PolySinger: Singing-Voice to Singing-Voice Translation from English to Japanese

Silas Antonisen and Iván Lopéz-Espejo

25th International Society for Music Information Retrieval (ISMIR)
University of Granada

santon@ugr.es

Thursday 24th October, 2024

Signal Processing, Multimedia Transmission
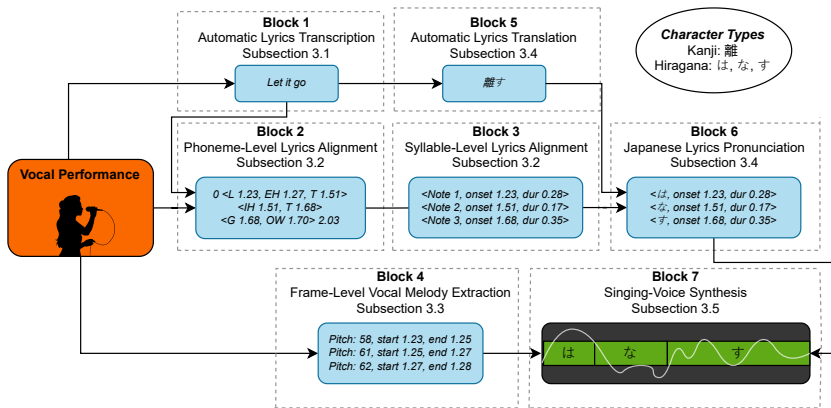and Speech/Audio Technologies

ugr | Universidad
de Granada

Introduction

- Modern **singing voice synthesis** (**SVS**) voicebanks can sing cross-lingually
- If a musician doesn't speak the language, they can't write lyrics in the language!
- Speech has more focus and greater performance than singing applications
- We present the first complete solution for **Singing-Voice to Singing-Voice Translation** (**SV2SVT**)
  1. Sing a song
  2. Translate the song
  3. Synthesize the song

Proposed Method

Implementation of English to Japanese **SV2SVT**

Vocal Performance

Block 1
Automatic Lyrics Transcription
Subsection 3.1

Let it go

Block 5
Automatic Lyrics Translation
Subsection 3.4

離す

Character Types
Kanji: 離
Hiragana: は, な, す

Block 2
Phoneme-Level Lyrics Alignment
Subsection 3.2

0 <L 1.23, EH 1.27, T 1.51>
<IH 1.51, T 1.68>
<G 1.68, OW 1.70> 2.03

Block 3
Syllable-Level Lyrics Alignment
Subsection 3.2

<Note 1, onset 1.23, dur 0.28>
<Note 2, onset 1.51, dur 0.17>
<Note 3, onset 1.68, dur 0.35>

Block 6
Japanese Lyrics Pronunciation
Subsection 3.4

<は, onset 1.23, dur 0.28>
<な, onset 1.51, dur 0.17>
<す, onset 1.68, dur 0.35>

Block 4
Frame-Level Vocal Melody Extraction
Subsection 3.3

Pitch: 58, start 1.23, end 1.25
Pitch: 61, start 1.25, end 1.27
Pitch: 62, start 1.27, end 1.28

Block 7
Singing-Voice Synthesis
Subsection 3.5

は　な　す

Post processing of translated lyrics in **Block 5** *and* **Block 6** *are subject to change depending on the language!*

Mean Opinion Score (MOS) Test

- 6 native Japanese participants
- Tested with both baseline and a fine-tuned translation system
- *Examples can be found at*
  https://silasantonisen.github.io/polysinger/
- 5 point scale

ID	Question	Baseline	Fine-tuned
Q1	How much sense do the lyrics make?	2.53 ± 0.49	2.17 ± 0.46
Q2	How natural is the Japanese used in the lyrics?	2.57 ± 0.48	2.30 ± 0.48
Q3	How well is the meaning of the original lyrics preserved?	2.47 ± 0.44	2.10 ± 0.44
Q4	How singable are the generated lyrics?	2.40 ± 0.41	2.23 ± 0.44
Q5	How well are the lyrics and melody aligned?	2.50 ± 0.52	2.10 ± 0.40
Q6	What is the overall quality of the generated Japanese singing?	2.33 ± 0.45	2.13 ± 0.41

- High variance in scores
- No statistically significant differences
- Baseline methodology and framework for **SV2SVT**
- Further research needed for robust **SV2SVT**