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

SIGMAT
Signal Processing, Multimedia Transmission and Speech/Audio Technologies

SILAS ANTONISEN AND IVÁN LOPÉZ-ESPEJO

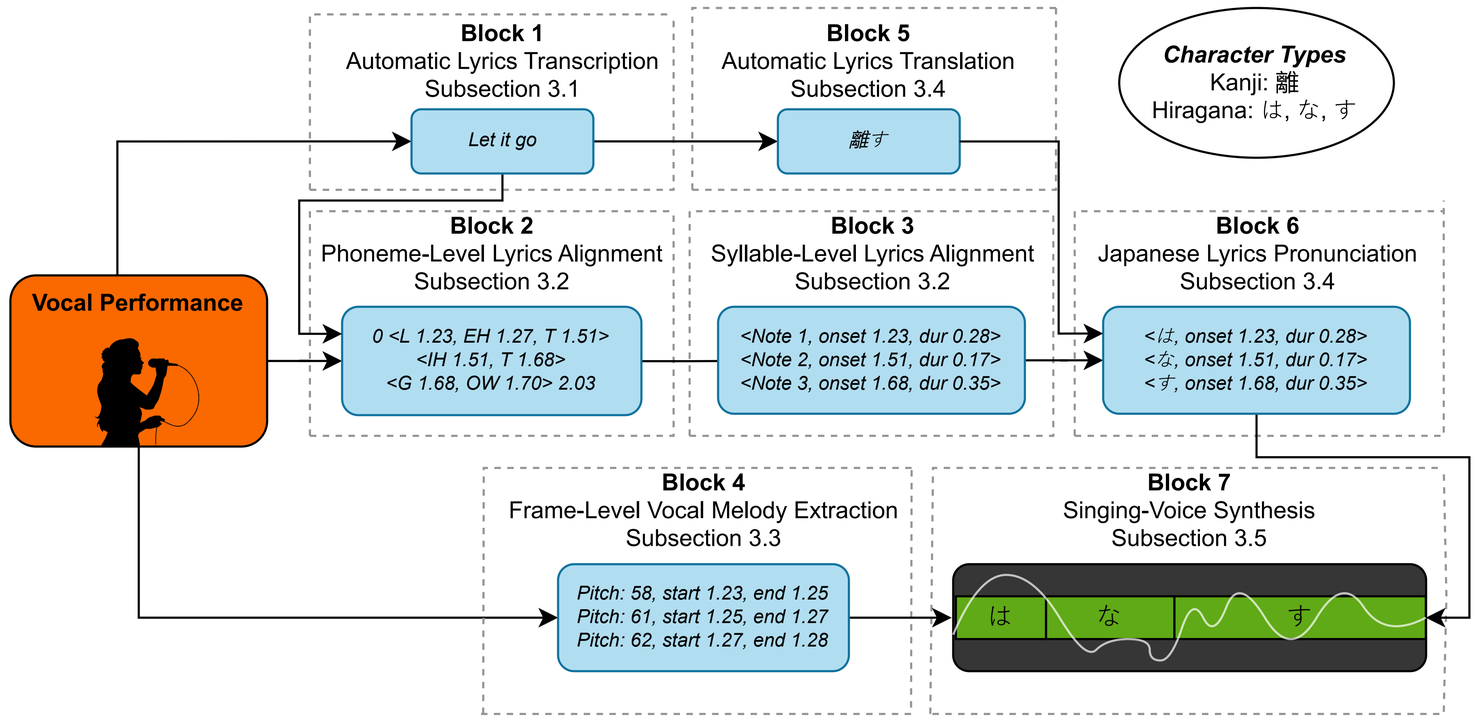
DEPT. OF SIGNAL THEORY, TELEMATICS AND COMMUNICATIONS,
UNIVERSITY OF GRANADA, SPAIN

UNIVERSIDAD DE GRANADA

INTRODUCTION

- Modern singing voice synthesis *(SVS)* voice-banks can sing cross-lingually.
- SVS and music generation applications have no feature for automatic song translation.
- Potential of cross-lingual SVS is untapped without automatic song translation.
- Vocal and lingual processing methods have more focus and greater performance in speech than singing.
- No ***complete system*** for transcription, singable translation and synthesis faithful to the original song exists.
- We present the first singing-voice to singing-voice translation ***(SV2SVT)*** system.

PROPOSED METHOD



PIPELINE BREAKDOWN

Record English vocal performance.

- **Block 1:** Transcribe lyrics.
- **Block 2:** Align phonetic sequence of the transcribed lyrics to audio.
- **Block 3:** Categorize phoneme sequences into syllables with onsets and durations.
- **Block 4:** Frame-level vocal-contour transcription.
- **Block 5:** Generate a list of possible Japanese translations.
- **Block 6:**
  - Break down all kanji in all translated sequences into their hiragana-form pronunciations and separate the words (Japanese uses no whitespace).
  - Choose the translated sequence with the most appropriate amount of syllables as best translation.
  - Align Japanese syllables (mora) with prior onsets and durations of English syllables.
- **Block 7:** Define notes in SVS engine with Japanese syllables bounded by onsets and durations. Automate the pitch values of notes with vocal-contour.

EXPERIMENTS

Method	Module
`Whisper-Large-V3`	Lyrics transcription *(Block 1)*
`Schufo lyrics-aligner`	Phoneme-level lyrics alignment *(Block 2)*
`CMU Pronunciation Dictionary`	Defining syllables *(Block 3)*
`Omnizart Vocal-Contour`	Vocal-contour transcription *(Block 4)*
`nllb-200-distilled-600M`	Baseline model for translation *(Block 5)*
`nllb-200-distilled-600M (fine-tuned)`	Fine-tuned model for translation *(Block 5)*
`pyKAKASI`	Convert kanji to hiragana *(Block 6)*
`Nagisa`	Word separation *(Block 6)*
`Synthesizer V`	SVS *(Block 7)*

- No standard evaluation metric for SV2SVT or song translation exists.
- Mean opinion score test of 6 questions with 6 native Japanese speakers, testing both the baseline and fine-tuned translation models. Examples used in evaluation can be found at `silasantonisen.github.io/polysinger`.

1 = *Very Poor*, 2 = *Poor*, 3 = *Neutral*, 4 = *Good*, and 5 = *Very Good*.

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 opinion scores.
- No statistically significant differences between the baseline and fine-tuned translation systems.
- Most recurring comments from participants revolved around quality and naturalness of the Japanese lyrics, e.g., incorrect pronunciations or word separations, not necessarily the translation.

CONCLUSION

- We have presented the first **SV2SVT** system in the literature.
- Our method facilitates a baseline framework for future development in English to Japanese **SV2SVT** as well as a baseline methodology for other language combinations.
- Further research is required for robust **SV2SVT**. Alternative technologies and methods must be tested, and robust tools must be developed. This will primarily be our future work.

CONTACT INFORMATION

E-mail: **santon@ugr.es**
Signal Processing, Multimedia Transmission and Speech/Audio Technologies Group, **SigMAT**.
Dept. of Signal Theory, Telematics and Communications, University of Granada, Spain.