Singing Expression Transfer from One Voice to Another for a Given Song

Korea Advanced Institute of Science and Technology
Sangeon Yong, Juhan Nam
Introduction
Introduction

diff source to target
Related Works

Antares Autotune 8 graphical mode

Steinberg Variaudio
Related Works

- Cano et al. (ICMC, 2000)
  - Voice morphing system with source and target voice
  - Score information is used for temporal alignment

- Nakano et al. (SMC, 2009)
  - Similar with above but using a singing synthesizer instead of the source voice (i.e. Vocaloid)
  - Tune synthesizer parameter with the lyric information of the song

However, they require additional score information!
Research Goal

Transfer musical expressions without any additional information
System Structure

Target Singing Voice

Feature Extraction

DTW

Smoothing

stretching ratio

Source Singing Voice

Time-Scale Modification

modified

stretched

Temporal Alignment

S

Target Singing Voice

Pitch Alignment

HPSS

Pitch Detector

harmonic signal

Pitch Shifting

Dynamics Alignment

Envelope Detector

gain ratio

Gain

Modified Singing Voice

smoothed stretching ratio
System Structure

Temporal Alignment
- Feature Extraction
- DTW
- Smoothing
  - stretching ratio
  - smoothed stretching ratio
- Time-Scale Modification

Pitch Alignment
- HPSS
- Pitch Detector
  - harmonic signal
  - pitch ratio
- Pitch Shifting
- Gain
  - gain ratio
- Envelope Detector

Dynamics Alignment

Target Singing Voice
Source Singing Voice
Modified Singing Voice

\[ s \rightarrow S \rightarrow S_T \rightarrow STP \rightarrow S_{TPE} \]
System Structure

Target Singing Voice

Source Singing Voice

Feature Extraction

DTW

Smoothing

stretching ratio

Temporal Alignment

Time-Scale Modification

smoothed stretching ratio

Pitch Alignment

HPSS

Pitch Detector

harmonic signal

Dynamics Alignment

Pitch Shifting

Gain

Envelope Detector

gain ratio

Modified Singing Voice

\[ s \]

\[ s_t \]

\[ s_{tp} \]

\[ s_{tpe} \]
Temporal Alignment

Singer A

Lyrics

Let it go

let it go

Singer B
Temporal Alignment – Dynamic Time Warping
System Structure

Target Singing Voice

Source Singing Voice

Temporal Alignment

Feature Extraction

DTW

Smoothing

stretching ratio

smoothed stretching ratio

Time-Scale Modification

Source Singing Voice

S

Pitch Alignment

HPSS

Pitch Detector

harmonic signal

pitch ratio

Dynamics Alignment

Envelope Detector

gain ratio

Gain

Modified Singing Voice

S_{TPE}

\( s \)

\( s_{T} \)

\( s_{TP} \)
Temporal Alignment – Feature Extraction

Spectrogram of Source

Spectrogram of Target
Temporal Alignment – Feature Extraction

Similarity matrix with spectrogram
Temporal Alignment – Feature Extraction

Spectrogram of Source

Spectrogram of Target
Feature Extraction Strategy

- Preserving common elements
  - Note-level melody
  - Lyrics

- Suppressing different characteristics
  - Vibrato or other pitch-related articulations
  - Singer timbre
Proposed Features

- Max-filtered Constant-Q transform
  - Semi-tone pitch resolution: vibrato with less than one semi-tone
  - Frequency-wise max-filtering: vibrato with more than one semi-tone
Proposed Features

- Phoneme score (phoneme classifier posteriorgram)
  - Frame-level features for accurate temporal alignment
  - Singer invariant lyrical features
Temporal Alignment – Feature Comparison

Spectrogram

Max-filtered Constant-Q Transform
Temporal Alignment – Feature Comparison

Spectrogram

phoneme score
Temporal Alignment – Feature Comparison

Spectrogram

Phoneme Score + Const-Q Trans
System Structure

Singing Voice

Feature Extraction

DTW

Smoothing

Stretching ratio

HPSS

Pitch Detector

Harmonic signal

Envelope Detector

Gain

Modified Singing Voice

Source Singing Voice

Time-Scale Modification

Temporal Alignment

Pitch Alignment

Dynamics Alignment

\[ S \rightarrow \text{Feature Extraction} \rightarrow \text{DTW} \rightarrow \text{Smoothing} \rightarrow \text{HPSS} \rightarrow \text{Pitch Detector} \rightarrow \text{Envelope Detector} \rightarrow \text{Gain} \rightarrow S_{TPE} \]

\[ S_{T} \rightarrow \text{Pitch Shifting} \rightarrow S_{TP} \]

\[ S \rightarrow \text{smoothed stretching ratio} \rightarrow \text{smoothing} \rightarrow \text{stretching ratio} \rightarrow \text{Dynamics Alignment} \]

\[ s_{s} \rightarrow \text{Temporal Alignment} \rightarrow s_{t} \rightarrow \text{Pitch Alignment} \rightarrow s_{tp} \rightarrow \text{Dynamics Alignment} \rightarrow s_{tp} \]

\[ s_{s} \rightarrow \text{Source Singing Voice} \rightarrow \text{Feature Extraction} \rightarrow \text{DTW} \rightarrow \text{Smoothing} \rightarrow \text{HPSS} \rightarrow \text{Pitch Detector} \rightarrow \text{Envelope Detector} \rightarrow \text{Gain} \rightarrow S_{TPE} \]
Temporal Alignment – Path Smoothing
Temporal Alignment – Path Smoothing
Temporal Alignment – Path Smoothing
**System Structure**

1. **Target Singing Voice**
2. **Source Singing Voice**
3. **Temporal Alignment**
   - Feature Extraction
   - DTW (stretching ratio)
   - Smoothing (smoothed stretching ratio)
4. **Pitch Alignment**
   - HPSS (harmonic signal)
   - Pitch Detector (pitch ratio)
5. **Dynamics Alignment**
   - Envelope Detector
   - Gain (gain ratio)
6. **WSOLA**

- **Modified Singing Voice**
System Structure

Temporal Alignment
- Feature Extraction
- DTW
- Smoothing
  - stretching ratio
  - smoothed stretching ratio

Time-Scale Modification

Pitch Alignment
- HPSS
- Pitch Detector
  - harmonic signal
  - pitch ratio
- Pitch Shifting

Dynamics Alignment
- Envelope Detector
- Gain
  - gain ratio

Modified Singing Voice

Source Singing Voice $S$

Target Singing Voice $S_T$

Modifying $S_T$ with $S$ results in $S_{TP}$

$S_{TPE}$ is the final modified singing voice.
Pitch Alignment

- Harmonic-Percussion Source Separation (HPSS)
  - Pre-processing of pitch detection to increase detection accuracy

- Pitch Detector
  - YIN

- Pitch shifting
  - Pitch-Synchronous Overlap-Add (PSOLA)
  - Formant preservation
Pitch Alignment

source

![Graph showing frequency distribution for source frames.]

target

![Graph showing frequency distribution for target frames.]

result

![Graph showing frequency distribution for result frames.]

Source Frame [32 samples]

Target Frame [32 samples]

Result Frame [32 samples]
System Structure

Target Singing Voice -> Feature Extraction -> DTW -> Smoothing

Source Singing Voice -> Time-Scale Modification

Temporal Alignment

Pitch Alignment

Dynamics Alignment

HPSS -> Pitch Detector

Envelope Detector

Gain

Modified Singing Voice

$S$ -> $S_T$ -> $S_{TP}$ -> $S_{TP_E}$

Stretching ratio

Smoothed stretching ratio

Harmonic signal

Pitch ratio

Gain ratio
Dynamics Alignment

source

RMS Value of the Source Singing Voice

Amplitude

Time [Seconds]

target

RMS Value of the Target Singing Voice

Amplitude

Time [Seconds]

result

RMS Value of the Modified Source Singing Voice

Amplitude

Time [Seconds]
Evaluation

Datasets
- 4 recordings for each of 4 songs (total 16 recordings)
- One of 4 recordings is a target singing voice (professional or skilled)
- Totally 12 pairs of source-target singing voice

<table>
<thead>
<tr>
<th>Song 1</th>
<th>Song 2</th>
<th>Song 3</th>
<th>Song 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>female</td>
<td>male</td>
<td>male</td>
</tr>
<tr>
<td>No. of source</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Remarks</td>
<td>high pitch English</td>
<td>low pitch English</td>
<td>swing rhythm Korean</td>
</tr>
</tbody>
</table>
Evaluation

- **Temporal alignment**
  - Better alignment has less fluctuation of the DTW slope
  - Standard deviation of slope angle $\theta = \arctan(slope)$
Evaluation

- Pitch alignment

<table>
<thead>
<tr>
<th>Gender</th>
<th>Song 1</th>
<th>Song 2</th>
<th>Song 3</th>
<th>Song 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of source</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Remarks</td>
<td>high pitch English</td>
<td>low pitch English</td>
<td>swing rhythm Korean</td>
<td>swing rhythm Korean</td>
</tr>
</tbody>
</table>

![Pitch Difference with Target (semitones) Chart]

- Chart: Comparison of pitch alignment across songs.
Evaluation

- Dynamics alignment

<table>
<thead>
<tr>
<th>Song 1</th>
<th>Song 2</th>
<th>Song 3</th>
<th>Song 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>female</td>
<td>male</td>
<td>male</td>
</tr>
<tr>
<td>No. of source</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Remarks</td>
<td>high pitch English</td>
<td>low pitch English</td>
<td>swing rhythm Korean</td>
</tr>
</tbody>
</table>
## Audio Examples

<table>
<thead>
<tr>
<th>source</th>
<th>target</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>let it go</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cherry blossom ending</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

More examples are available on [https://seyong92.github.io/ICASSP2018](https://seyong92.github.io/ICASSP2018)
Summary

- Proposed a method to transfer vocal expressions from one voice to another in terms of tempo, pitch and dynamics without any additional information.

- Showed the proposed method effectively transformed the source voices so that they mimic singing skills from the target voice.
Future Plan

- The limitation of this work is that the target voice must be available

- A possible solution is to model a target singer model (e.g. singing synthesizer with natural expressions) and generate a target example using melody and lyrics information extracted from the source voice

- Improve the audio quality using other time-scale/pitch modification algorithms
Thank you