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

2024/02

About me

Xueyao Zhang (张雪遥)

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

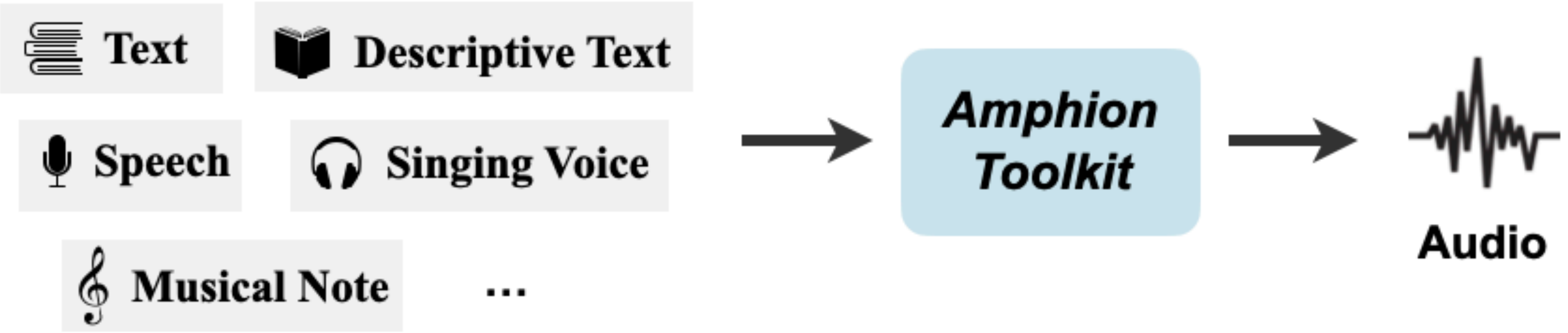
- ✦ **Amphion v0.1's co-founder**
  Project: https://github.com/open-mmlab/Amphion **(3.5k stars)**

- ✦ **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...

[Amphion + Sora](#)

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

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

Parallel corpus is hard to collect!

Non-Parallel Singing Voice Conversion

Professional Singer1

→

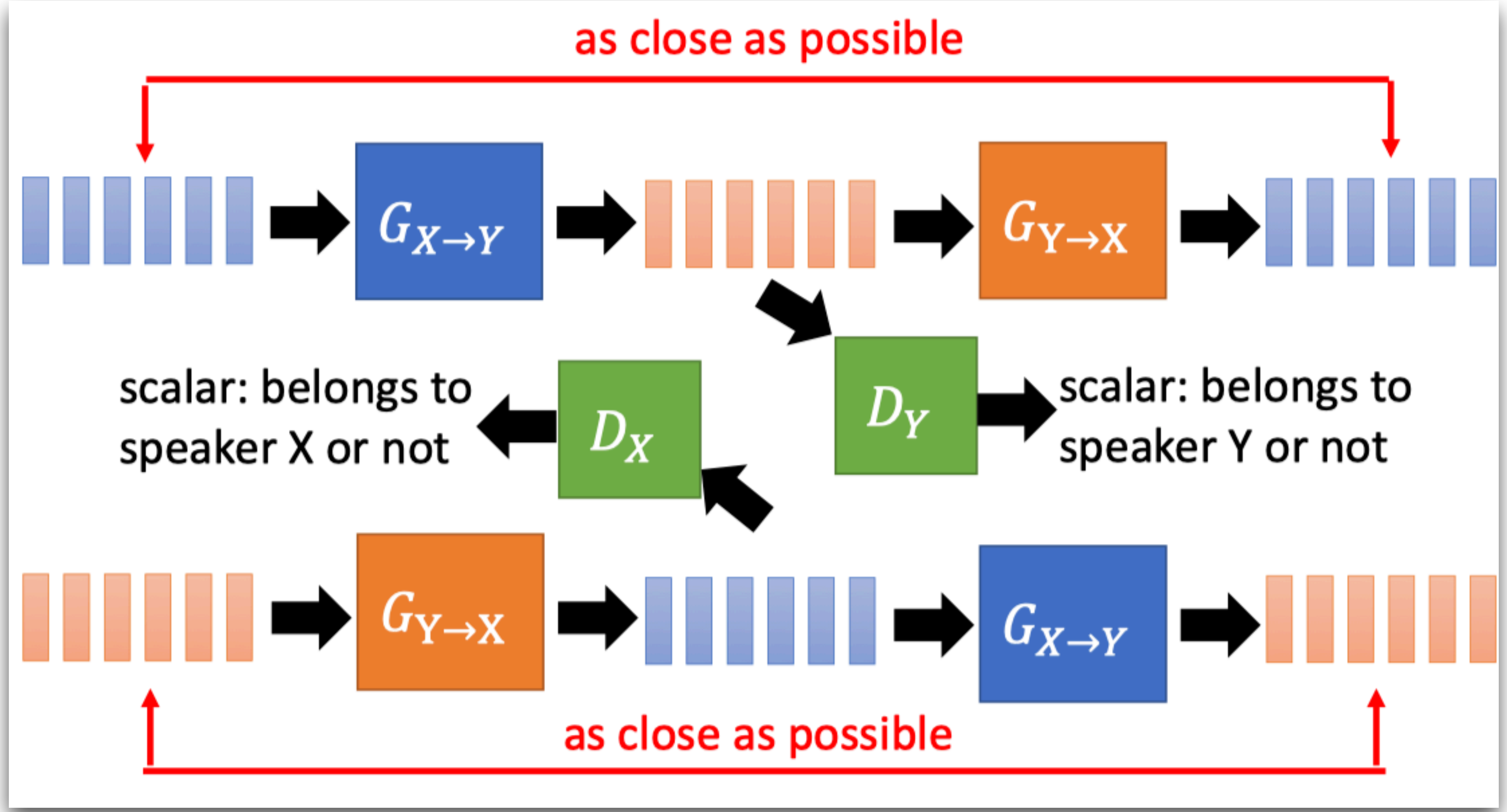
Professional Singer2

X

Singer1's Songs Singer2's Songs

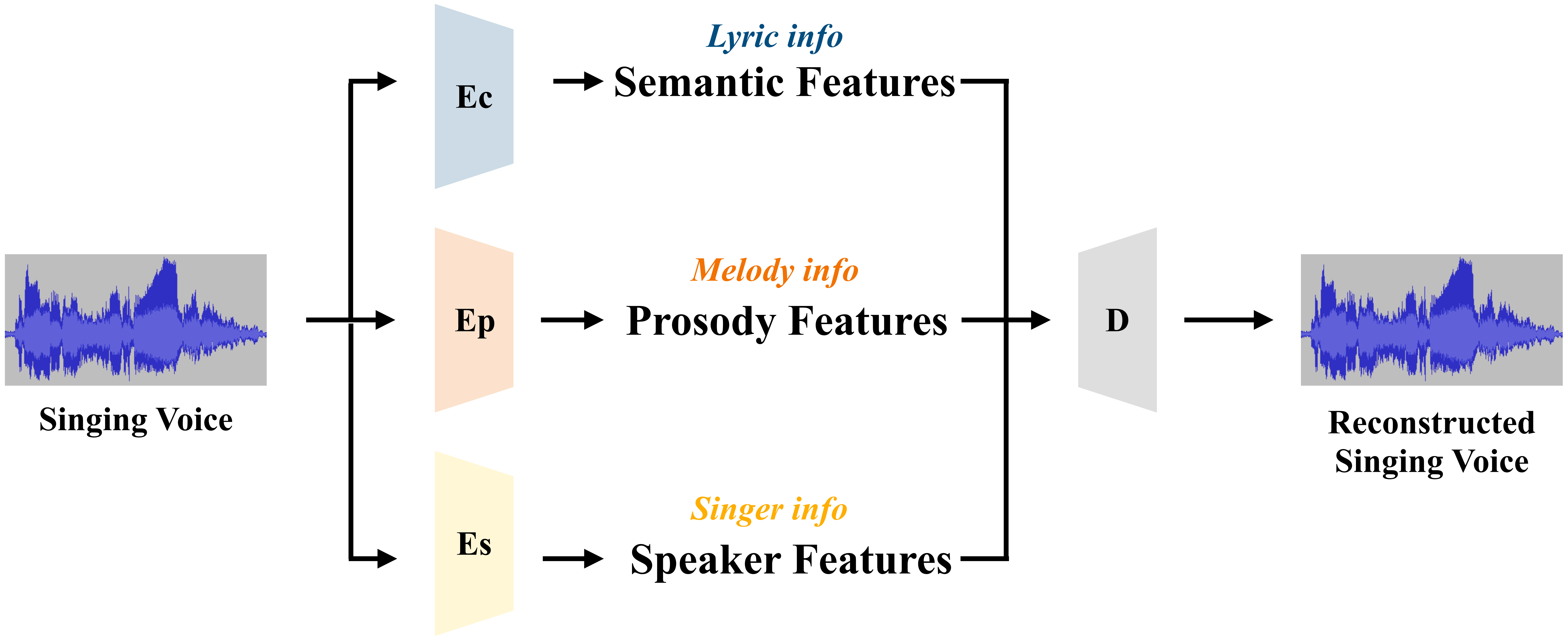
How to decouple the singer identity?

Non-Parallel SVC: GAN School



Credit: Voice Conversion, Hung-yi Lee.

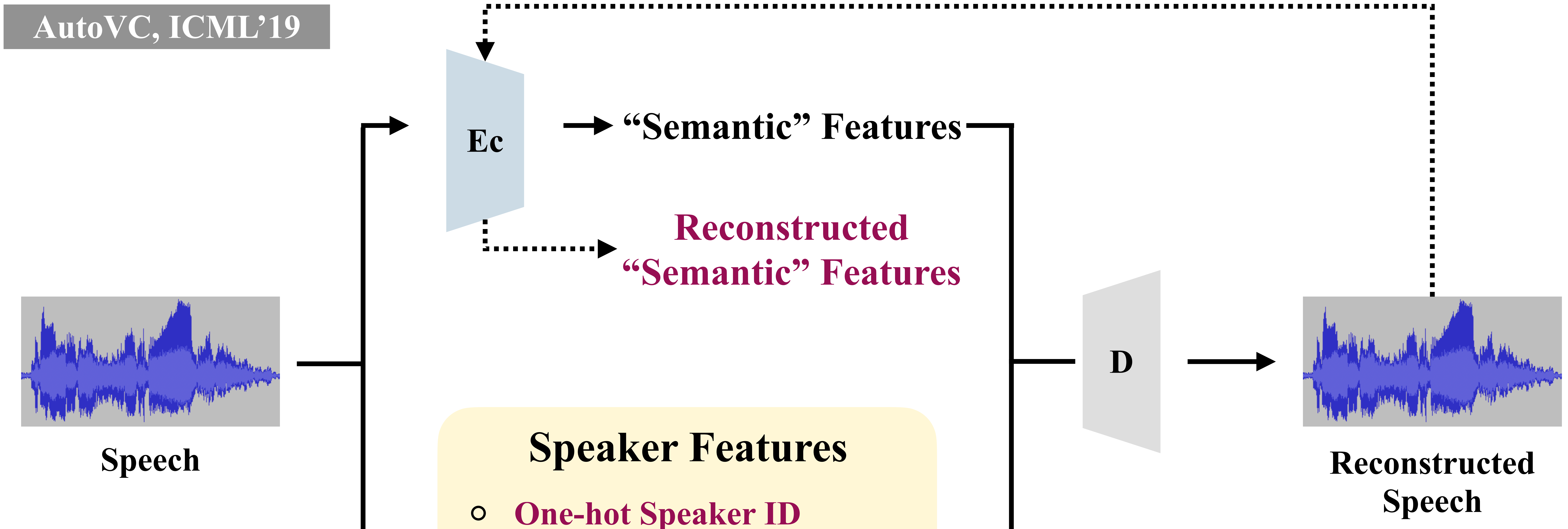
Non-Parallel SVC: Auto-Encoder School



- 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



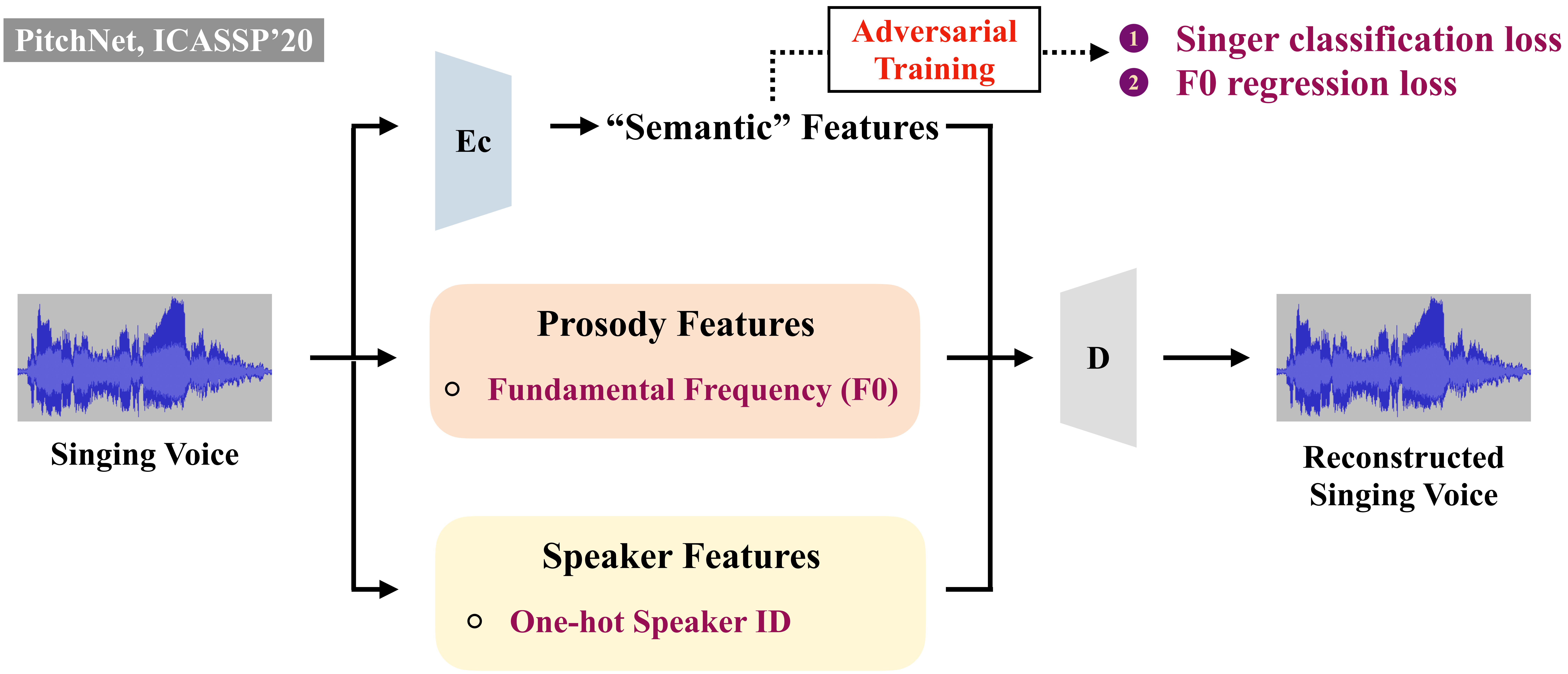
Speaker Features

- One-hot Speaker ID
- Features extracted from speaker verification model

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

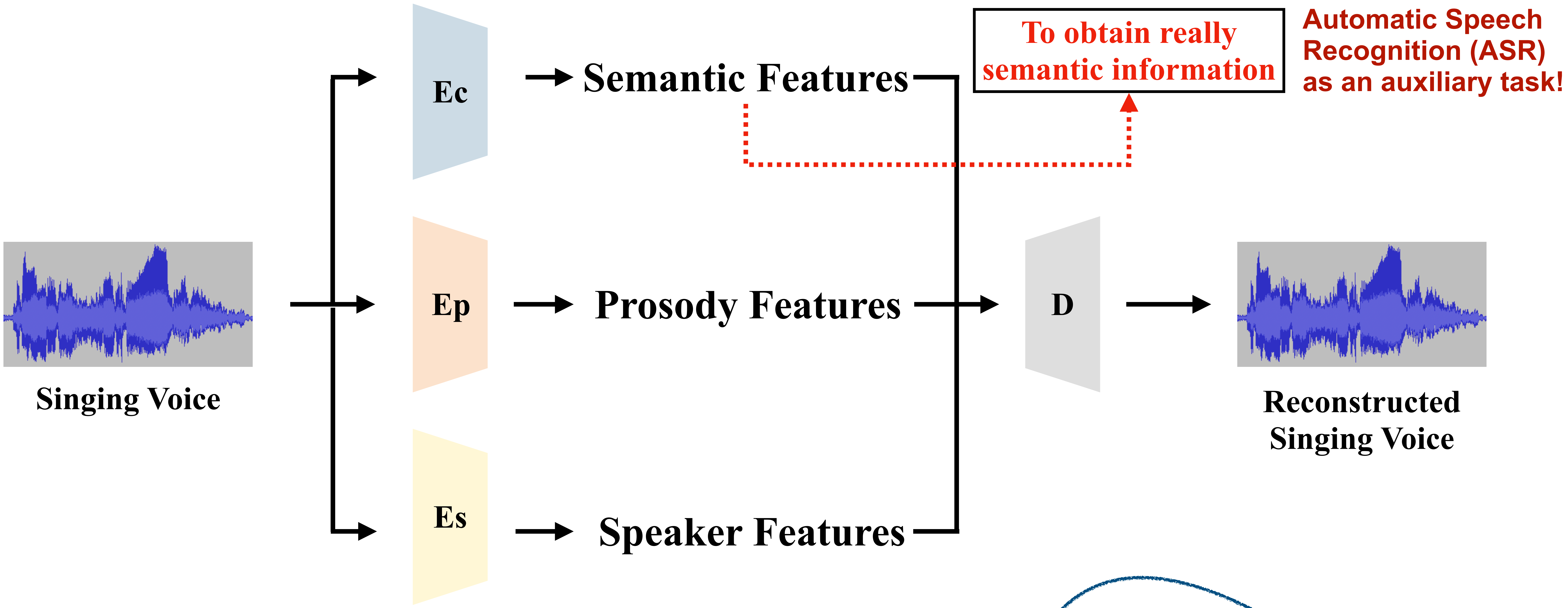
Auto-Encoder SVC: The Early Researches

PitchNet, ICASSP'20



PitchNet: "Adopt adversarial training to disentangle better"

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

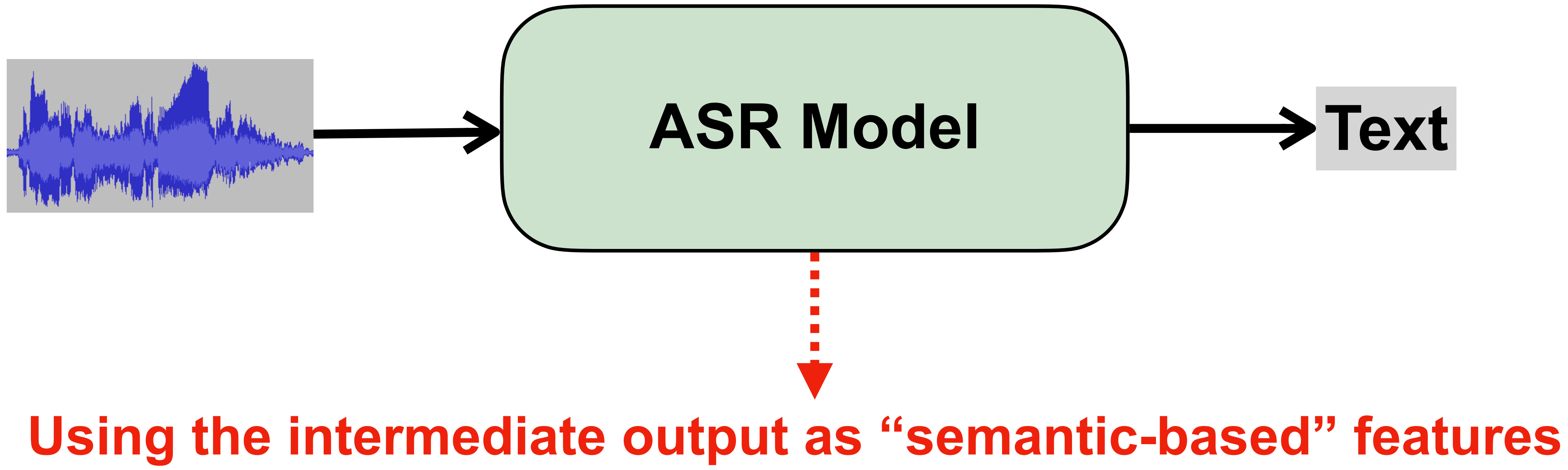


- How to ensure the disentanglement of different features?
- How to ensure there is enough information of each features?

🎯 Solved to some extent

🤔 How to address?

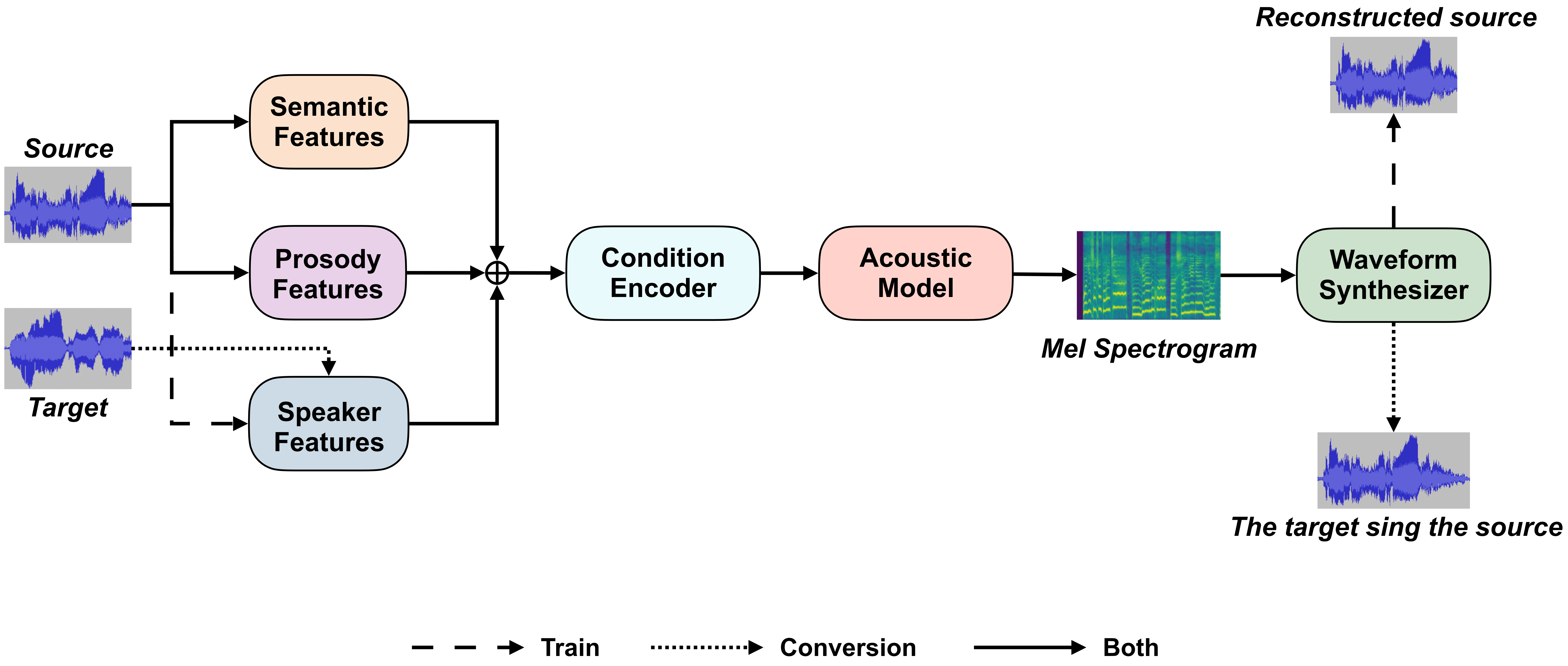
Non-Parallel VC/SVC — a.k.a Recognition & Synthesis VC/SVC



🤔 Why do we use the *dense 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

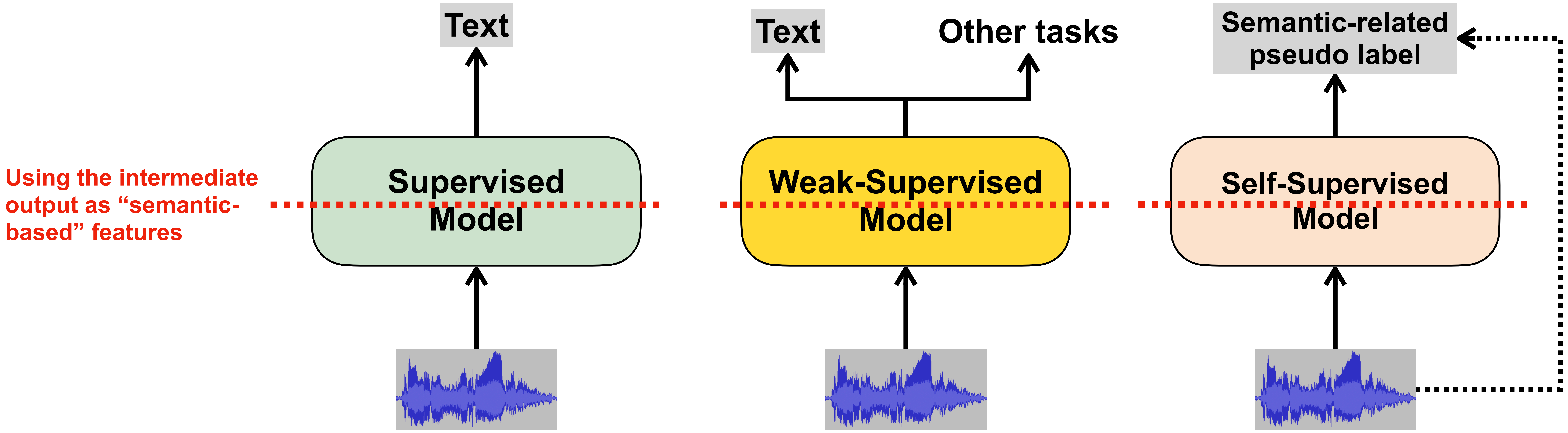
- **Singing Voice Conversion**
  - Definition, Classic Works, and Modern Pipeline
- **Singing Voice Conversion in Amphion**
  - Supported Model Architectures
  - Our research: *Leveraging Diverse Semantic-based Audio Pretrained Models for Singing Voice Conversion*
- **Amphion's Philosophy**
  - Unique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

Amphion SVC: Supported Model Architectures

- **Semantic Features Extractor**
  - WeNet, Whisper, ContentVec
  - 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

Semantic Features: Why Joint Usage of Multiple Extractors?

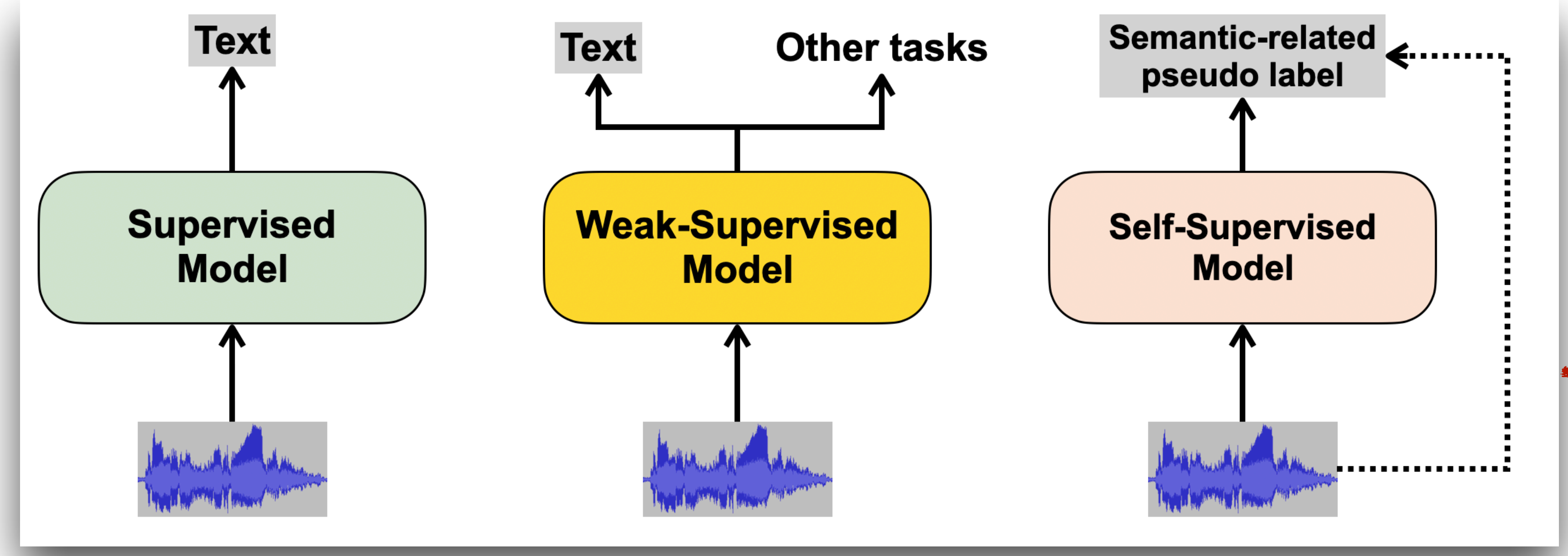
- **Background of Semantic-based Pretrained Models:**
  - **Varied choices**: Classic ASR models, Whisper, HuBERT, Wav2Vec, WavLM, ContentVec…
  - **High pretraining cost**: E.g., Whisper-large (1.5B parameters, 680k hours training data)



Semantic Features: Why Joint Usage of Multiple Extractors?

- **However, for the downstream tasks (such as SVC), the underlying knowledge of these models remains largely unknown:**

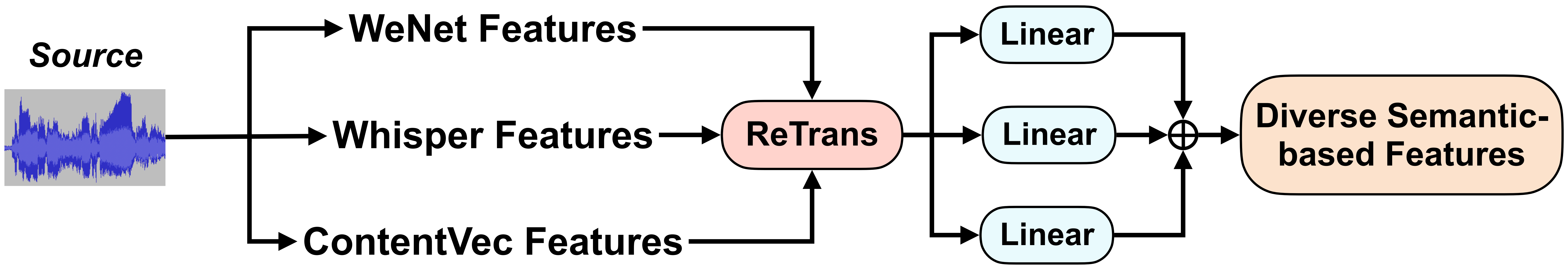
Requirements of SVC	Capability of the Semantic-based Features
To model melody	**Whether could or not** remains unknown
To model lyrics	Could. But **exactly how much** remains unknown
To model auxiliary acoustic information	Could. But **whether the information is speaker-agnostic or not** remains unknown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



Different pretraining ways will yield different underlying knowledge

🤔 **Could diverse pretrained models be complementary for SVC?**

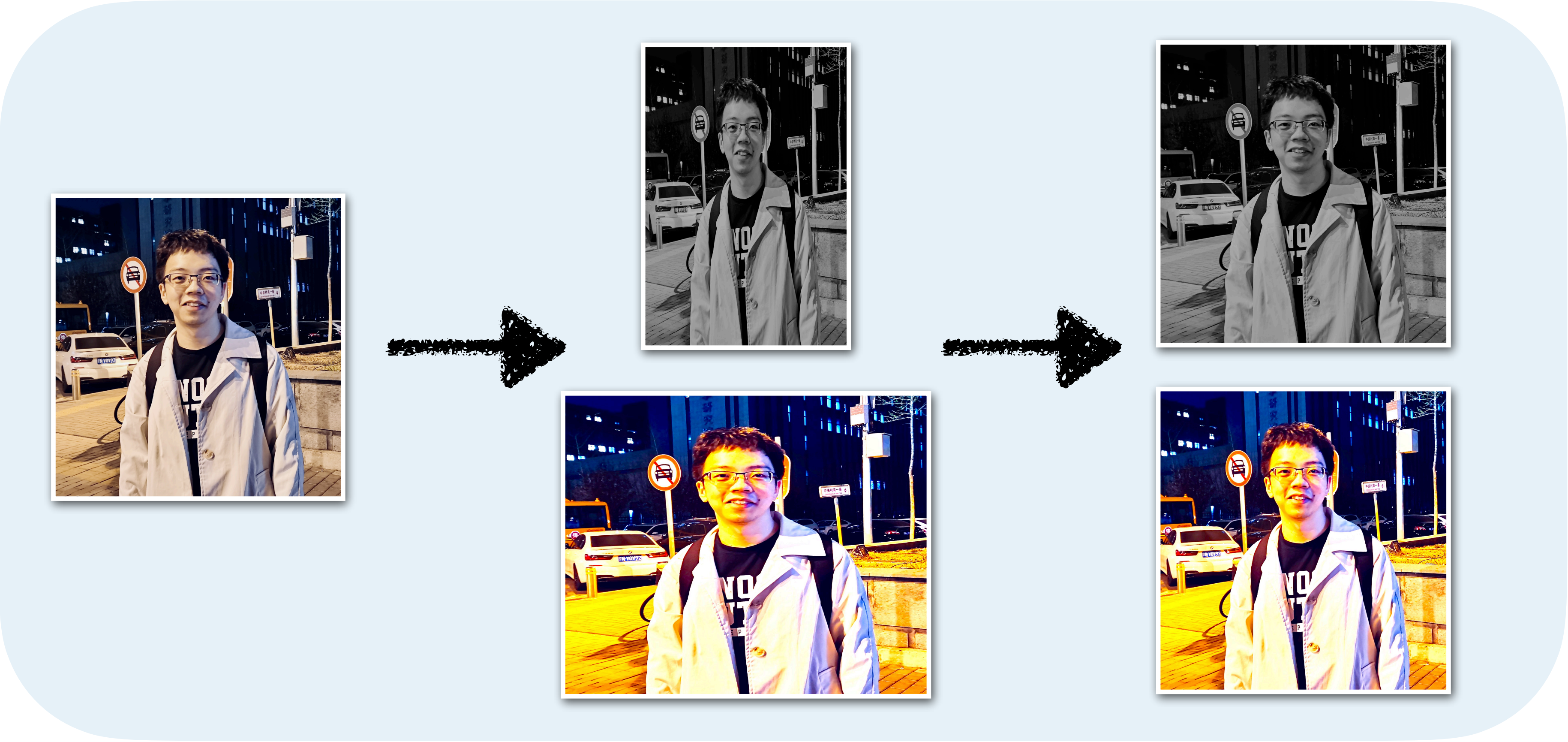
Challenge: Time Resolution Mismatch of Multiple Semantic Features



Algorithm 1 ReTrans for features of audio pretrained models

Input: $\mathbf{f}$ – the features to be transformed whose frame rate is r_1 Hz.
Parameter: r_2 – The desired frame rate to transform
Output: $\mathbf{f}'$ – the transformed features whose frame rate is r_2 Hz

1: $c \leftarrow$ **gcd** (r_1, r_2) ▷ The greatest common divisor
2: $\mathbf{f}' \leftarrow$ **upsample** $(\mathbf{f}, r_2/c)$ ▷ Upsampling
3: $\mathbf{f}' \leftarrow$ **downsample** $(\mathbf{f}', r_1/c)$ ▷ Downsampling
4: **return** $\mathbf{f}'$



High efficiency;
No more training cost

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**
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 **To model melody:**

Whisper > ContentVec > WeNet

② **To model lyrics:**

Whisper > ContentVec > WeNet

③ **To be speaker-agnostic:**

All the three is good

- ✻ **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

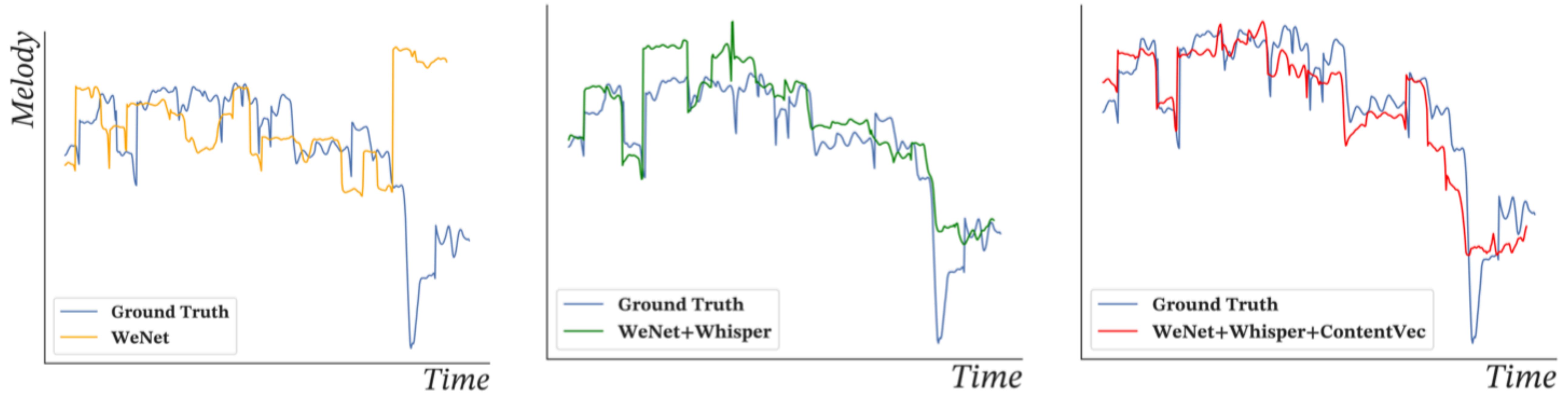
②

After Introducing F0

Using diverse semantic-based features:

① Most results are promoted stage by stage

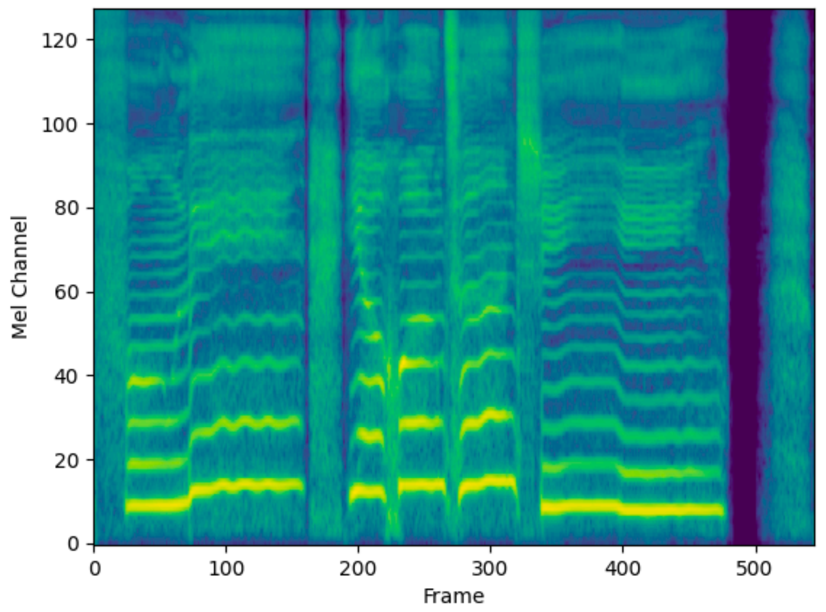
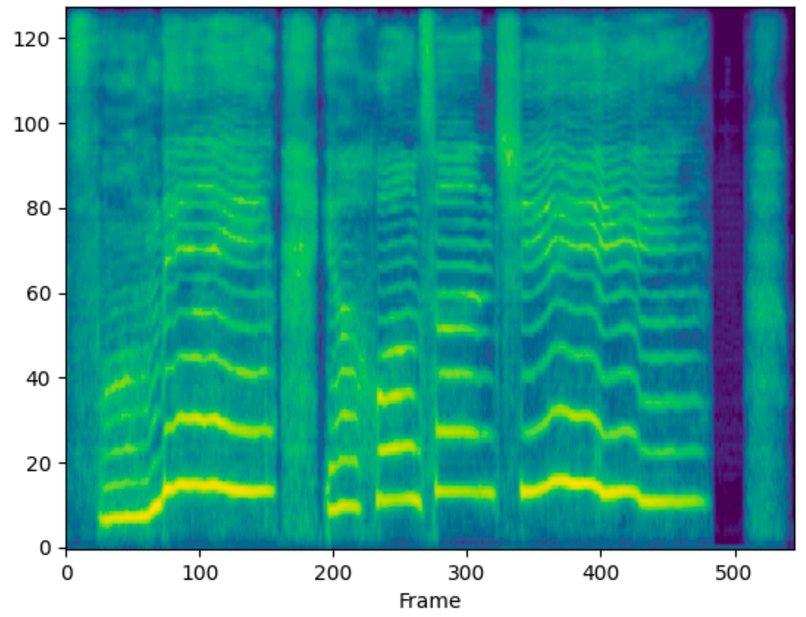
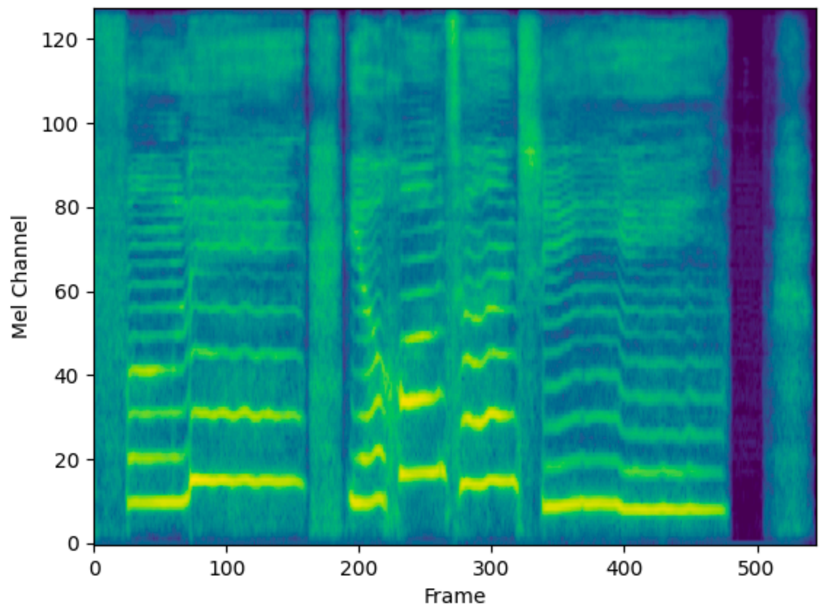
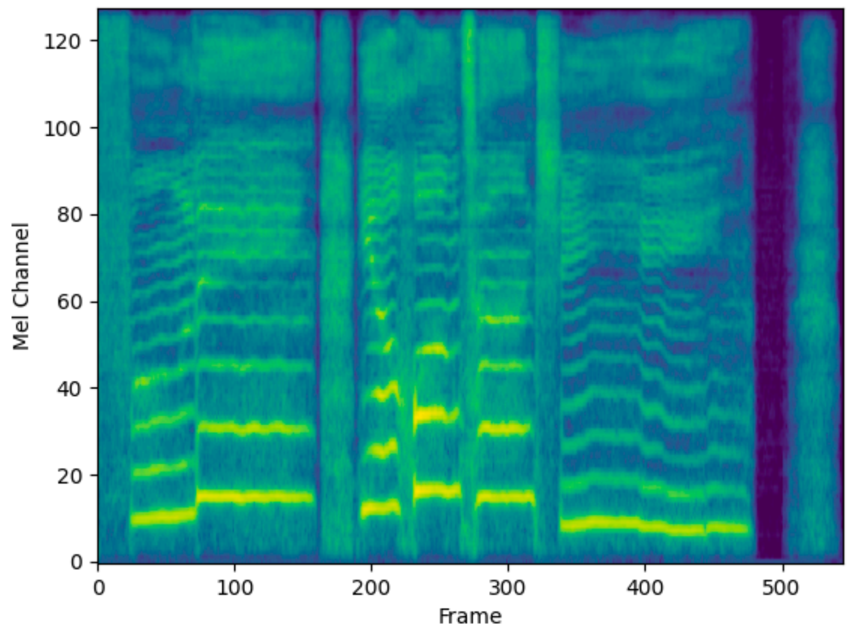
② Introducing explicit melody modeling for SVC remains necessary



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

Results: Complementary roles of Diverse Semantic-based Features

Spectrogram Reconstruction

	WeNet	WeNet + Whisper	WeNet + Whisper + ContentVec
			
Ground Truth	MCD: 12.56	MCD: 8.00	MCD: 5.02

Intelligibility

	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%

Results: SVC Framework based on Diverse Semantic-based Features Fusion

Objective Evaluation

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	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	21.1%	0.865
	WeNet + Whisper + ContentVec	0.940	55.2	15.7%	0.884	0.919	79.9	23.3%	0.867

Subjective Evaluation for DiffWaveNetSVC

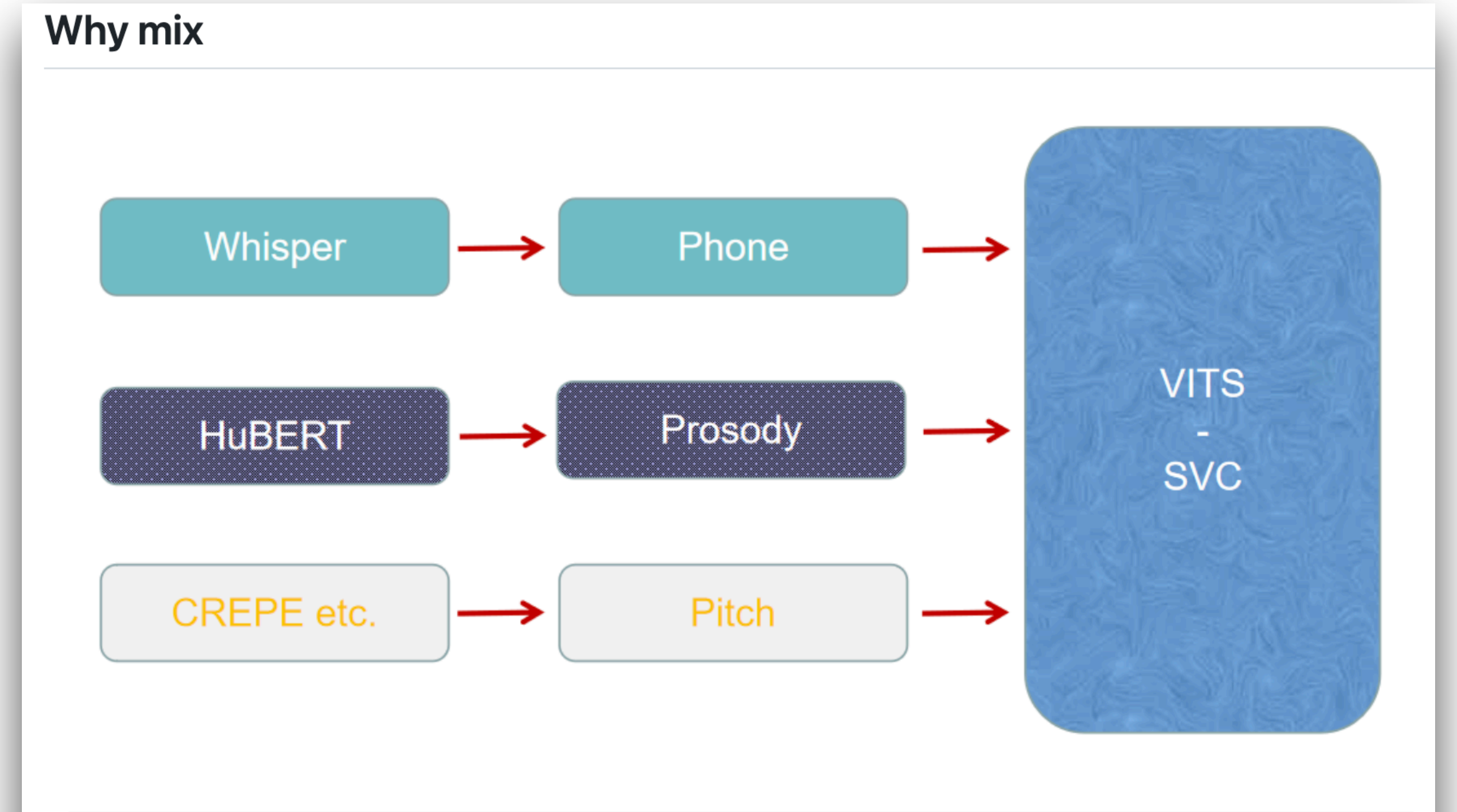
Semantic-based Features	Recording Studio Setting		In-the-Wild Setting	
	Naturalness (↑)	Similarity (↑)	Naturalness (↑)	Similarity (↑)
WeNet	2.72 $\pm$0.22	2.64 $\pm$0.21	2.85 $\pm$0.21	2.34 $\pm$0.20
WeNet + Whisper	4.02 $\pm$0.18	3.13 $\pm$0.17	3.70 $\pm$0.18	2.86 $\pm$0.23
WeNet + Whisper + ContentVec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

The full scores of Naturalness and Similarity are 5 and 4

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, such solution is also effective.

AI Singer Demo and Impact



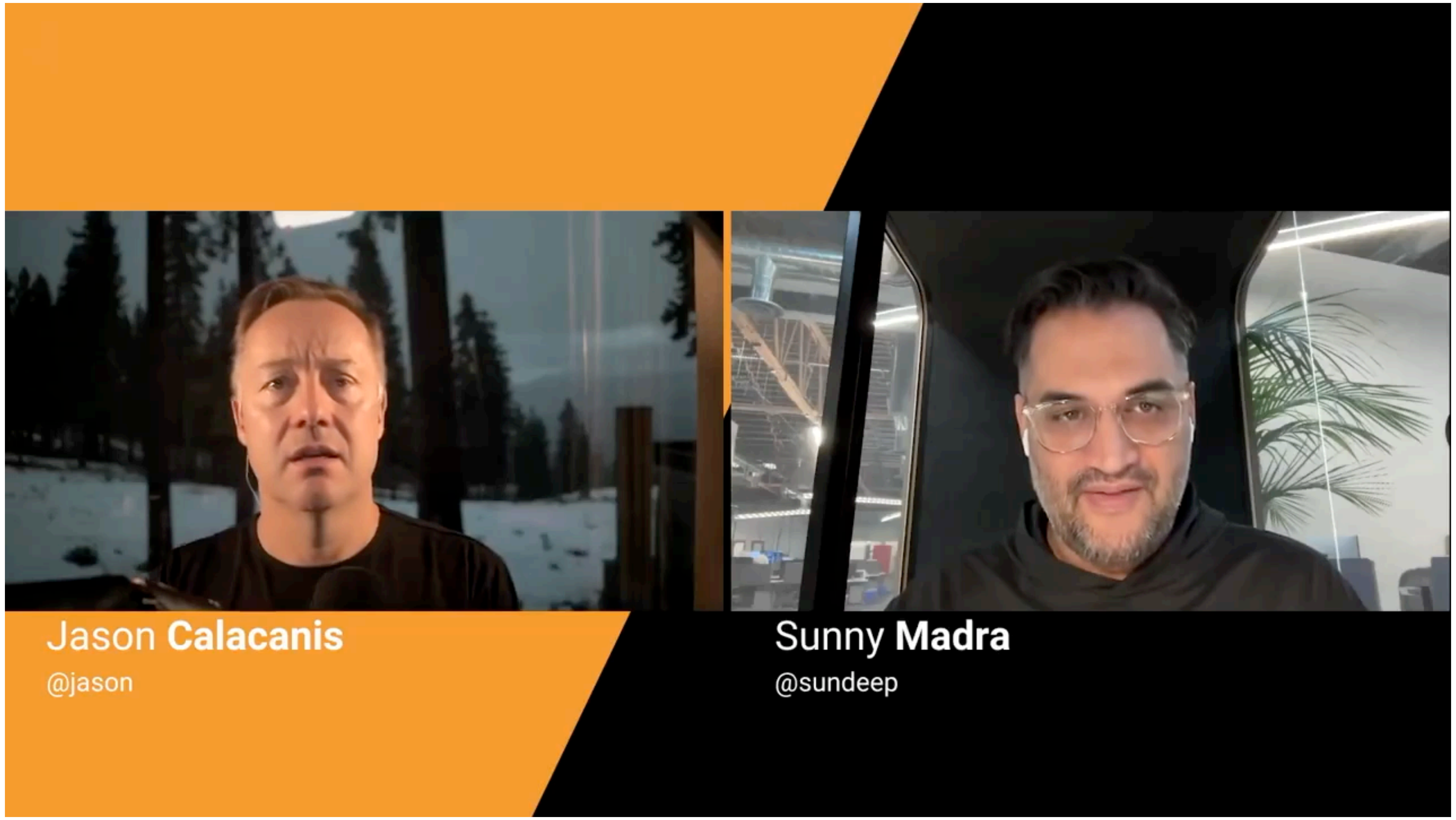
System	Naturalness (↑)	Similarity (↑)
Ground Truth	4.67 ± 0.18	3.17 ± 0.29
FastSVC	1.02 ± 0.04	1.23 ± 0.15
SoVITS	2.98 ± 0.31	2.74 ± 0.31
Ours (Whisper + ContentVec)	$\mathbf{3.52 \pm 0.26}$	$\mathbf{2.95 \pm 0.30}$
Ours (WeNet + Whisper + ContentVec)	3.48 ± 0.29	2.56 ± 0.35

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

- Make Taylor Swift sing Mandarin song!

- **Our idea of using multiple content features has been borrowed and integrated into So-VITS-SVC 5.0 (Github over 2k stars)**

AI Singer Demo and Impact

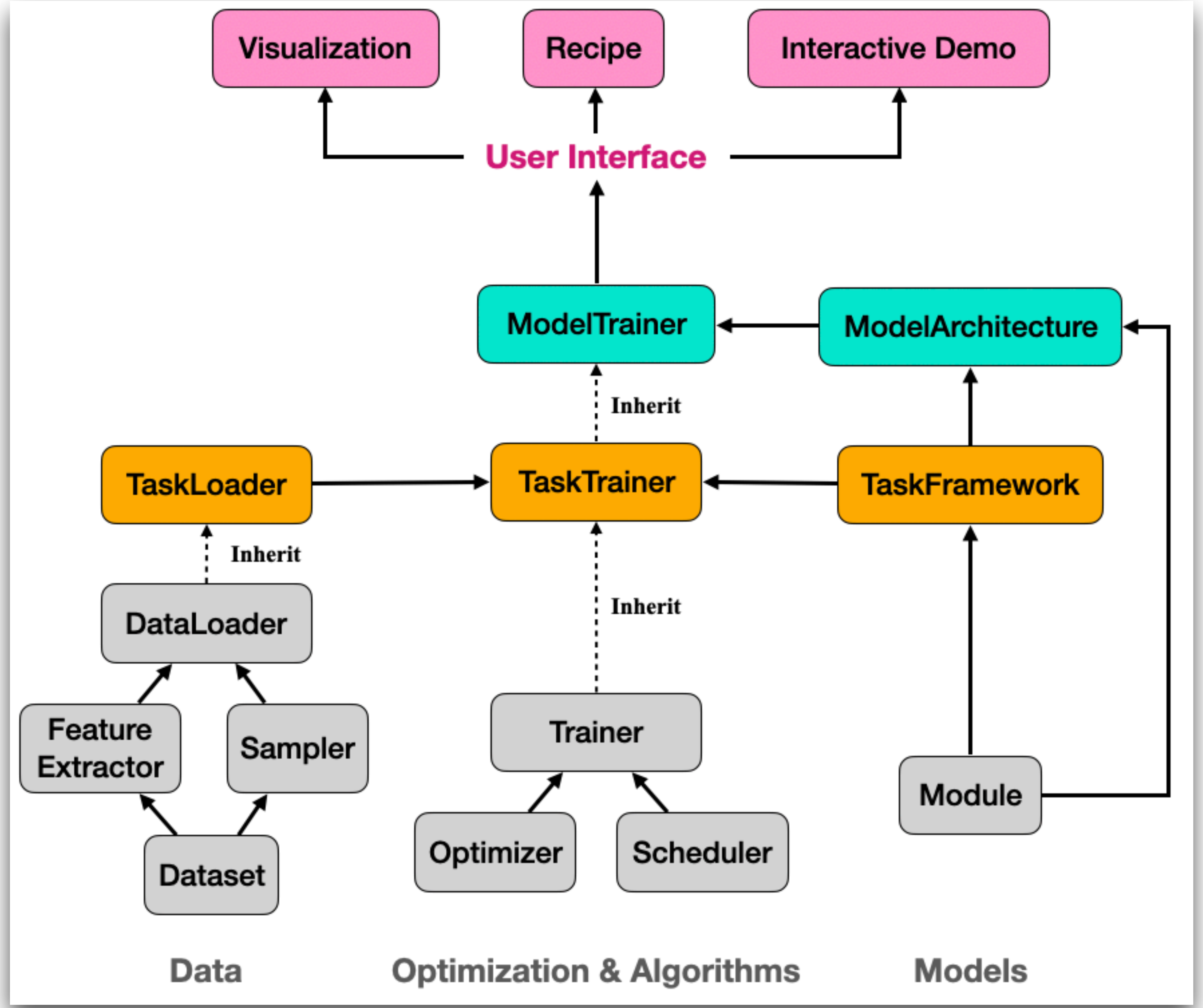


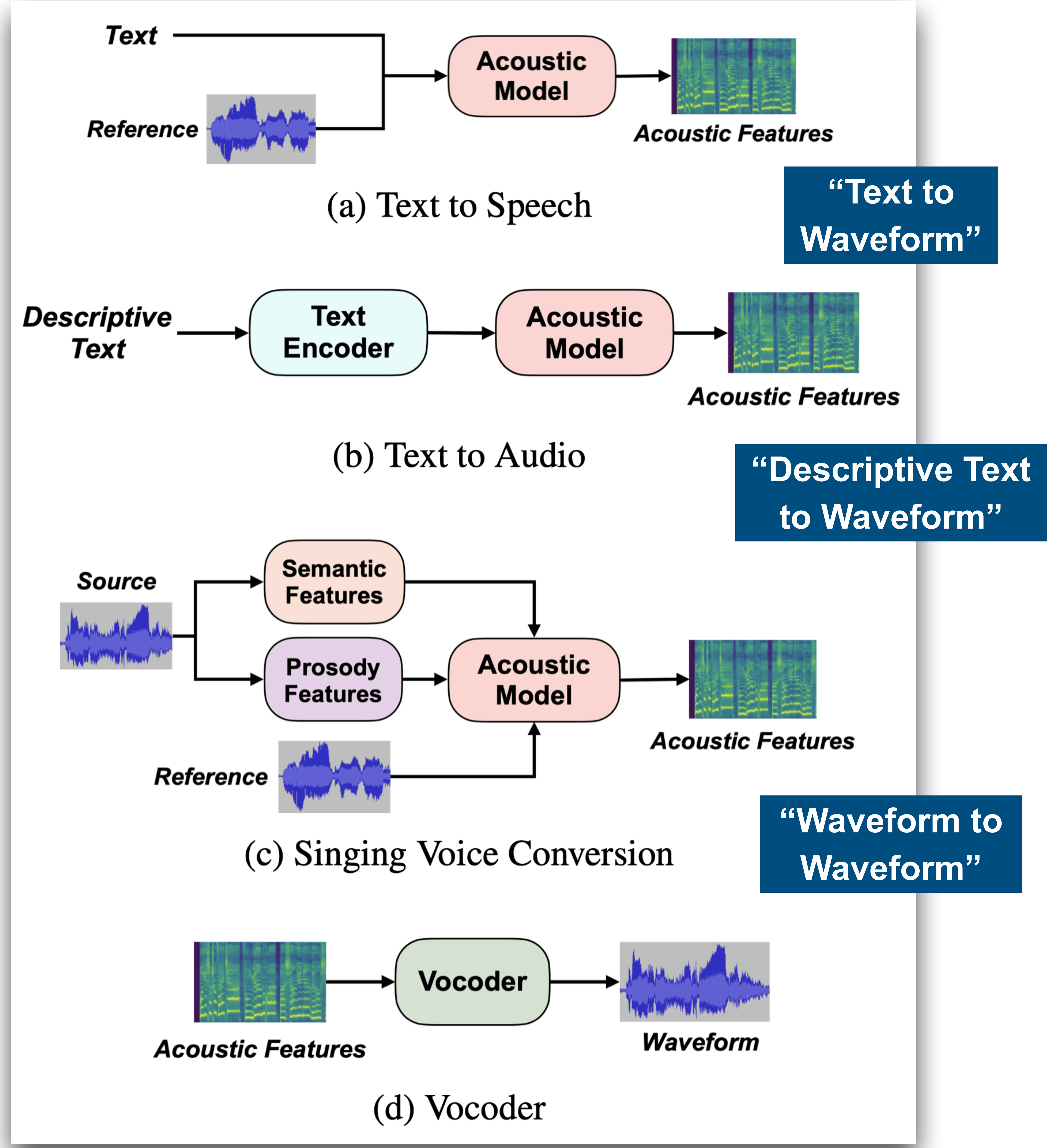
- **Highly positive comments from the market**

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

Strength1: Unified Audio Generation Framework





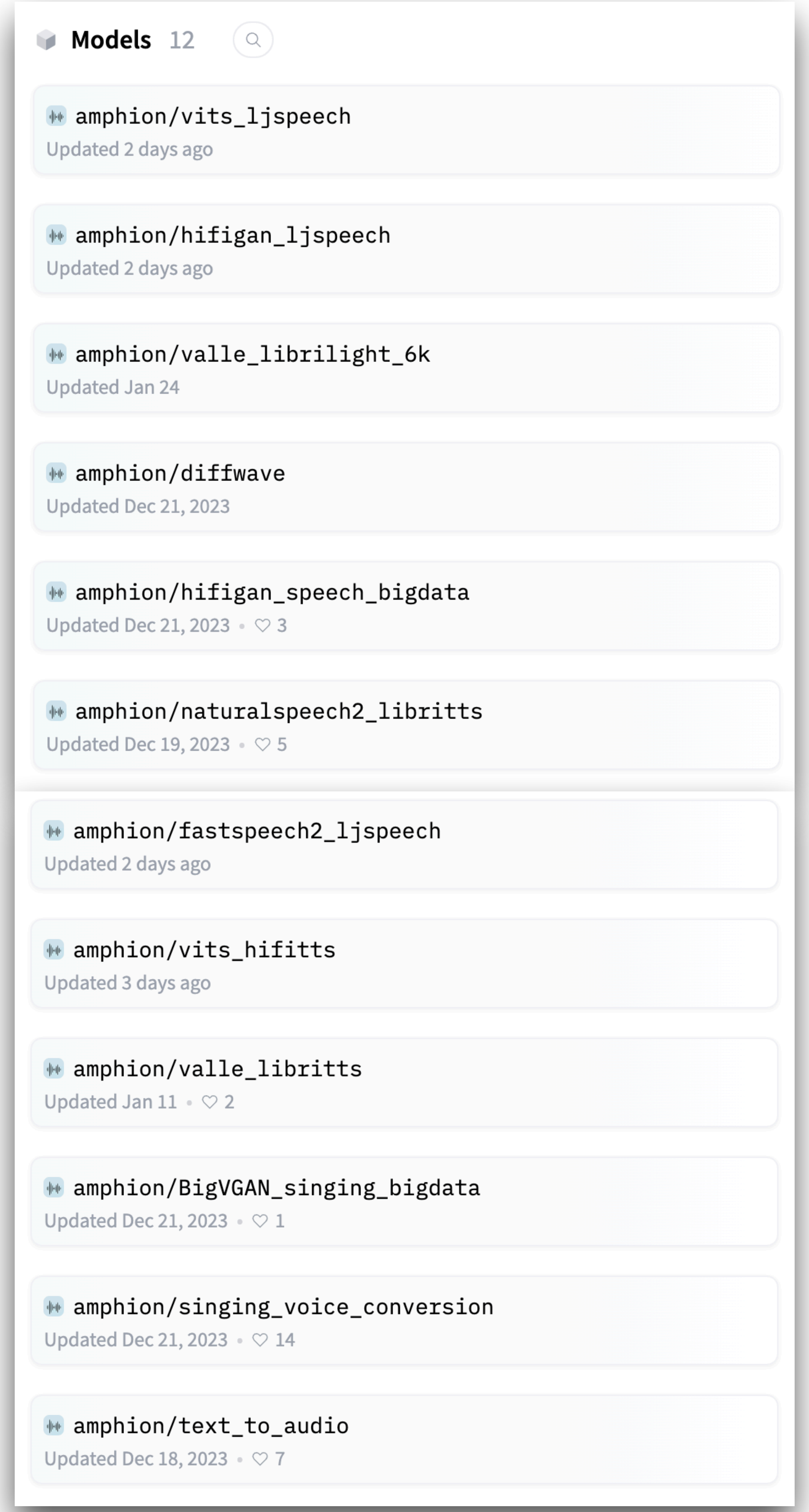
Strength2: Beginner-friendly End-to-End Workflow

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

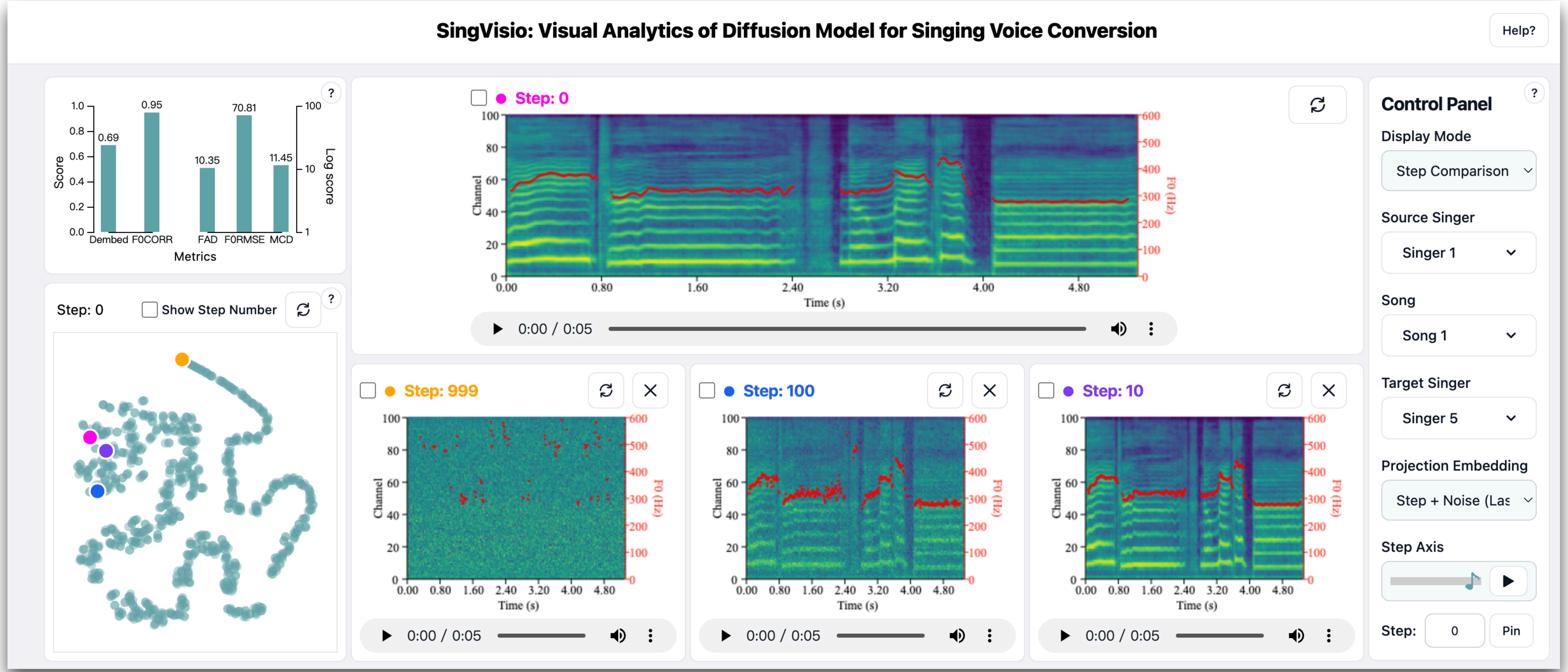
Strength3: Open Pre-trained Models

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.*

THANKS

Xueyao Zhang (张雪遥)

- **Second-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 **(3.5k stars)**

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

Amphion Offical 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

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