Singing Voice Conversion based on Target Waveform Mapping

Jiaying Li, Nat Condit-Schultz
Georgia Institute of Technology, Atlanta, GA, USA

Georgia Tech Center for Music Technology

INTRODUCTION

Singing voice conversion (SVC) is a technique that converts the source sound of singer A to singer B's target sound without changing the lyrics.

- Only modify singer-dependent features such as formant, intonation, intensity, and duration, while keeping singer-independent content such as melody and lyrics.
- In my research project, I only focus on changing the vocal timbres.
- Professional singers and music producers are able to use SVC techniques to produce music at a faster speed and at a lower cost.
- Amateurs are able to use SVC techniques to make their own songs.

COMPARE TO VOICE CONVERSION (VC)

Compare to VC, SVC techniques adapted to:

- Same length, same speed, and the same rhythm;
- Same pitch;
- Larger frequency range.

RELATED WORK

Data:

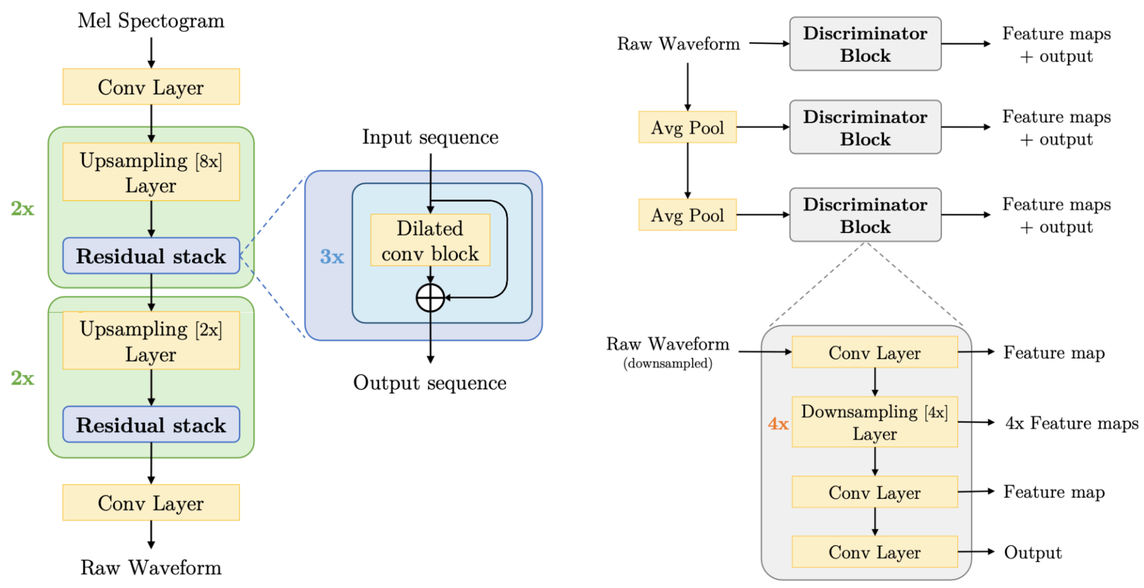
- Parallel Data: Same content, different speakers/singers;
  - **NUS-48 dataset** (collected by the National University of Singapore)
    - Folk songs and pop songs
    - 12 different singers
    - 4 songs sung by each singer
    - 108 minutes
- Non-Parallel Data: Different content, different speakers/singers.

Models:

- MelGAN
- WaveGlow
- WaveNet
- …
- Input MFCC, train in the model, and use vocoders to output waveform audio
- Estimation: Mean Opinion Score (MOS)

BASELINE MODEL: MELGAN

Model Architectures: MelGAN (Generator: Left; Discriminator: Right)



The **MelGAN model** uses Mel spectrogram features as input, gradually upsamples to the speech length, and adds convolution blocks between upsampling signals to calculate the transformation from the frequency domain to the time domain.

- The final output is the audio with a fixed number of frames.

- The entire upsampling process is used as the generator (Generator) part and embedded into the GAN framework for training The discriminator (Discriminator) and the objective function is adjusted according to the specific features of audio signals, making the training more stable and effective;
- Each upsampling layer is a transposed convolution with kernel size being twice the stride, which is the same as the upsampling ratio for the layer;
- Each residual dilated convolution stack has three layers, and the activation function is Leaky ReLu;
- Loss: STFT loss; Evaluation: Mel-cepstral distance, Mean Square Error.

OWN DATASET

- Find the original songs and covers from the *We Sing* website (3 octaves);
- Use *Spleeter* to separate the vocal parts;
- Label them with the singer ID and song ID.
  - 20 different songs, 270 minutes in total;
  - Each song is sung by 3 different singers (2 female singers and 1 male singer);

MELGAN RESULT

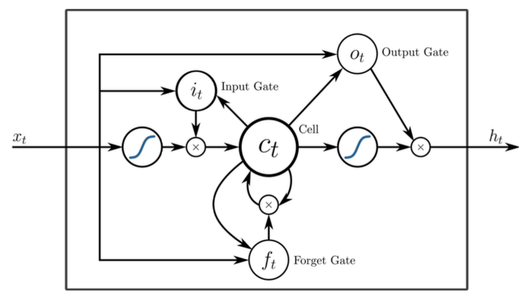
Objective Evaluation of Two Dataset

Dataset	Mel-cepstral Distance	RMSE
NUS-48 Dataset	7.3	3.49
New Dataset	5.2	1.23

- Trained on NUS-48 Dataset: Performs well in the speech frequency range, but not that well for large-frequency ranges.
- Trained on our own Dataset: Performs better.

WAVEFORM MAPPING

- Frame the singing voice audio into pitch blocks with a hop size;
  - Cut the audio by pitch class in each sentence;
    - e.g., In the same sentence, map “C3” sung by the male singer and “C4” sung by the female singer.
    - The minimum unit of a note: 1000 sample points per frame
    - 299 pieces of frames in a single song
    - 598 pieces of frames in total for a pair of parallel data
- Set up a network to train on parallel data
  - LSTM network
- Input: pitch (F0) and waveform frames;
- output: waveform frames
- Overlap-add the blocks



RESULTS

- The timbre maps well;
- The different fundamental frequency is kept;
- Constraint: only work for specific target conversion.

FUTURE WORK

- Implementing overlap-add
- The transition between notes and notes

CONTACT INFORMATION

Computational and Cognitive Musicology Lab, Georgia Tech Center for Music Technology
jli3269@gatech.edu
Dataset: https://github.com/JiayingLi0803/SingingVoiceConversion