香港中文大學(深圳) | 数据科学学院
The Chinese University of Hong Kong, Shenzhen | School of Data Science

A Review: Singing Voice Conversion with Non-Parallel Data

Xueyao Zhang

Human Language Technology Lab (HLT), CUHK-SZ

Outline

- **Background**
- **Challenges**
- **Review of the Existing Works**
- **Our Future Work**

Outline

- **Background**
  - Problem definition
  - Applications and user scenarios
- **Challenges**
- **Review of the Existing Works**
- **Our Future Work**

Problem: Voice Conversion & Singing Voice Conversion

- **Voice Conversion (VC)**
  - VC is a technique to **modify speech waveform** to **convert non-/para-linguistic information** while **preserving linguistic information**. [1]
- **Singing Voice Conversion (SVC)**
  - SVC make it possible **for a singer** to sing a song with the **desired voice timbre** beyond their own physical constraints. [2]
  - SVC make it possible to **convert a source singer's** singing voice into **another target singer's** singing voice. [2]

[1] Tomoki Toda. **Recent Progress on Voice Conversion: What is Next?** 2021.
[2] Kazuhiro Kobayashi, Tomoki Toda, et al. **Statistical Singing Voice Conversion with Direct Waveform Modification based on the Spectrum Differential**. InterSpeech 2014.

Problem: Singing Voice Conversion with Non-Parallel Data

- Parallel data:

There exists the (source audio, desired audio) pair.

Converse the source singing voice to one containing more *chest resonance* for increasing your singing's power — Jiawei Li

- Non-parallel data:

Source Reference Conversion Result [1]

- Non-parallel cross domain data:

Source Reference Conversion Result [2]

[1] Chao Wang, et al. **Towards High-Fidelity Singing Voice Conversion with Acoustic Reference and Contrastive Predictive Coding.** InterSpeech 2022.
[2] Heyang Xue, et al. **Learn2Sing 2.0: Diffusion and Mutual Information-Based Target Speaker SVS by Learning from Singing Teacher.** InterSpeech 2022.

Application and User Scenarios

Imitation and Entertainment

Impression Show to various singers
— Taking 姐就是女王 as an example

Singing Voice Beautification

Tone Tuning

Application and User Scenarios

Creative Art

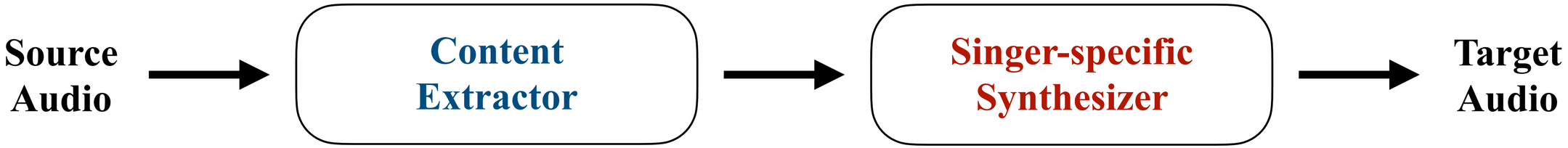
Music Montage (Michael Jackson feat. 曲比阿乌)

A novel morphing singing technique (merging *Pop* and *Folk*) of Jian Li.

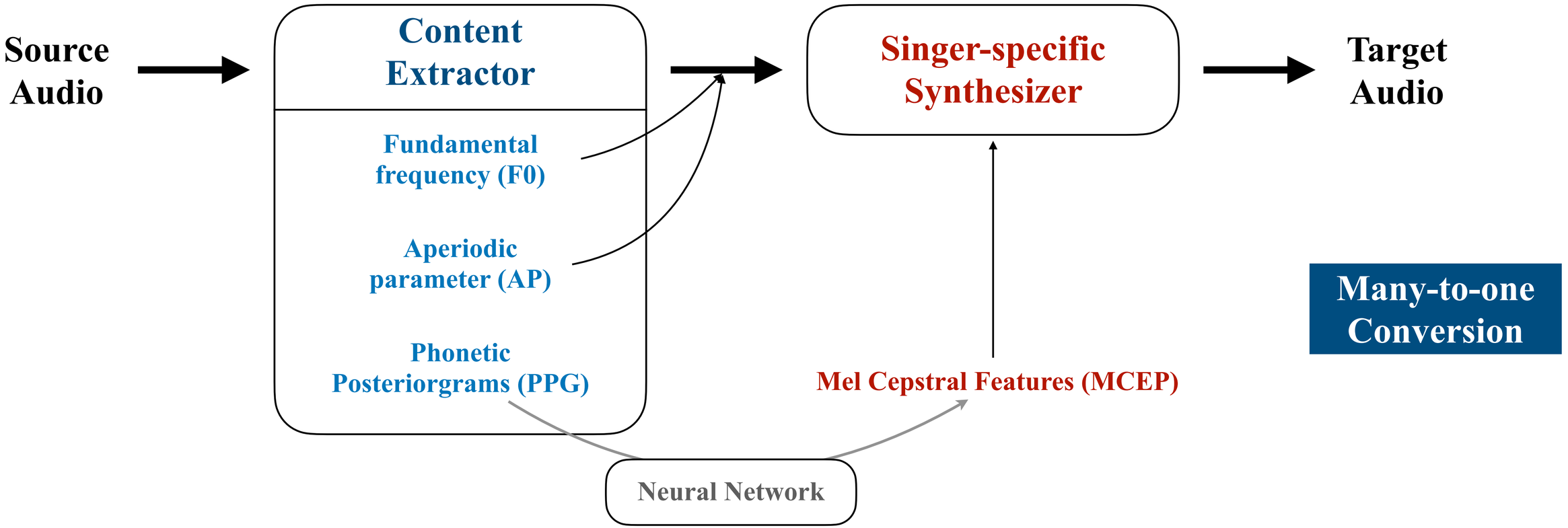
Outline

- **Background**
- **Challenges**
  - Paradigm of the conversion framework
  - Three main challenges
- **Review of the Existing Works**
- **Our Future Work**

Paradigm of the conversion framework

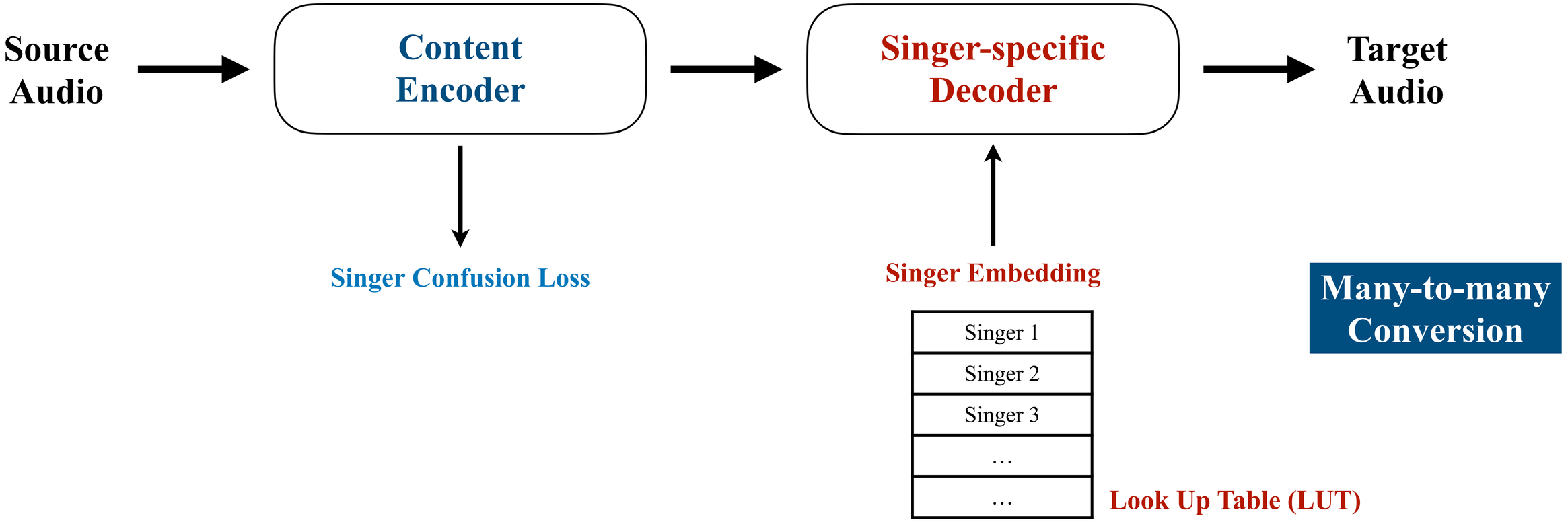


Paradigm of the conversion framework



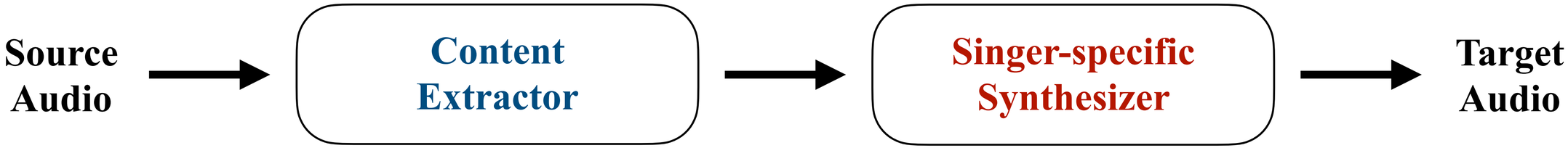
[1] Xin Chen, et al. **Singing Voice Conversion with Non-parallel Data**. IEEE MIPR 2019.
[2] Masanori Morise, et al. **WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications.** IEICE Trans. Inf. Syst. 2016

Paradigm of the conversion framework



Eliya Nachmani, et al. **Unsupervised Singing Voice Conversion**. InterSpeech 2019.

Three Main Challenges



Three main challenges:

① How to extract the **singer independent** features (i.e. **content info**)?

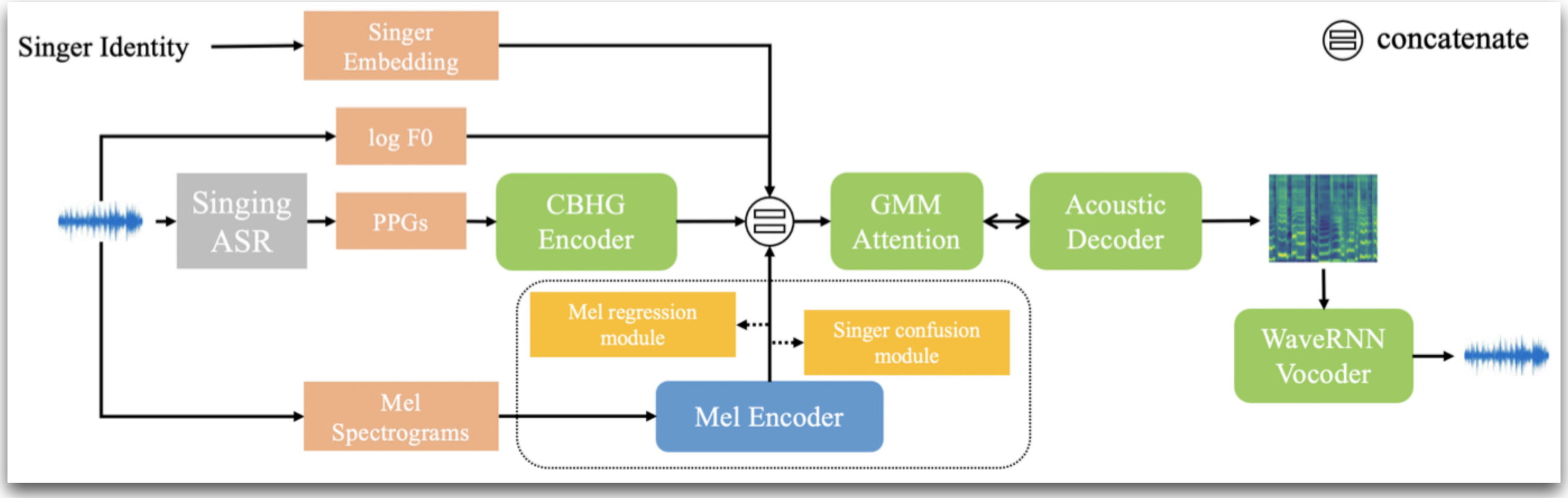
② How to model the **singer dependent** characteristics (i.e. **singer info**)?

③ How to make the general framework **special to singing voice** (i.e. to introduce **domain prior knowledge**)?

Outline

- **Background**
- **Challenges**
- **Review of the Existing Works**
  - **To model the singer independent features (3 papers chosen)**
  - To model the singer dependent characteristics (2 papers chosen)
  - To introduce the domain prior knowledge (3 papers chosen)
- **Our Future Work**

(1/3) Phonetic Posteriorgrams (PPG) as singer independent features



Content Info

① **PPG features**

② **Singer independent musical/acoustic info**

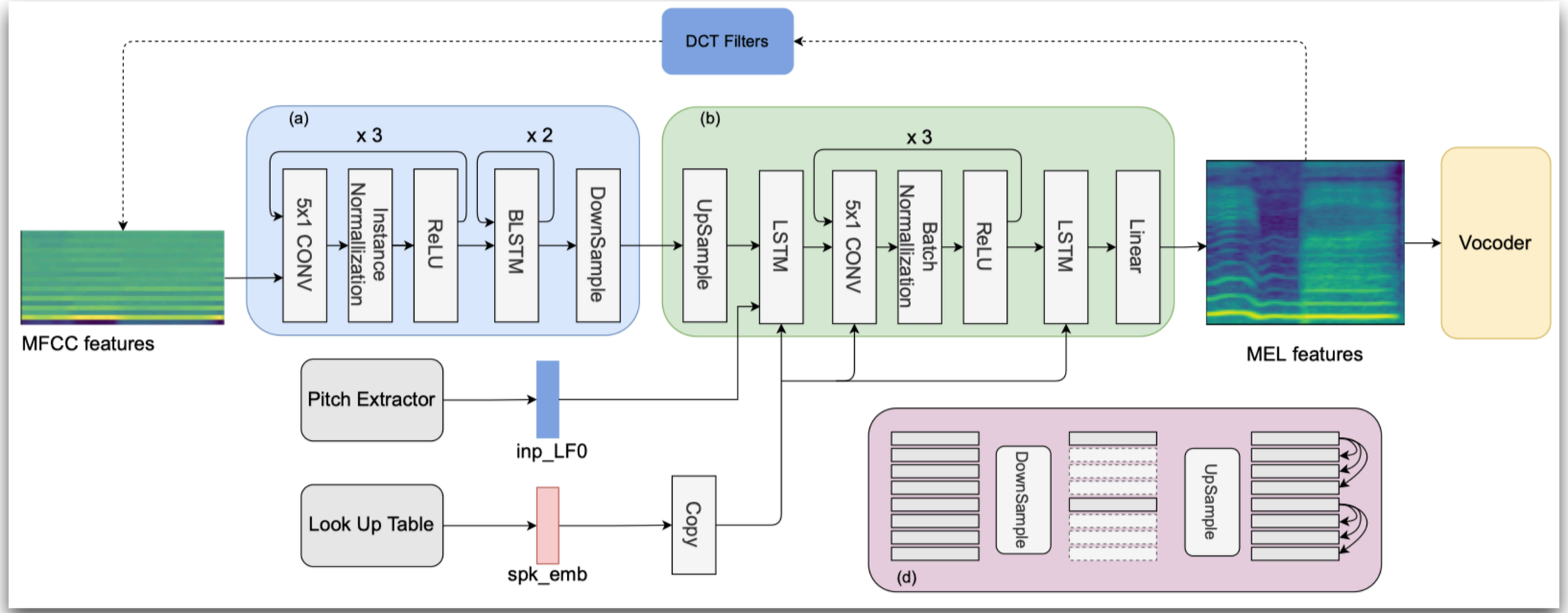
Singer Info **Singer Embedding**

Special to Singing Voice

Extract musical content from Mel Spectrograms

Zhonghao Li, et al. **PPG-Based Singing Voice Conversion with Adversarial Representation Learning**. ICASSP 2021.

(2/3) Low quefrencies of MFCC as singer independent features



Content Info

① **Low quefrencies of MFCC features (the first 20-dimension MFCCs)**

② **Pitch (F0) curve**

Singer Info

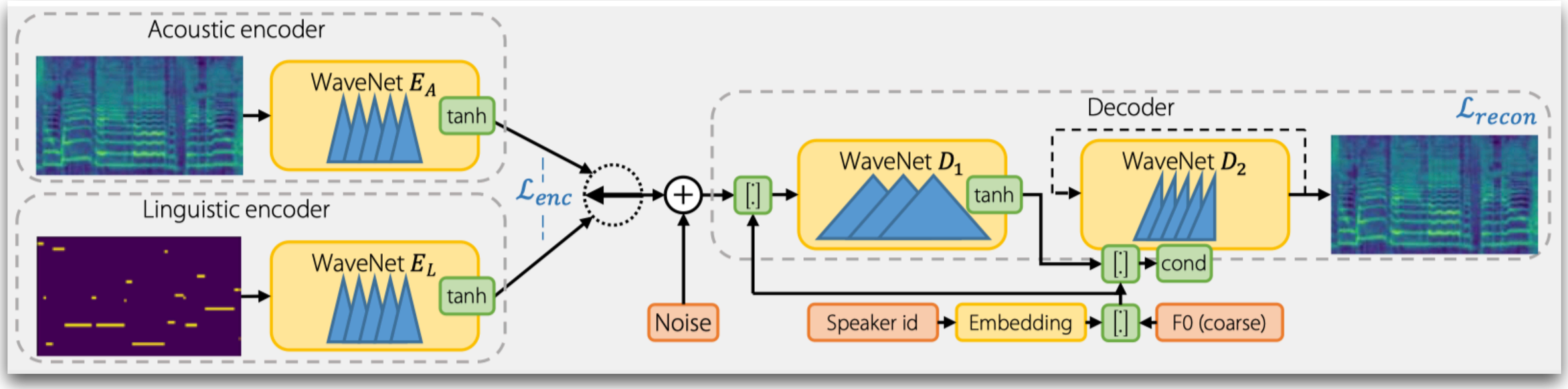
Singer Embedding

Special to Singing Voice

Low quefrencies of MFCC affects the linguistic info, while high quefrencies affects the F0 and the harmonics

Ying Zhang, et al. **K-Converter: An Unsupervised Singing Voice Conversion System**. ICASSP 2022.

(3/3) Learn singer independent Acoustic info from Linguistic info



Content Info

① **Representations that can be either from linguistic or acoustic info**

② **Pitch (F0) curve**

Singer Info

Singer Embedding

Special to Singing Voice

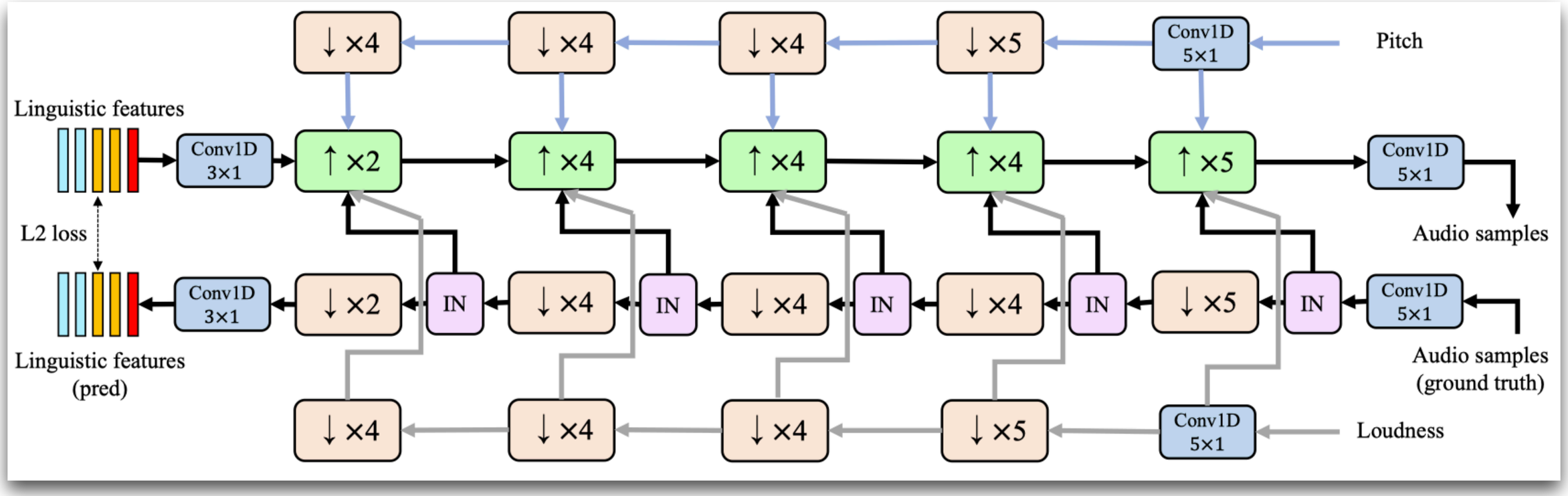
The paper gives an answer to "how to utilize the singing voice data which has well-aligned score with lyrics".

Jordi Bonada, et al. **Semi-supervised Learning for Singing Synthesis Timbre**. ICASSP 2021.

Outline

- **Background**
- **Challenges**
- **Review of the Existing Works**
  - To model the singer independent features (3 papers chosen)
  - **To model the singer dependent characteristics (2 papers chosen)**
  - To introduce the domain prior knowledge (3 papers chosen)
- **Our Future Work**

(1/2) IN/AdaIN for removing/capturing singer characteristics



Content Info

① **Representations after hierarchical Instance Normalization (IN)**

② **Pitch and Loudness**

Singer Info

Temporal statistics in every IN

Special to Singing Voice

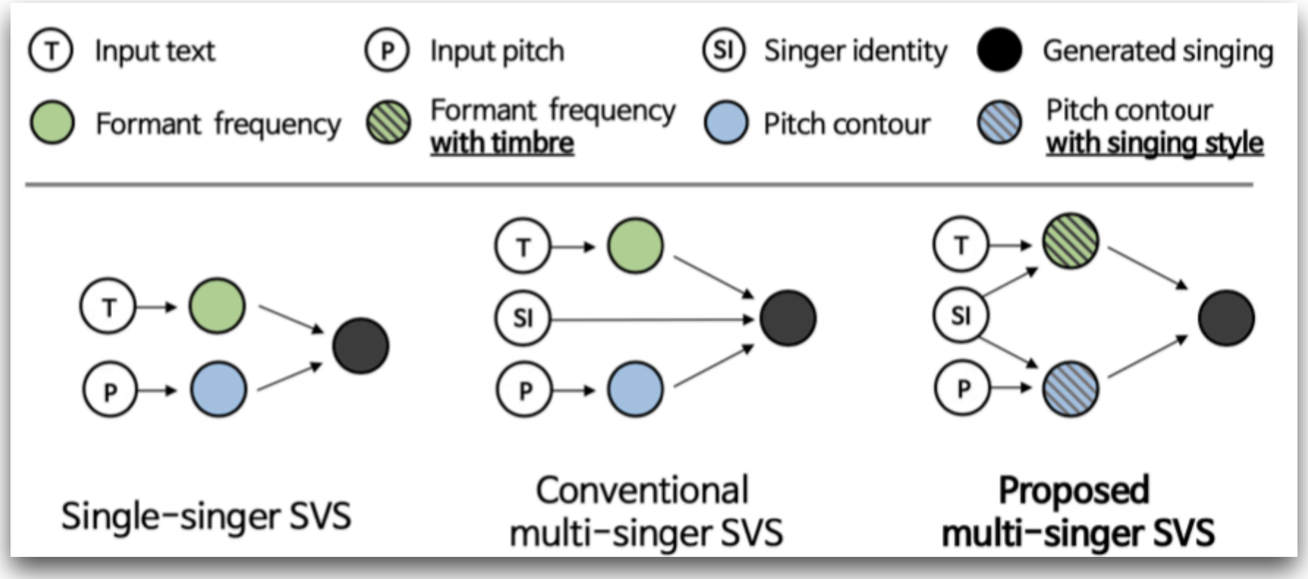
The hierarchical framework can capture fine-grained singer characteristics at different granularity.

Xu Li, et al. **A Hierarchical Speaker Representation Framework for One-shot Singing Voice Conversion**. InterSpeech 2022.

(2/2) Divide the singer characteristics into Timbre and Singing Style

Singer Info

Disentangled representations of timbre and singing style



Special to Singing Voice

The singer characteristics lies in two aspects:

① **"What to sing": special formant frequency;**

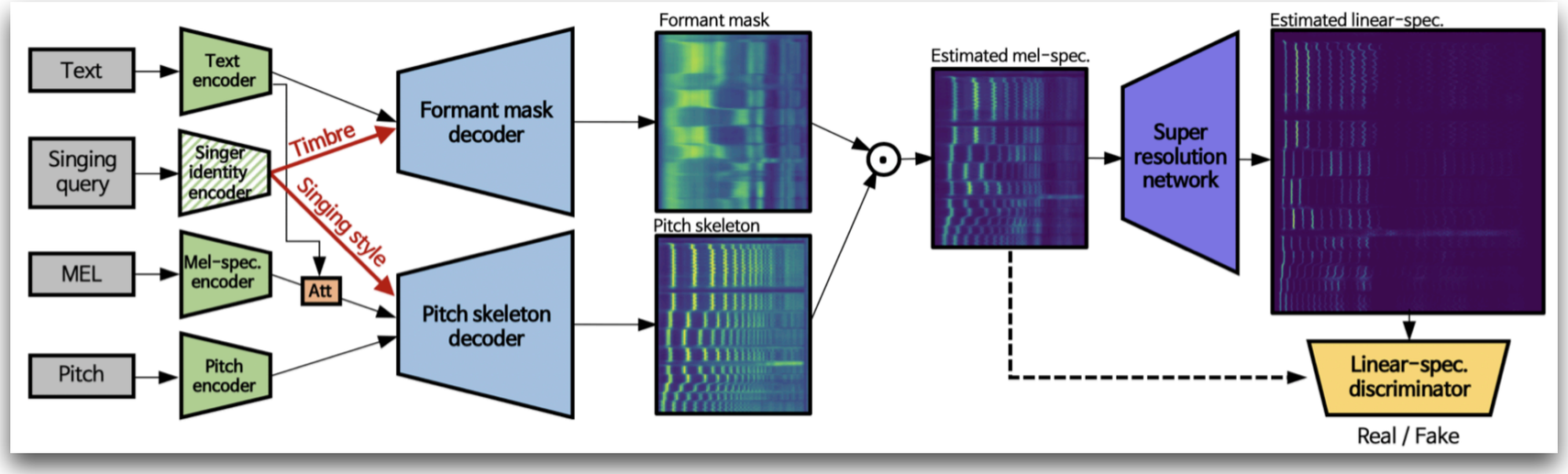
② **"How to sing": special singing pitch style**

Content Info

A given score:

① **pitch**

② **lyrics**

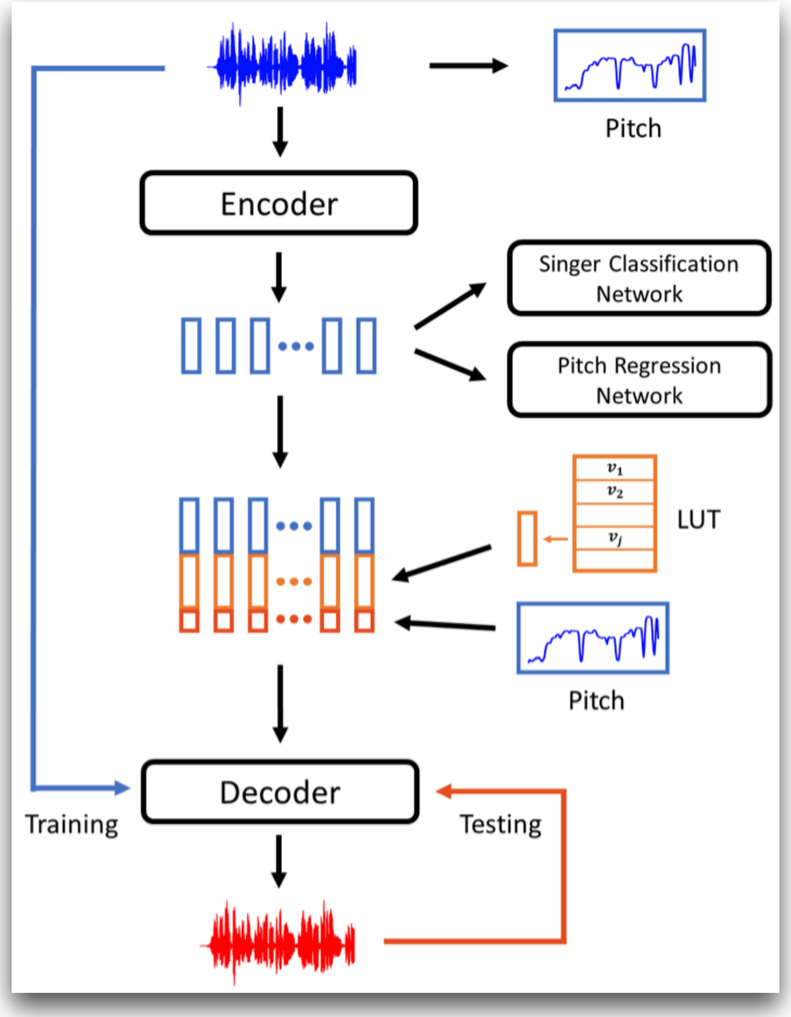


Juheon Lee et al. **Disentangling Timbre and Singing Style with Multi-Singer Singing Synthesis System**. ICASSP 2020.

Outline

- **Background**
- **Challenges**
- **Review of the Existing Works**
  - To model the singer independent features (3 papers chosen)
  - To model the singer dependent characteristics (2 papers chosen)
  - **To introduce the domain prior knowledge (3 papers chosen)**
- **Our Future Work**

(1/3) Enhance the modeling for Pitch



Content Info

① **Pitch curve**
② **Singer independent representations**

Singer Info

Singer Embedding

Special to Singing Voice

Pitch is a very strong feature for singing voice!

Source Baseline PitchNet

Chengqi Deng et al. **PitchNet: Unsupervised Singing Voice Conversion with Pith Adversarial Network**. ICASSP 2020.

(2/3) Enhance the modeling for Harmonic Signals

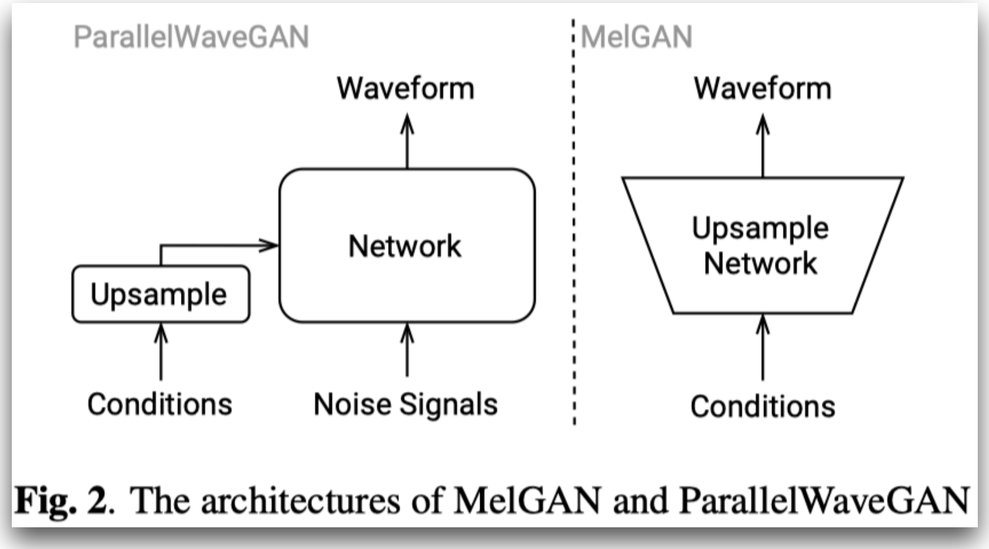


Fig. 2. The architectures of MelGAN and ParallelWaveGAN

Special to Singing Voice

Harmonic signals matters a lot for the smoothness and continuity of audio.

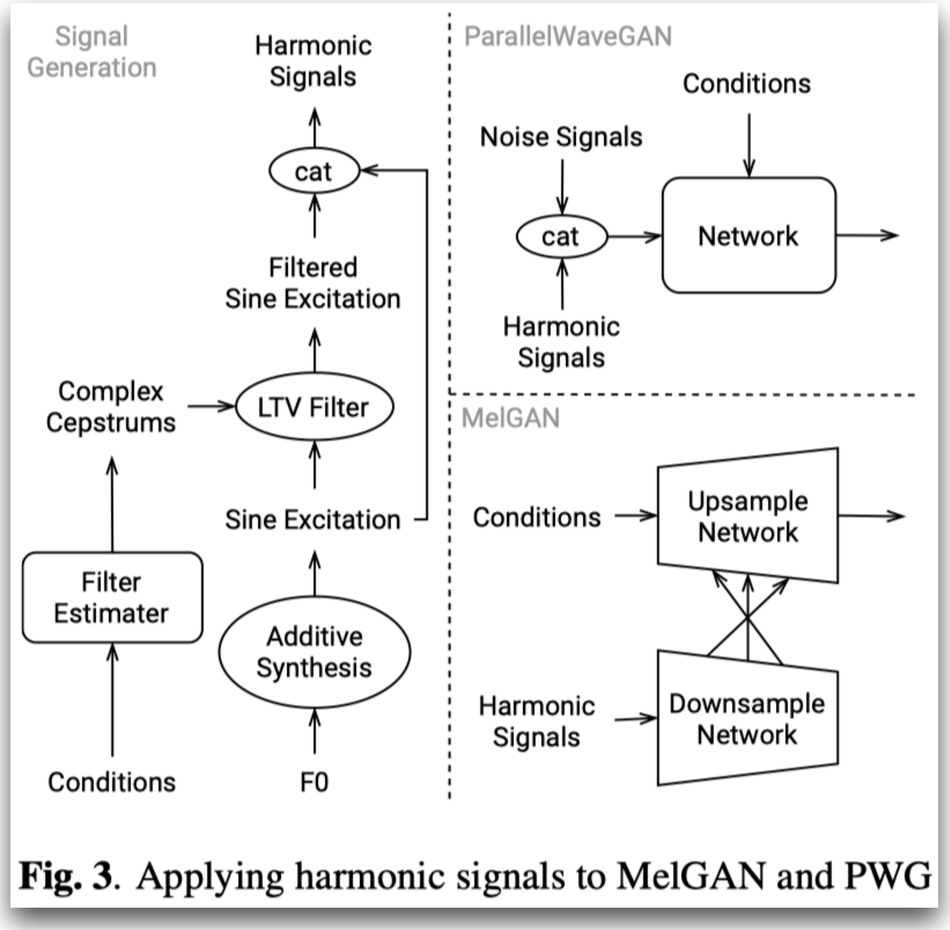
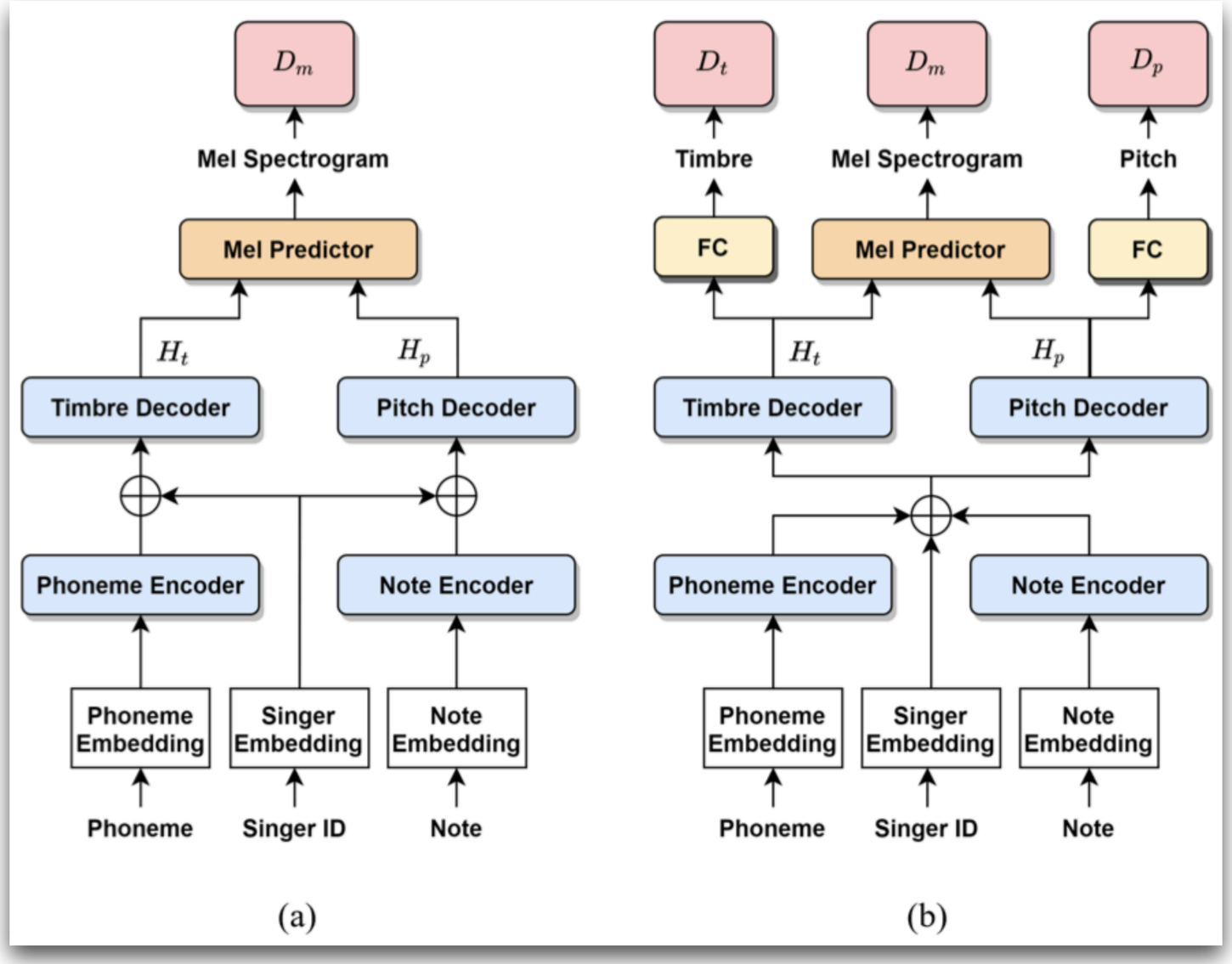


Fig. 3. Applying harmonic signals to MelGAN and PWG

MelGAN	MelGAN w/ Harmonic Signals	PWG	PWG w/ Harmonic Signals

Haohan Guo et al. **Improving Adversarial Waveform Generation Based Singing Voice Conversion with Harmonic Signals**. ICASSP 2022.

(3/3) Enhance the modeling for Timbre



Special to Singing Voice

Pitch and timbre are entangled closely.

Timbre features in this paper

① Mel-generalized Cepstrum

② Band Aperiodicity

③ Voiced/Unvoiced Flags

(which are extracted by WORLD)

Tae-Woo Kim et al. **Adversarial Multi-Task Learning for Disentangling Timbre and Pitch in Singing Voice Synthesis**. InterSpeech 2022.

Outline

- **Background**
- **Challenges**
- **Review of the Existing Works**
- **Our Future Work**
  - Promising directions
  - Our next step

Promising directions

- ✦ Well-organized evaluation
- ✦ Explorations for singer dependent characteristics
- ✦ More flexible and general conversion problems
- ✦ More sufficient modeling for music domain knowledge, such as:
  - Duration-Lyrics(-Pitch) alignment info
  - Music theory knowledge (for more flexible conversion)
  - Singing knowledge for different genres

Source | Conversion with a slightly altered score

An example of *Bel Canto*

香港中文大學(深圳) | 数据科学学院
The Chinese University of Hong Kong, Shenzhen | School of Data Science

THANKS