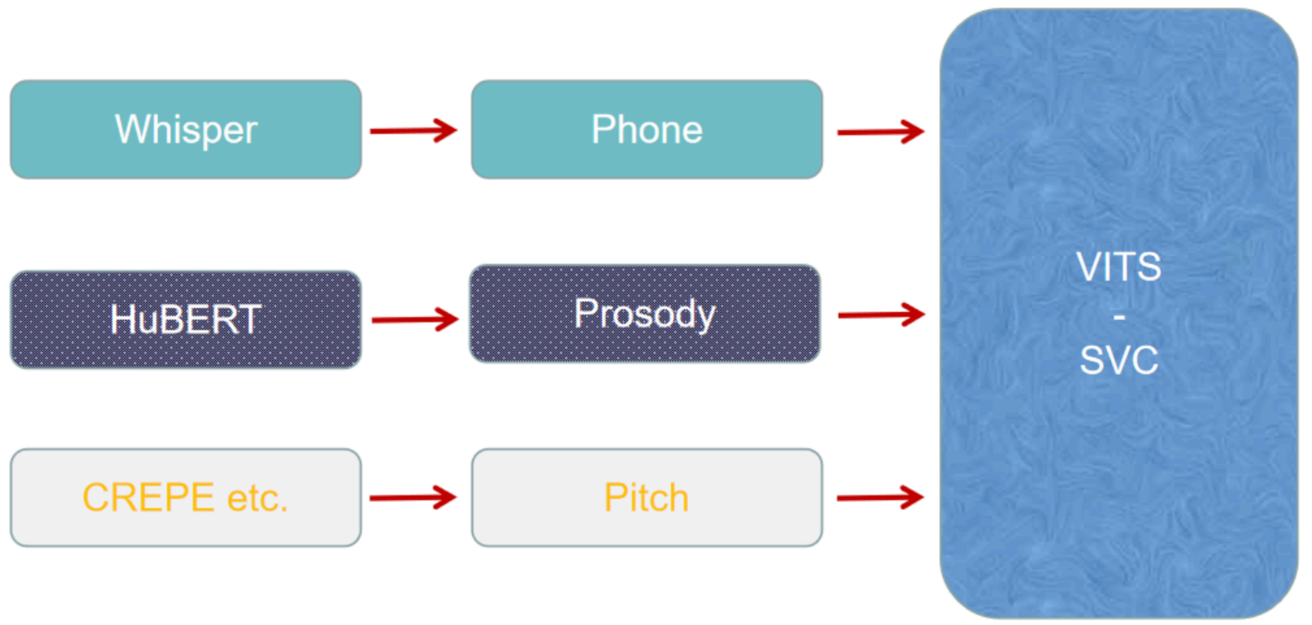
Conclusions

Requirements of SVC	Capability of the Semantic-based Features
To model melody	**Almost could not**
To model lyrics	The **pretraining data** effects the robustness
To model auxiliary (and speaker-agnostic) acoustic information	When using speaker ID, the information **“seems” to be speaker-agnostic**. *(However, there is timbre leakage issue especially for zero-shot setting.)*
To be robust for in-the-wild acoustic environment	The **pretraining data** effects the robustness

AI Singer Demo and Impact

- **Make Taylor Swift sing Mandarin song!**

Why mix



System	Naturalness (↑)	Similarity (↑)
Ground Truth	4.67 ± 0.18	3.17 ± 0.29
FastSVC	1.02 ± 0.04	1.23 ± 0.15
SoVITS	2.98 ± 0.31	2.74 ± 0.31
Ours (Whisper + ContentVec)	**3.52 ± 0.26**	**2.95 ± 0.30**
Ours (WeNet + Whisper + ContentVec)	3.48 ± 0.29	2.56 ± 0.35

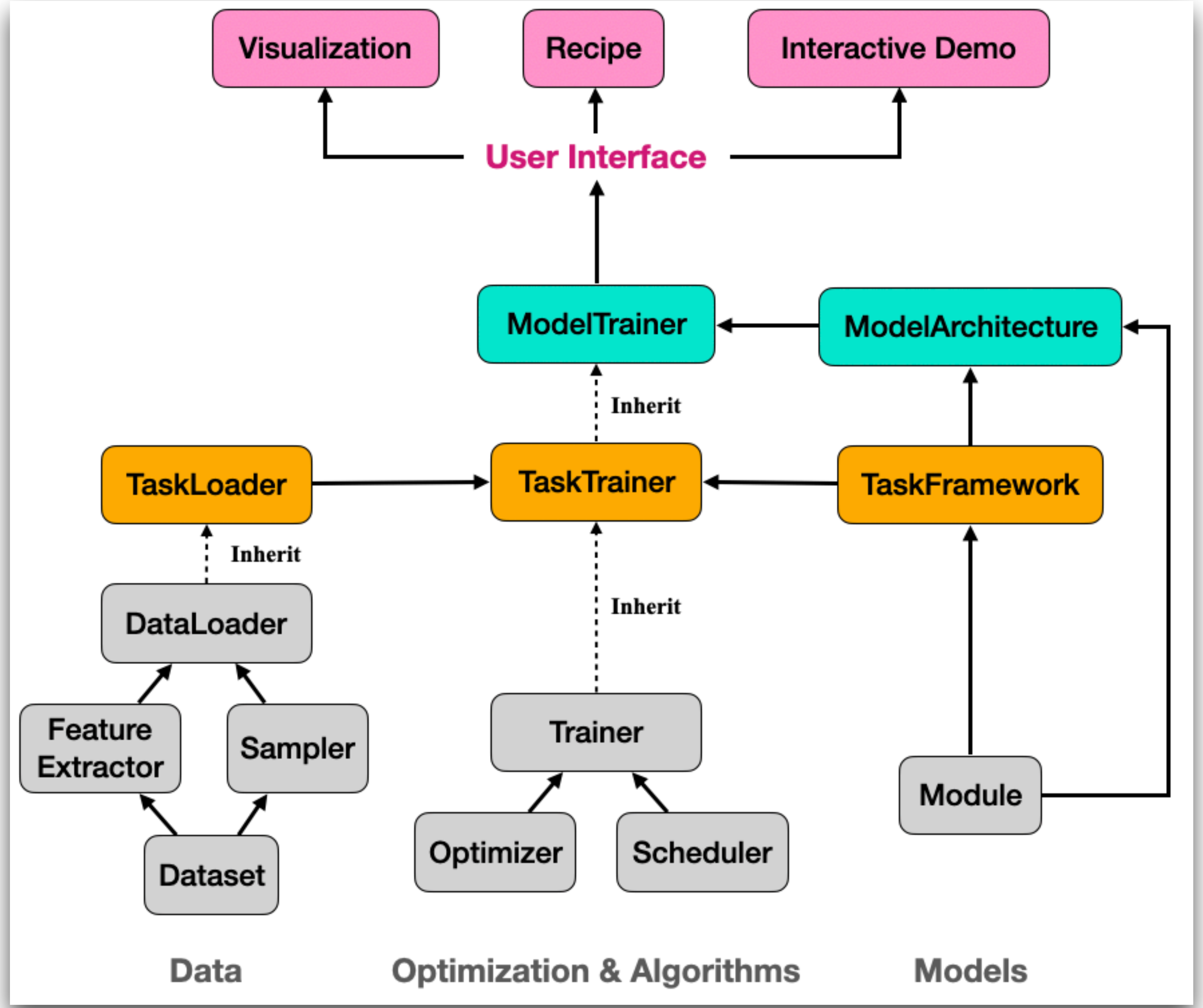
Leveraging Content-based Features from Multiple Acoustic Models for Singing Voice Conversion

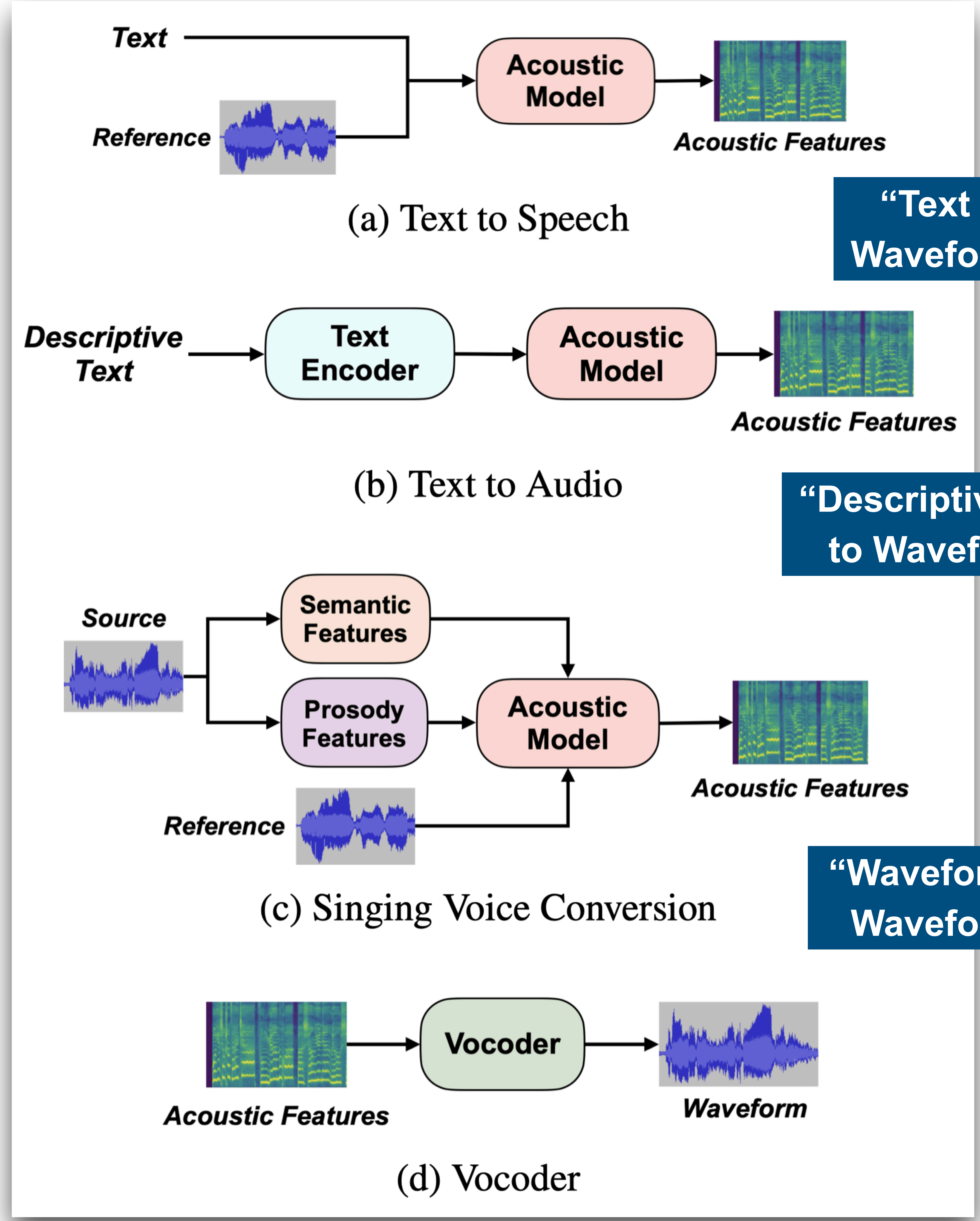
- **Our idea of using multiple content features has been borrowed and integrated into So-VITS-SVC 5.0 (Github over 2.7k stars)**

Roadmap

- **Singing Voice Conversion**
  - Definition, Classic Works, and Modern Pipeline
- **Singing Voice Conversion in Amphion**
  - Supported Model Architectures
  - Our research: *Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion*
- **Amphion's Philosophy**
  - Unique strengths, Supported Features, and Visualization
- **Singing Voice Conversion: Next Steps**

Strength1: Unified Audio Generation Framework





"Text to Waveform"

"Descriptive Text to Waveform"

"Waveform to Waveform"

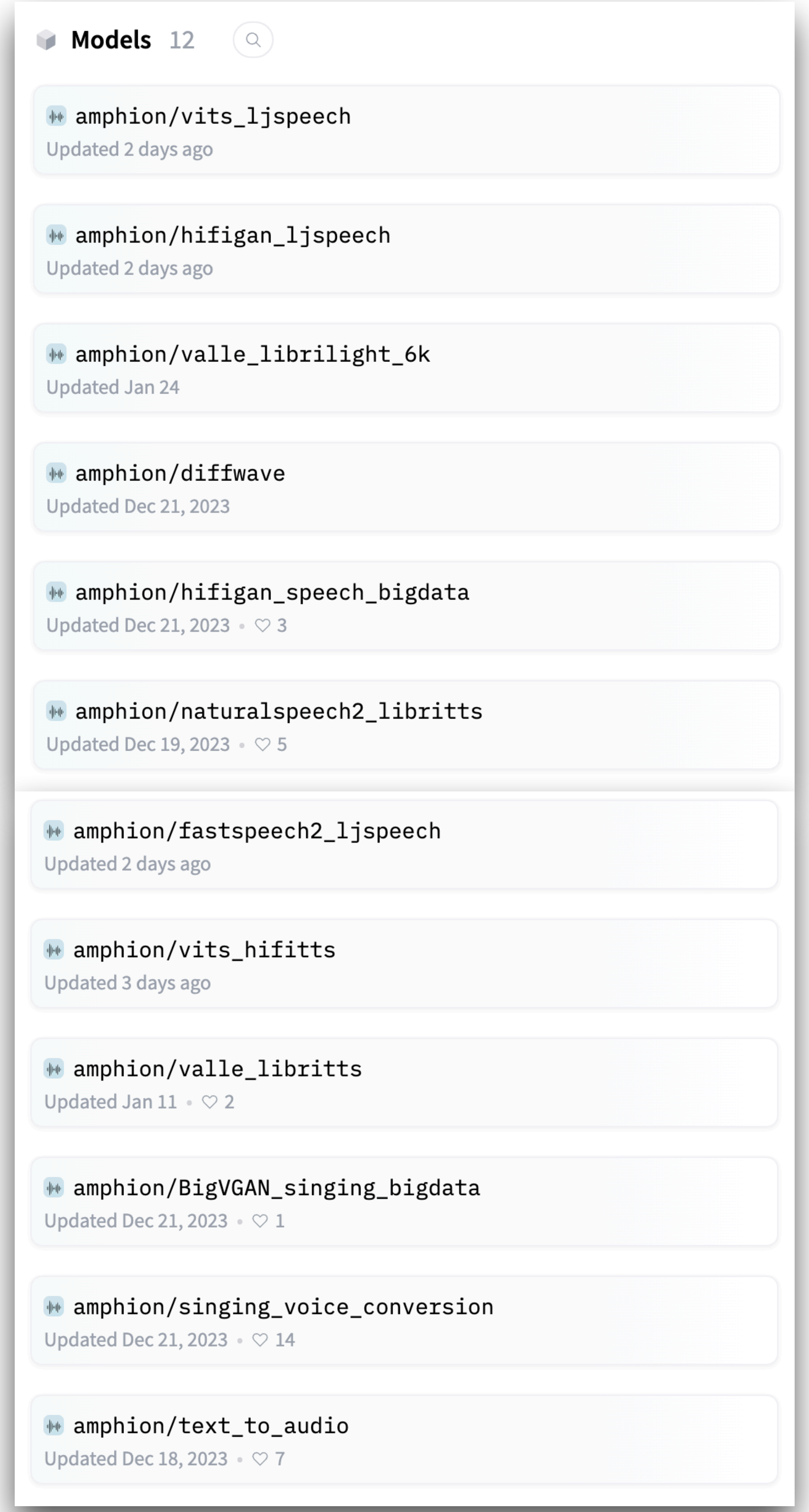
Strength2: Beginner-friendly End-to-End Workflow

Visualization & Education	Visual Analysis Tools	Interactive Demos	
Model Inference & Evaluation	High-quality Pre-trained Vocoders	Unified Evaluation Scripts	Various Metrics
Model Training	Multi GPU Training with DDP	Dynamic Batch Sizes	Training Progress Monitor
Feature Extraction	Offline Feature Extraction	On-the-fly Feature Extraction	
Data Preprocessing	Support both Academic and User Custom Datasets		

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

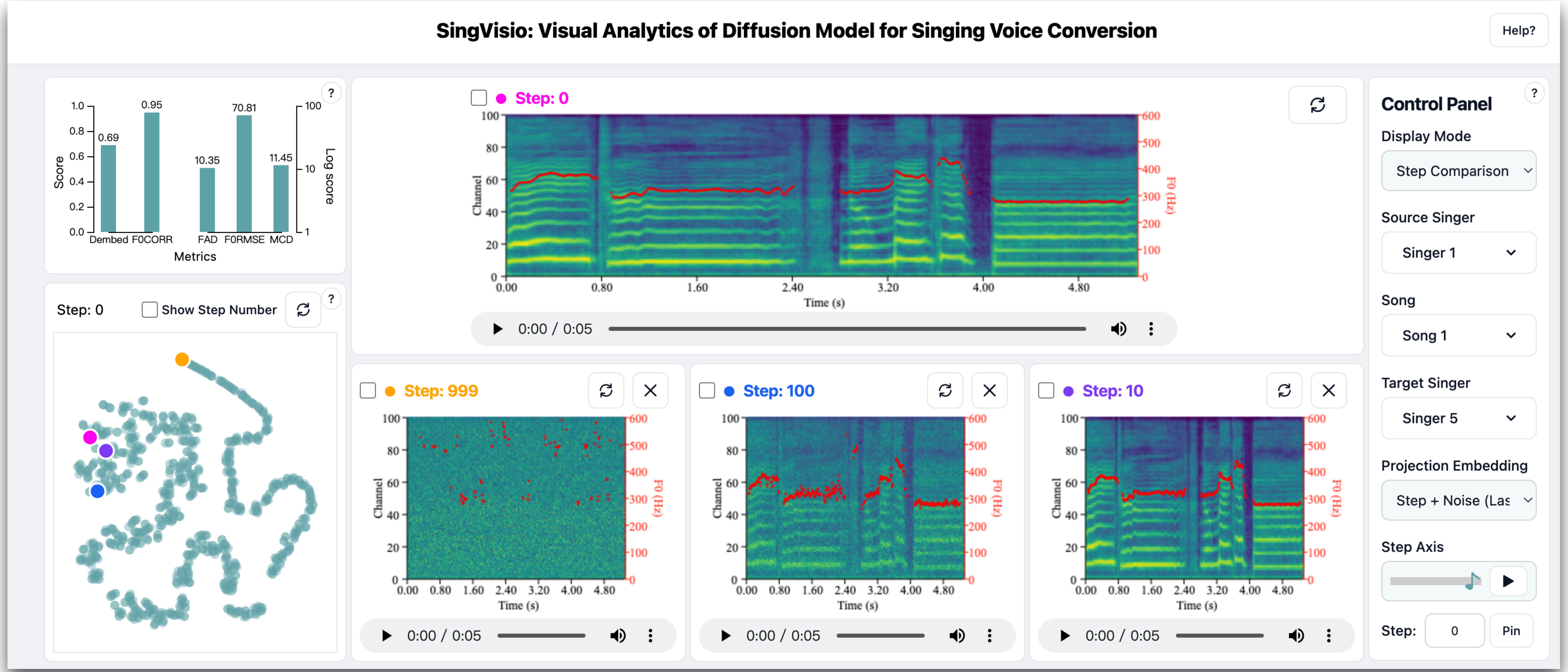
Strength3: Open Pre-trained Models

Release Criteria	Description
Model Metadata	Detail the model architecture and the number of parameters.
Training Datasets	List all the training corpus and their sources.
Training Configuration	Detail the training hyberparameters (like batch size, learning rate, and number of training steps) and the computational platform
Evaluation Results	Display the evaluation results and the performance comparison to other typical baselines.
Usage Instructions	Instruct how to inference and fine-tune based on the pre-trained model.
Interactive Demo	Provide an online interactive demo for users to explore.
License	Clear the licensing details including how the model can be utilized, shared, and modified.
Ethical Considerations	Address ethical considerations related to the model's application, focusing on privacy, consent, and bias, to encourage responsible usage.



Supported Pretrained Models (Updating)

Strength4: 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.

Roadmap

- **Singing Voice Conversion**
  - Definition, Classic Works, and Modern Pipeline
- **Singing Voice Conversion in Amphion**
  - Supported Model Architectures
  - Our research: *Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion*
- **Amphion's Philosophy**
  - Unique strengths, Supported Features, and Visualization
- **Singing Voice Conversion: Next Steps**

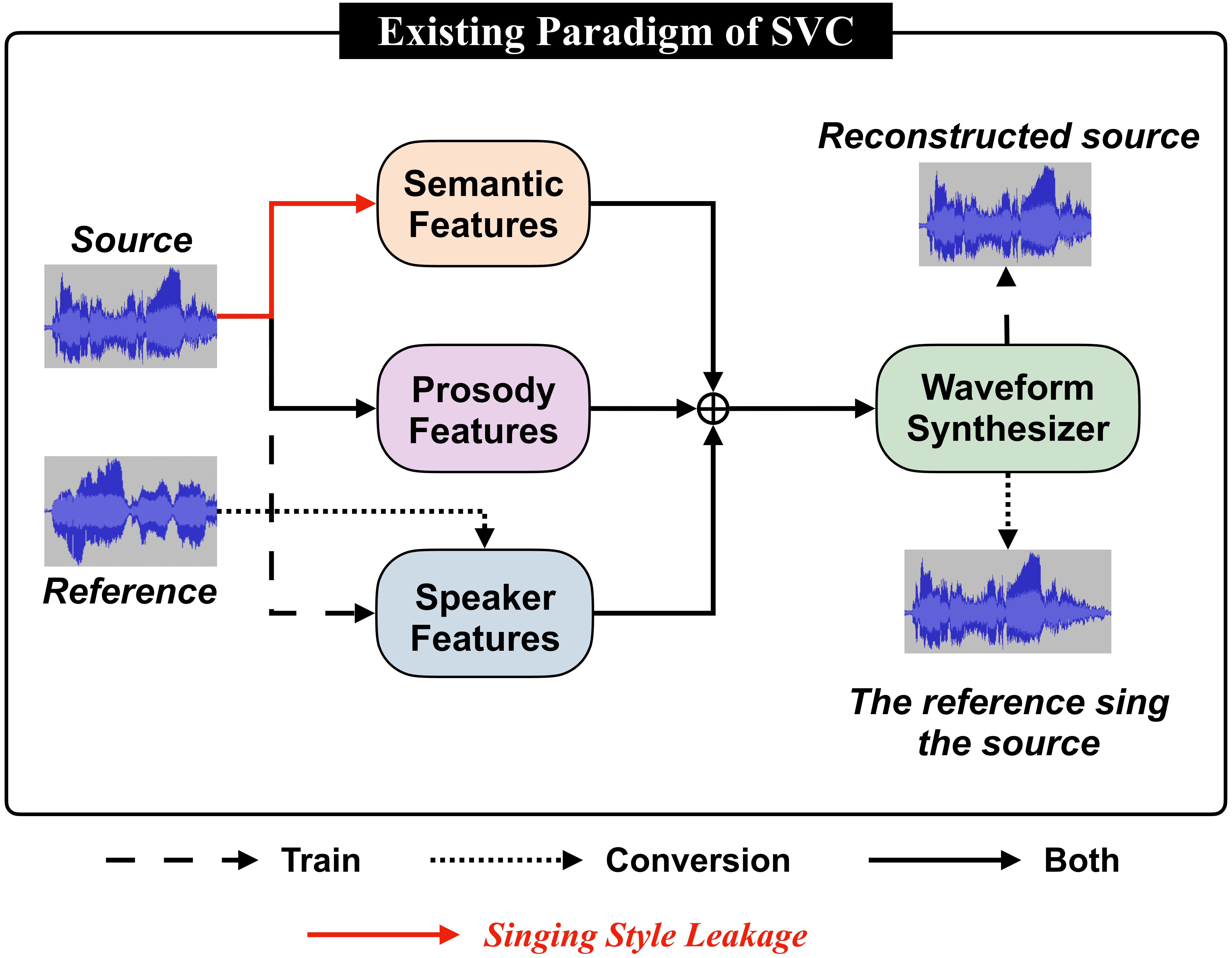
Challenges: To Clone *Singing Style* Beyond Timbre

	Source	Conversion Results [1]	Ground Truth
韩红 to 李健			
齐秦 to 李健			
张学友 to 李健			-
林志炫 to 李健			-
陶喆 to 李健			-

***Timbre* (音色) has been cloned, but the imitation of *singing style* (唱法) still has a long way to go.**

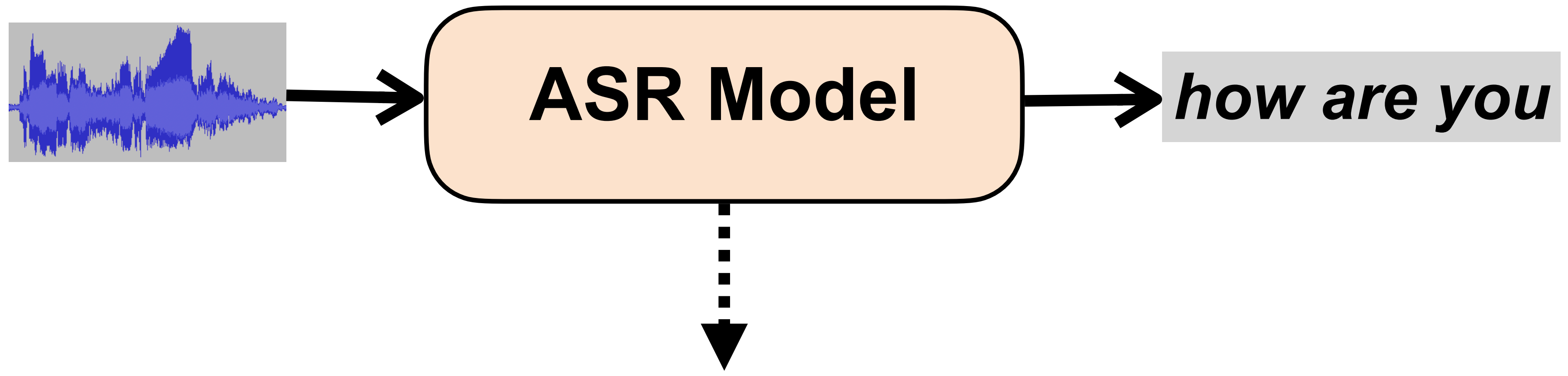
[1] **Xueyao Zhang**, et al. *Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion.* IEEE SLT 2024.

Singing style is just repeating like source!



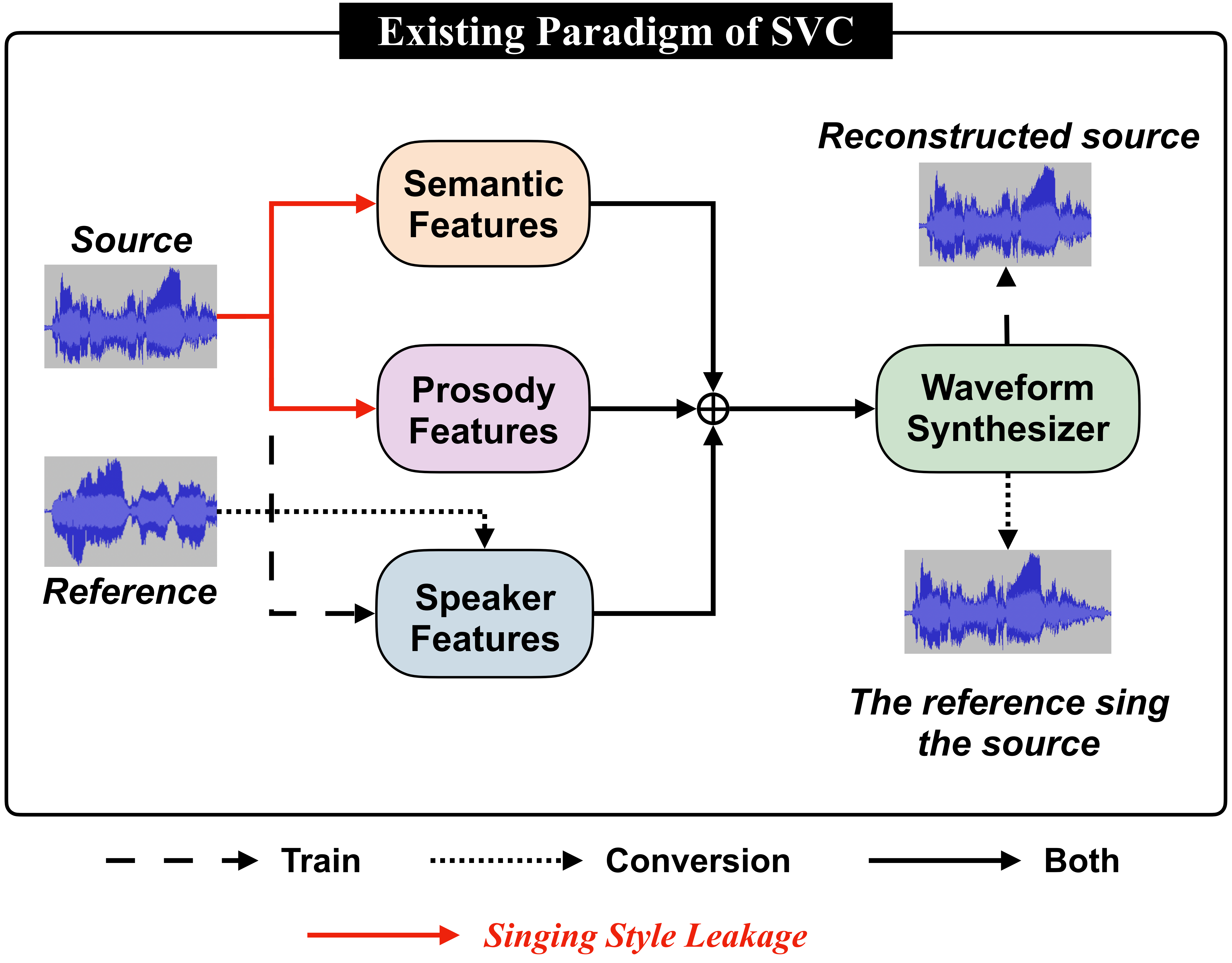
🤔 Where are singing style from?

① Semantic Features



- Using the intermediate output as "semantic-*based*" features (a.k.a., PPG or BNF), there could be singing expression leakage including ***phonetic timing*** and ***articulation patterns***.

Singing style is just repeating like source!

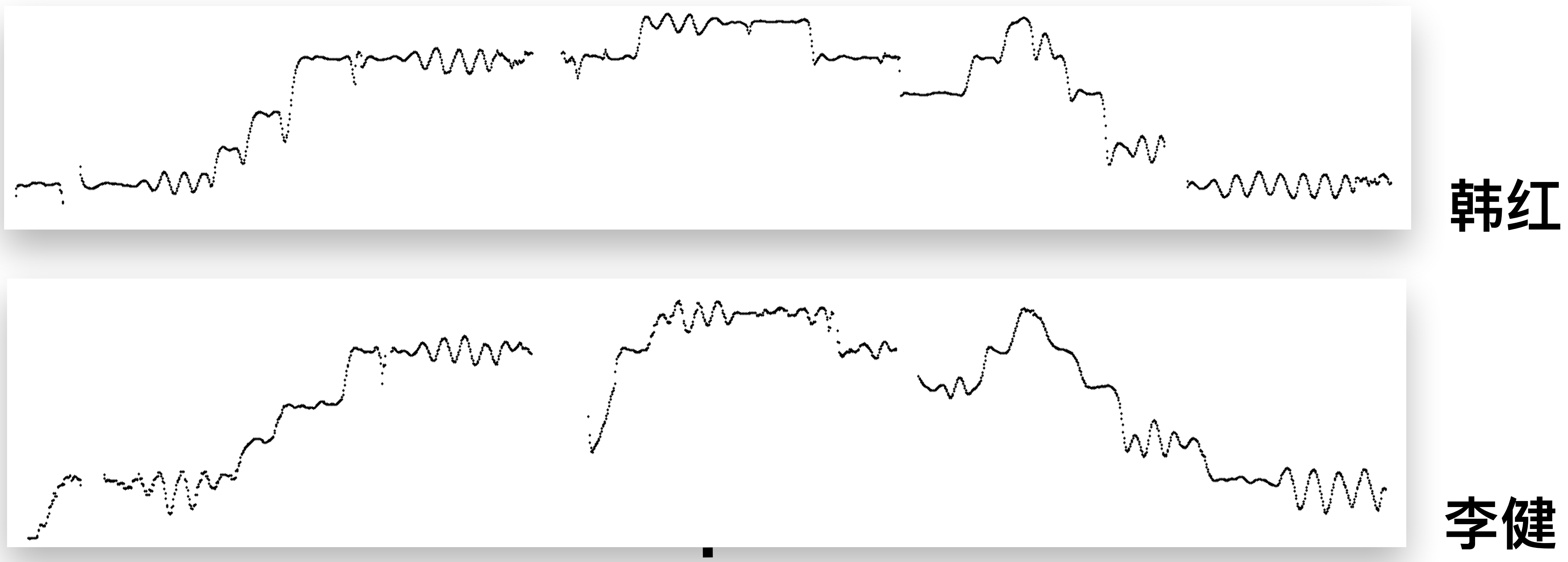


🤔 **Where are singing style from?**

① Semantic Features

② Prosody Features

Same song sung by different singers



- Using fundamental frequency (f0) as prosody features, there could be singing expression style including ***musical note timing*** and ***vibrato patterns***.

THANKS

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 v0.1's co-founder**
  Project: https://github.com/open-mmlab/Amphion **(7.8k stars)**

- ✦ **Research interest: "AI + Music"**, especially on:
  - Singing Voice Processing
  - Music Generation

Amphion Official Account

📎 **Amphion Technical Report:** https://arxiv.org/abs/2312.09911
🧑‍💻 **Amphion GitHub:** https://github.com/open-mmlab/Amphion
🎯 **Amphion Demos/Models/Datasets:** https://huggingface.co/amphion