POLYSINGER: SINGING-VOICE TO SINGING-VOICE TRANSLATION FROM ENGLISH TO JAPANESE

Silas Antonisen, Iván López-Espejo

Department of Signal Theory, Telematics and Communications, University of Granada, Spain

`{santon,iloes}@ugr.es`

ABSTRACT

The speech domain prevails in the spotlight for several natural language processing (NLP) tasks while the singing domain remains less explored. The culmination of NLP is the speech-to-speech translation (S2ST) task, referring to translation and synthesis of human speech. A disparity between S2ST and the possible adaptation to the singing domain, which we describe as singing-voice to singing-voice translation (SV2SVT), is becoming prominent as the former is progressing ever faster, while the latter is at a standstill. Singing-voice synthesis systems are overcoming the barrier of multi-lingual synthesis, despite limited attention has been paid to multi-lingual songwriting and song translation. This paper endeavors to determine what is required for successful SV2SVT and proposes PolySinger (**Poly**glot **Singer**): the first system for SV2SVT, performing lyrics translation from English to Japanese. A cascaded approach is proposed to establish a framework with a high degree of control which can potentially diminish the disparity between SV2SVT and S2ST. The performance of PolySinger is evaluated by a mean opinion score test with native Japanese speakers. Results and in-depth discussions with test subjects suggest a solid foundation for SV2SVT, but several shortcomings must be overcome, which are discussed for the future of SV2SVT.

1. INTRODUCTION

Speech-to-speech translation (S2ST) is a method for translating human speech into another language using synthetic speech. To do this, the conventional approach is to concatenate technologies that process separate parts of human speech into a complete system, where the cornerstones are speech recognition, machine translation and speech synthesis [1–3]. Although the use of end-to-end (E2E) solutions for S2ST has been studied thanks to the emergence of seq-to-seq models [4–6], neither E2E nor cascaded solutions have been attempted in the singing domain.

Singing-voice synthesis (SVS) systems have in recent years become very capable of human-like singing [7,8] and have even accomplished multi-lingual synthesis [9]. However, while the synthetic voice can sing cross-lingually, the songwriter might not be able to write cross-lingually.

Lyrics translation is a complex task which strives for inter-cultural comprehension of what makes a song suitable for singing. Prose translation, also called direct translation [10–12], differs greatly in its application from poetry and lyrics translation, as prose translation does not respect rules regarding rhythm and rhyme [13,14]. A few attempts at automatic lyrics translation have been made by transforming standard music notation from one language to another, which shows promising results [15, 16]. However, the necessity for standard music notation becomes a glaring restriction. From the perspective of a songwriter with interest in writing foreign-language lyrics, the creation of standard music notation is a labor-intensive task begging for automation. Therefore, to overcome the present limitations in adapting S2ST methods to the singing domain, we propose PolySinger: the first system for singing-voice to singing-voice translation (SV2SVT). PolySinger is a concatenated system of music information retrieval (MIR) technologies with the goal of directly translating a vocal performance in a source language into a synthetic vocal performance in a target language. PolySinger is made publicly available [1].

Automatic recognition of note-level events in a vocal melody is a complex and vaguely defined task [17]. Nonetheless, standard music notation is required for lyrics translation, and as such, this work proposes a simple yet effective approach to defining note-level events by assistance from syllable alignment.

State-of-the-art (SOTA) in the following technologies are structured into a complete SV2SVT system for PolySinger: *1)* automatic lyrics transcription, *2)* phoneme-level lyrics alignment, *3)* frame-level vocal melody extraction, *4)* automatic lyrics translation, and *5)* singing-voice synthesis. PolySinger is proposed as a concatenated solution instead of E2E to represent a modular framework facilitating research in SV2SVT. In this paper, PolySinger is presented for English to Japanese SV2SVT, which, to the best of our knowledge, also constitutes the first attempt at automatic lyrics translation from English to Japanese.

A series of native Japanese speakers participated in a mean opinion score (MOS) test to evaluate the perceptual quality of PolySinger for English to Japanese SV2SVT. Results show a promising fundamental structure

---

[1] `https://silasantonisen.github.io/polysinger/`

for SV2SVT, but also that our translated Japanese lyrics have not yet reached ideal naturalness.

2. RELATED WORK

Convolutional neural networks and expanded pronunciation dictionaries have been used for automatic lyrics transcription in monophonic recordings [18], along with time-delay neural networks in polyphonic recordings [19]. The latest reported SOTA in automatic lyrics transcription was achieved by adapting a Wav2Vec 2.0 [20] speech recognizer to the singing domain by transfer learning [21]. However, in our preliminary tests we found the current SOTA speech recognition system, Whisper [22], to outperform the SOTA in automatic lyrics transcription [21] when transcribing a vocal performance. Therefore, Whisper [22] is used for automatic lyrics transcription in PolySinger.

Limited success has been achieved using speech alignment systems for lyrics alignment [23]. However, great results have been obtained in word-level lyrics alignment by training a polyphonic acoustic model in [24], but it is not until in [25] that a direct attempt is made at phoneme-level lyrics alignment without sacrificing competitive performance in word-level alignment. Recent approaches have exploited the correlation between phoneme onset and note pitch by joint representation learning [26] or cross-modal embedding in the audio and text domain through contrastive learning [27]. For PolySinger we use [25] due to its documented performance in the specific task of phoneme-level lyrics alignment and accessibility to a pre-trained model.

Defining note-level events in a vocal melody is a complex task, and thus there is a lack of datasets and trained neural networks for note-level vocal melody extraction (VME) [17]. On the other hand, frame-level VME is an extensively researched field with robust frameworks [28–30]. Considering the high accuracy of most modern frame-level VME systems, [30] is used in PolySinger due to the streamlined implementation available through the MIR toolkit Omnizart [31]. Instead of defining the note-level events by VME, we define them by syllable-wise boundaries delimited by the phoneme-level lyrics alignments, and guide the pitch of those notes with frame-level VME.

In [15], a rule-based approach is suggested for translating from English to Chinese lyrics with respect to the original lyrics, melody and rhythm, as well as the tonal properties of Chinese. In [16], a system for bidirectional translation between English and Chinese is proposed which incorporates an alignment decoder for determining the amount of syllables to write in the translation and how they should align to the melody. Additionally in [16], the evaluation process of the system is assisted by synthesizing the translation via SVS. For PolySinger, we take inspiration from [15] by going for a simple rule-based approach for English to Japanese lyrics translation due to data scarcity and an interest in unraveling the implications of processing Japanese lyrics. To do so, we exploit the pre-trained SOTA model for multi-lingual translation `nllb-200` [12] by transferring it to the singing domain.

Early work on SVS created concatenated singing libraries of sampled vocal sounds in a wide range of pitches, from which a synthesizer chose the samples for synthesis based on a musical score [32, 33]. More recent approaches use acoustic models trained on vocal performances from a singer to replicate the way he/she would perform a song given a musical score [7]. Furthermore, cross-lingual synthesis has become possible even when only training on mono-lingual singers [9]. The open-source scene has entered the SVS consumer market, e.g., by use of the ENUNU [2] plugin to enable usage of the NNSVS toolkit [8] in the OpenUTAU editor. Synthesizer V [3] is gathering a common consensus of being one of the best consumer products for SVS with a wide range of high-quality neural singing libraries capable of cross-lingual synthesis in a user-friendly environment with scripting capabilities. Therefore, Synthesizer V is used for SVS in PolySinger. Similarly to [16], we synthesize the translated lyrics, but we want to emphasize that, *differently from [16], PolySinger automates the intermediate link between automatic lyrics translation and SVS.*

3. PROPOSED SINGING-VOICE TO SINGING-VOICE TRANSLATION SYSTEM

Figure 1 illustrates a flowchart of the proposed SV2SVT system. This section will systematically break down the technology, implementation and functionality of each block presented in this figure.

3.1 Automatic Lyrics Transcription

Whisper [22] is a Transformer-based model [10] originally pre-trained on 680k hours of weakly-labeled audio for multi-task learning; there among the main task being multi-lingual automatic speech recognition. The most recent checkpoint, `Whisper-large-V3`, is trained on 1M hours of weakly-labeled audio and 4M hours of audio which was pseudo-labeled by `Whisper-large-V2`. We collect `Whisper-large-V3` from HuggingFace [4] for automatic lyrics transcription. The model has 1,550M parameters and was trained for 2 epochs on the dataset. We have not fine-tuned Whisper on singing data due to Whisper's great ability to generalize across several domains. To keep the memory usage of Whisper within ∼8 GB, a chunking algorithm segments the vocal performance into 30-second segments which are processed individually with a batch size of 4. Block 1 in Figure 1 is facilitated by Whisper to transcribe an English string of text from an English vocal performance.
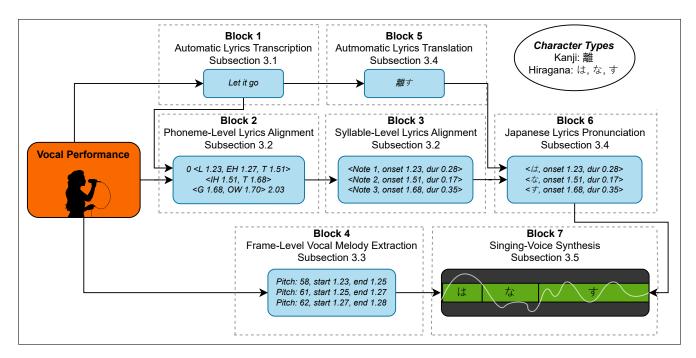
3.2 Phoneme-Level Lyrics Alignment and Syllable-Level Lyrics Alignment

In Western languages, poetry and lyrics are very reminiscent of each other. Poetry has a rhythmic structure called

---

[2] https://github.com/oatsu-gh/ENUNU
[3] https://dreamtonics.com/synthesizerv/
[4] https://huggingface.co/openai/
whisper-large-v3

Figure 1. Overview of our proposed SV2SVT system, PolySinger. Provided an English vocal performance, a synthetic vocal performance is created in Block 7, defined by notes with onsets, durations and Japanese lyrics, guided by a frame-level melody. Every numeric value is in seconds and "< >" illustrates the boundaries of notes. The process of segmenting words into syllables is illustrated in Table 1. Fundamentals of Japanese writing are explained in Subsection 3.4, and the process of converting kanji to hiragana is illustrated in Table 2.

meter. This structure can be dissected into a syllabic pattern [34]. Therefore, in this work, we define the onset and duration of notes by aligning the sung syllables to the vocal performance. To obtain syllable-level lyrics alignments, we first align the sequence of phonemes present in the vocal performance. The phoneme sequence is extracted with the pre-trained phoneme-level lyrics aligner informed in [25]. This model is a deep neural network trained for joint phoneme-level lyrics alignment and singing-voice separation. Text and audio are encoded separately. Text features and audio features are aligned by dynamic time warping-attention to minimize the total distance between audio frames and phonemes. The list of possible phonemes is provided by CMUdict [5]. Block 2 in Figure 1 uses the phoneme-level lyrics aligner to align the transcribed lyrics provided by Block 1 to the vocal performance. The phoneme alignments are informed in seconds. The phoneme sequence is concatenated into syllables by simple rules: *1)* it is assumed that each phoneme corresponding to a vowel makes an individual syllable, and *2)* consonants are merged with their closest neighboring vowel, gravitating towards the rightmost vowel in case of both neighboring phonemes corresponding to vowels. The process of breaking a word into phonemes according to CMUdict and concatenating them into syllables is illustrated in Table 1. In Block 3 of Figure 1, in order to perform syllable-level lyrics alignment, we define the onset of a note as the start of the first phoneme in a syllable, and we define the duration as the time difference between the onset and the end of the last phoneme in the syllable.

Word	BLUEBERRY						
Phonemes	B	L	UW1	B	EH2	R	IY0
Syllables		BLUW		BEH		RIY	

Table 1. Example of the word "blueberry" being deconstructed into phonemes with respect to CMUdict and reconstructed into syllables. An integer ranging 0-2 is associated with each vowel to indicate the type of vowel stress.

3.3 Vocal Melody Extraction

The notes in Block 3 of Figure 1 with timings defined by phoneme boundaries are not individually associated with a unique pitch. Instead, to preserve the melody from the vocal performance as much as possible, the melody is extracted at a frame level. This frame-level pitch contour is used to automate the pitch over time for the note sequence. The notes are set to a standard pitch of 60, and the contour is used to describe the deviation from 60 at each frame. The frame-level VME system presented in [30] deploys a deep convolutional neural network for semantic segmentation across a time-frequency image. Additionally, a progressive neural network is used for cross-domain transfer learning between the audio domain (frequencies) and the symbolic domain (pitch). Block 4 in Figure 1 utilizes this model for extracting the frame-level melody contour as pitches from the vocal performance. The start and end of pitches are in seconds. The model is used through the Omnizart toolkit [31].

3.4 Automatic Lyrics Translation and Japanese Lyrics Pronunciation

`nllb-200` [12] is a Transformer [10]-based mixture-of-experts (MoE) multi-lingual translation model that achieves SOTA results in many languages. This success is largely owed to the parallel development of datasets: *1)* the expertly-annotated Flores-200 dataset (which consists of 3,001 English sentences translated into 204 languages), and *2)* automatically-generated datasets by web-scraping for either mono-lingual sentences with high probability of being each other's translation, or mono-lingual sentences for back-translation. The biggest model, `nllb-200-MoE`, has 54B parameters, which is infeasible to run locally. Therefore, we collect the smaller checkpoint `nllb-200-distilled-600M` from HuggingFace [6] and fine-tune the model for English to Japanese lyrics translation. Since neither high-quality nor high-quantity dataset of paired English and Japanese lyrics exists, we take inspiration from [15] and scrape the web for lyrics translations that are not necessarily singable. Our dataset consists of ∼213k paired lines, thereof ∼80% (∼20%) being Japanese⇒English (English⇒Japanese) [7]. With PolySinger, we perform English⇒Japanese translation, so we invert the Japanese⇒English lyrics pairs into English⇒Japanese lyrics pairs. We hypothesize that this inversion does not raise an issue, but might in fact incentivize the model to produce high-quality Japanese lyrics even when provided with low-quality English lyrics. Besides, unlike in [15], we attempt fine-tuning with no prior self-supervised training on mono-lingual lyrics due to our larger paired dataset.

Ideally, the output of our fine-tuned model should have the same amount of syllables as established in Block 3 of Figure 1. However, counting syllables is not as simple in Japanese as in English. Japanese has moraic syllabaries in the form of kana. Kana characters have specific pronunciations that take up one mora (Japanese syllable) each. Japanese also uses kanji as logograms, that is, characters that convey a certain meaning. Kanji characters have multiple readings depending on the context, and, as such, it becomes a challenging task to decide the pronunciation of a Japanese sentence. To get the correct pronunciation of kanji characters, `pyKAKASI` [8] is used to decode kanji into their hiragana (a type of kana) readings. An illustration of the relation between kanji and Japanese pronunciation can be seen in Table 2. `pyKAKASI` is dictionary-based, and it can therefore be difficult to convert sentence-wise instead of word-wise. Japanese does not use blank space to separate words, therefore, we use `Nagisa` [9], a recurrent neural network trained for Japanese word segmentation.

During inference, a beam search is applied to the output

Kanji character	離	
Hiragana readings	り	はな
Roman readings	RI	HA NA
Mora count	1	2

Table 2. Example of two possible hiragana readings for a kanji character.

of our fine-tuned `nllb-200-distilled-600M` with as many beams as memory will allow (∼50 beams in our tests). The beams are biased towards a token count lower than the number of syllables in Block 3 of Figure 1 due to a kanji always corresponding to at least one syllable. Each generated sentence becomes word-separated with `Nagisa` and the kanji are converted into hiragana readings with `pyKAKASI`. The sentence with the lowest non-negative difference between mora count and syllable count gets selected and assigned to the notes in Block 6 of Figure 1.

3.5 Singing-Voice Synthesis

Synthesizer V is a SVS system with growing popularity among musicians. The technology behind Synthesizer V is kept proprietary. Based on related literature [7, 8], it is assumed that AI singing-voices in Synthesizer V are acoustic models trained on phoneme-level annotated vocal performances. With this training scheme, the model recognizes patterns in a singer's vocal performances, e.g., articulation of phoneme sequences, transitions between pitches and tendencies to use vibrato. Additionally, Synthesizer V AI voices have parameters for vocal modes, which can be included in training by annotating vocal samples with a singing style, e.g., nasal, powerful, soft, and whisper. AI singing-voices are usually only trained on vocal performances by a mono-lingual or bilingual singer, however AI voices in Synthesizer V are capable of cross-lingual synthesis in English, Japanese, Mandarin Chinese, Cantonese, and, recently, Spanish. It is assumed, based on related literature [9], that cross-lingual synthesis is achieved by unifying phoneme representations across languages with the international phonetic alphabet and training on data labeled with language identification such that the acoustic model can learn language-specific features. As illustrated in Figure 1, the notes with Japanese lyrics provided by Block 6 are plotted into Synthesizer V at a standard pitch of 60. The vocal contour provided by Block 4 is used to automate the deviation from pitch 60 over time. We use the AI singing-voice Mai in Synthesizer V to generate the Japanese vocal performance.

4. EXPERIMENTS

Objective measures in machine translation such as BLEU [35] are typically used for word-wise similarity with respect to a ground truth. Such a method does not suit lyrics translation as there should rather be a focus on semantic interpretation rather than precise word choice. More-

---

[6] https://huggingface.co/facebook/nllb-200-distilled-600M

[7] Both English⇒Japanese and Japanese⇒English are collected from https://lyricstranslate.com/. Extra Japanese⇒English is also collected from https://www.animelyrics.com/. More information about this dataset can be found at https://silasantonisen.github.io/polysinger/

[8] https://codeberg.org/miurahr/pykakasi

[9] https://github.com/taishi-i/nagisa

Score	1	2	3	4	5
Meaning	Very poor	Poor	Neutral	Good	Very good

Table 3. Five-point scale for MOS test.

ID	Question
Q1	How much sense do the lyrics make?
Q2	How natural is the Japanese used in the lyrics?
Q3	How well is the meaning of the original lyrics preserved?
Q4	How singable are the generated lyrics?
Q5	How well are the lyrics and melody aligned?
Q6	What is the overall quality of the generated Japanese singing?

Table 4. Questions asked to the test subjects in the MOS test of Section 4.

over, PolySinger has to be evaluated on the overall performance achieved for English⇒Japanese SV2SVT rather than solely on the translation quality. Therefore, we evaluate PolySinger subjectively by means of a MOS test.

4.1 Methodology

Six native Japanese speakers participated in a MOS test to evaluate the perceptual quality of English⇒Japanese SV2SVT using PolySinger on 5 different vocal performances. All test subjects were females ranging from 24 to 39 years old with no hearing impairment. The test subjects were asked to self-report their English speaking level. Two participants reported complete fluency (5/5), one reported near fluency (4/5), two more indicated advanced comprehension (3/5), and the final one reported intermediate comprehension (2/5). Using the inference procedure described in Subsection 3.4, PolySinger was alternately tested with the original `nllb-200-distilled-600M` (*Baseline*) and our fine-tuned `nllb-200-distilled-600M` (*Fine-tuned*) on every vocal performance. The test subjects were asked to first listen to an English vocal performance, followed by the synthetic performances generated by the two PolySinger versions (i.e., Baseline and Fine-tuned). Participants were not informed which synthetic vocal performance was generated by which system variant. Using the 5-point scale shown in Table 3, the test subjects were asked to assess each generated performance by the 6 MOS questions displayed in Table 4. The average time a participant spent on the evaluation was 53 min. The audio samples used for evaluation can be accessed here [10].

After the participants submitted their MOS scores, we additionally had a brief discussion with them individually about their general opinions and observations.

4.2 Results

Table 5 shows the MOS test results along with 95% confidence intervals from the Student's t-distribution [36]. Both system variants (i.e., Baseline and Fine-tuned) lie somewhere between poor and neutral in all 6 MOS questions Q1–Q6. The relatively large confidence intervals in Table

---

[10] https://silasantonisen.github.io/polysinger/

5 suggest a high variance in opinion scores. We investigate this variance in Figure 2 by representing per-question score's relative frequency. While it is true that the majority of opinion scores lies in the mid-to-low end of the spectrum, several evaluations have also resulted in good or very good opinion scores. This emphasizes the very subjective nature of the SV2SVT problem.

Given a MOS question Q1–Q6, we determine if there is a statistically significant difference between the opinion scores for Baseline and Fine-tuned. A Kolmogorov-Smirnov test [37] generally rejects, at a standard significance level of 5%, the null hypothesis that our opinion score sample populations follow Gaussian distributions. Therefore, we use a Wilcoxon rank-sum test [38] to determine whether there are statistically significant differences in MOS between the two system variants. The p-values shown in Table 6 demonstrate that the performance of the two systems is rather equivalent. Specifically, these p-values indicate that there are no statistically significant differences between Baseline and Fine-tuned at a standard significance level of 5% given any of the 6 MOS questions.

During discussions conducted after the MOS test, the test subjects generally conveyed a positive reaction towards SV2SVT being possible with PolySinger. However, as anticipated, the participants mainly assessed PolySinger by the naturalness of the Japanese language used in the context of singing and the pronunciation of words. The most recurring observations from the participants, that were suggested as crucial improvements needed for the pursuit of natural Japanese singing, are summarized in Table 7. In the next section, we will discuss the comments in Table 7 as to why our processing and synthesis of Japanese might not have been of ideal quality, along with our plan for improving them in future. Moreover, the statistically insignificant difference between Baseline and Fine-tuned is also discussed along with techniques and technologies that may assist in improving PolySinger.
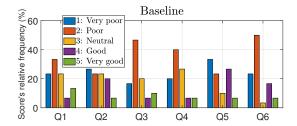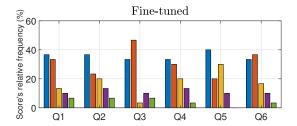
5. DISCUSSION

To produce natural Japanese speech synthesis, the front-end of a text-to-speech system requires phonetic and prosodic features [39]. Phonetic features, i.e., pronunciations, are typically acquired by grapheme-to-phoneme (G2P) conversion, and prosodic features, i.e., rhythm and intonation, are in Japanese typically acquired by phrase break prediction and accent estimation [40–42]. G2P conversion is particularly difficult in Japanese, since kanji characters can have multiple pronunciations. As indicated by our test subjects (C1 in Table 7) and discussed in [40], the accuracy obtained by dictionary-based G2P conversion in Japanese is not satisfactory. Japanese has no word separators, which also makes it difficult to determine phrase breaks. In our work, we performed word segmentation with `Nagisa` to avoid intra-word breaking, and attempted to define phrase breaks as the pauses transcribed by lyrics alignment on an English vocal performance. However, according to our test subjects (C5 in Table 7), these methods yielded limited success. As future work, we will inves-

System / Question	Q1	Q2	Q3	Q4	Q5	Q6
Baseline	2.53 ± 0.49	2.57 ± 0.48	2.47 ± 0.44	2.40 ± 0.41	2.50 ± 0.52	2.33 ± 0.45
Fine-tuned	2.17 ± 0.46	2.30 ± 0.48	2.10 ± 0.44	2.23 ± 0.44	2.10 ± 0.40	2.13 ± 0.41

Table 5. MOS quality test results, broken down by question, with 95% confidence intervals.

Figure 2. Bar plots representing per-question score's relative frequency from the MOS quality test for Baseline (left) and Fine-tuned (right).

Question	Q1	Q2	Q3	Q4	Q5	Q6
p-value	0.228	0.389	0.115	0.509	0.279	0.557

Table 6. *p*-values, broken down by question, from a Wilcoxon rank-sum test comparing MOS scores from Baseline and Fine-tuned.

ID	Comment
C1	Incorrect readings of kanji
C2	Usage of keigo in casual language
C3	Direct translations where interpretations are needed
C4	Occasionally, lyrics are not entirely translated
C5	Both intra- and inter-word separation at unnatural places
C6	Missing keywords important to the song
C7	Wrong word order
C8	Improper mixture of feminine and masculine language

Table 7. Comments from discussions with the test subjects on essential improvements that could lead to more natural synthetic Japanese singing.

tigate the adaptation of SOTA methodologies in Japanese text-to-speech to SV2SVT such as phrase break prediction with large language models (LLMs) [41] and G2P conversion via machine translation [40].

In [15], they demonstrated an improvement in automatic lyrics translation by fine-tuning on paired lyrics that were not necessarily singable, but also by pre-training on mono-lingual lyrics. In this work, we avoided pre-training on mono-lingual lyrics and only fine-tuned on paired lyrics that were not necessarily singable, which resulted in no statistically significant improvement with respect to the baseline model (see Table 6). We applied a beam search to find translated lyrics that fit well into the syllable count of the original lyrics. The selected lyrics were occasionally not a full translation of the original lyrics (C4 in Table 7). Apart from the use of keigo (honorific language) being inappropriate for the inherent casual nature of song lyrics (C2 in Table 7), we conjecture that keigo could also be a major cause of incomplete lyrics translations. This is because keigo will usually incorporate more characters than casual language, which means that it will be harder to fit the lyrics into the fixed syllable count.

In [16], they achieve SOTA results by training on a dataset created by back-translating mono-lingual lyrics and automatically aligning automatically-generated melodies that fit both the source and target lyrics. As future work, creating such a dataset and training an alignment decoder similarly to [16] could very well be adapted to Japanese. However, we hypothesize that translation systems have an inherent limitation towards cross-lingual songwriting that hinders them from rivaling professional human translators due to a lack of abstract interpretation and "imagination". Hence, as future work, we will also investigate the usage of LLMs for sentiment analysis and feature extraction to exploit poetry/lyrics generation models. By lyrics generation, guided by keyword spotting, we can also address the issue of missing keywords (C6 in Table 7).

6. CONCLUSION

The goal of this paper has been to adapt conventional S2ST to the singing domain. To do so, we have built the first SV2SVT system, PolySinger, by cascading SOTA MIR technologies facilitating a modular tool for extended research in SV2SVT. We have conducted a MOS test with native Japanese speakers to evaluate PolySinger's performance for English to Japanese SV2SVT. Results indicate that we have created a fundamentally-coherent structure for SV2SVT, but the translation of English lyrics into Japanese and the automatic synthesis of it is not yet natural enough. To further develop SV2SVT, our future work will investigate —to facilitate creative lyrics generation— the usage of sentiment analysis and feature extraction for abstract meaning representation of lyrics as opposed to translation. Finally, we will also investigate the necessities for autonomous generation of natural Japanese lyrics.

8. REFERENCES

[1] A. Lavie, A. Waibel, L. Levin, M. Finke, D. Gates, M. Gavalda, T. Zeppenfeld, and P. Zhan, "Janus-III: speech-to-speech translation in multiple languages," in *Proc. of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, pp. 99–102.

[2] S. Nakamura, K. Markov, H. Nakaiwa, G. Kikui, H. Kawai, T. Jitsuhiro, J.-S. Zhang, H. Yamamoto, E. Sumita, and S. Yamamoto, "The ATR multilingual speech-to-speech translation system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, pp. 365–376, 2006.

[3] International Telecommunication Union, "Functional requirements for network-based speech-to-speech translation services," *ITU-T F.745*, 2016.

[4] Y. Jia, R. J. Weiss, F. Biadsy, W. Macherey, M. Johnson, Z. Chen, and Y. Wu, "Direct speech-to-speech translation with a sequence-to-sequence model," in *Proc. of Interspeech 2019*, 2019.

[5] X. Li, Y. Jia, and C.-C. Chiu, "Textless direct speech-to-speech translation with discrete speech representation," in *Proc. of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[6] Seamless Communication, L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, C. Klaiber, P. Li, D. Licht, J. Maillard, A. Rakotoarison, K. R. Sadagopan, G. Wenzek, E. Ye, B. Akula, P.-J. Chen, N. E. Hachem, B. Ellis, G. M. Gonzalez, J. Haaheim, P. Hansanti, R. Howes, B. Huang, M.-J. Hwang, H. Inaguma, S. Jain, E. Kalbassi, A. Kallet, I. Kulikov, J. Lam, D. Li, X. Ma, R. Mavlyutov, B. Peloquin, M. Ramadan, A. Ramakrishnan, A. Sun, K. Tran, T. Tran, I. Tufanov, V. Vogeti, C. Wood, Y. Yang, B. Yu, P. Andrews, C. Balioglu, M. R. Costa-jussà, O. Celebi, M. Elbayad, C. Gao, F. Guzmán, J. Kao, A. Lee, A. Mourachko, J. Pino, S. Popuri, C. Ropers, S. Saleem, H. Schwenk, P. Tomasello, C. Wang, J. Wang, and S. Wang, "SeamlessM4T: Massively multilingual & multimodal machine translation," *arXiv*, vol. abs/2308.11596, 2023.

[7] J. Liu, C. Li, Y. Ren, F. Chen, and Z. Zhao, "Diff-Singer: Singing voice synthesis via shallow diffusion mechanism," *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 10, pp. 11 020–11 028, 2022.

[8] R. Yamamoto, R. Yoneyama, and T. Toda, "NNSVS: A neural network-based singing voice synthesis toolkit," in *Proc. of the 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[9] X. Wang, C. Zeng, J. Chen, and C. Wang, "CrossSinger: A cross-lingual multi-singer high-fidelity singing voice synthesizer trained on monolingual singers," in *Proc. of the 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, and A. Fan, "Multilingual translation with extensible multilingual pretraining and finetuning," *arXiv*, vol. abs/2008.00401, 2020.

[12] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. L. Spruit, C. Tran, P. Y. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang, "No language left behind: Scaling human-centered machine translation," *arXiv*, vol. abs/2207.04672, 2022.

[13] J. Franzon, "Three dimensions of singability. An approach to subtitled and sung translations," in *Text and Tune. On the Association of Music and Lyrics in Sung Verse*. Peter Lang, 2015, pp. 333–346.

[14] M. Ghazvininejad, Y. Choi, and K. Knight, "Neural poetry translation," in *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 67–71.

[15] F. Guo, C. Zhang, Z. Zhang, Q. He, K. Zhang, J. Xie, and J. Boyd-Graber, "Automatic song translation for tonal languages," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 729–743.

[16] C. Li, K. Fan, J. Bu, B. Chen, Z. Huang, and Z. Yu, "Translate the beauty in songs: Jointly learning to align melody and translate lyrics," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 27–39.

[17] J.-Y. Hsu and L. Su, "VOCANO: A note transcription framework for singing voice in polyphonic music," in *ISMIR 2021: Proceedings of the 22nd International Society for Music Information Retrieval Conference*, 2021, pp. 293–300.

[18] E. Demirel, S. Ahlbäck, and S. Dixon, "Automatic lyrics transcription using dilated convolutional neural networks with self-attention," in *Proc. of the 2020 International Joint Conference on Neural Networks (IJCNN)*, 2020.

[19] E. Demirel, S. Ahlbäck, and S. Dixon, "MSTRE-Net: Multistreaming acoustic modeling for automatic lyrics transcription," in *Proc. of ISMIR 2021, International Society for Music Information Retrieval*, 2021.

[20] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[21] L. Ou, X. Gu, and Y. Wang, "Transfer learning of wav2vec 2.0 for automatic lyric transcription," in *Proc. of ISMIR 2022, International Society for Music Information Retrieval*, 2022.

[22] A. Radford, J. wook Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. of the 40th International Conference on Machine Learning: ICML 2023*, Hawaii (USA), 2023, pp. 28 492—28 518.

[23] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *Proc. of ISMIR 2016, International Society for Music Information Retrieval*, 2016.

[24] C. Gupta, E. Yılmaz, and H. Li, "Automatic lyrics alignment and transcription in polyphonic music: Does background music help?" in *Proc. of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 496–500.

[25] K. Schulze-Forster, C. S. J. Doire, G. Richard, and R. Badeau, "Phoneme level lyrics alignment and text-informed singing voice separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2382–2395, 2021.

[26] J. Huang, E. Benetos, and S. Ewert, "Improving lyrics alignment through joint pitch detection," in *Proc. of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2022, pp. 451–455.

[27] S. Durand, D. Stoller, and S. Ewert, "Contrastive learning-based audio to lyrics alignment for multiple languages," in *Proc. of ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2023.

[28] R. M. Bittner, J. Salamon, S. Essid, and J. P. Bello, "Melody extraction by contour classification," in *Proc. of the International Conference on Music Information Retrieval (ISMIR)*, 2015.

[29] S. Kum, C. Oh, and J. Nam, "Melody extraction on vocal segments using multi-column deep neural networks," in *ISMIR 2016: Proceedings of the 17th International Society for Music Information Retrieval Conference*, 2016, pp. 819–825.

[30] W.-T. Lu and L. Su, "Vocal melody extraction with semantic segmentation and audio-symbolic domain transfer learning," in *ISMIR 2018: Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018, pp. 521–528.

[31] Y.-T. Wu, Y.-J. Luo, T.-P. Chen, I.-C. Wei, J.-Y. Hsu, Y.-C. Chuang, and L. Su, "Omnizart: A general toolbox for automatic music transcription," *Journal of Open Source Software*, vol. 6, no. 68, p. 3391, 2021.

[32] M. Macon, L. Jensen-Link, E. B. George, J. Oliverio, and M. Clements, "Concatenation-based MIDI-to-singing voice synthesis," in *Proc. of the Audio Engineering Society Convention 103*, 1997.

[33] H. Kenmochi and H. Ohshita, "VOCALOID – commercial singing synthesizer based on sample concatenation," in *Proc. of Interspeech 2007*, 2007.

[34] E. Greene, T. Bodrumlu, and K. Knight, "Automatic analysis of rhythmic poetry with applications to generation and translation," in *Proc. of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010, pp. 524–533.

[35] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[36] N. Blachman and R. Machol, "Confidence intervals based on one or more observations," *IEEE Transactions on Information Theory*, vol. 33, no. 3, pp. 373–382, 1987.

[37] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[38] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, vol. 1, no. 6, pp. 80–83, 1945.

[39] T. Fujimoto, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Impacts of input linguistic feature representation on Japanese end-to-end speech synthesis,"

in *Proc. of the 10th ISCA Speech Synthesis Workshop. ISCA, Vienna, Austria*, 2019.

[40] N. Kakegawa, S. Hara, M. Abe, and Y. Ijima, "Phonetic and prosodic information estimation from texts for genuine Japanese end-to-end text-to-speech," in *Proc. of Interspeech*, 2021, pp. 126–130.

[41] K. Futamata, B. Park, R. Yamamoto, and K. Tachibana, "Phrase break prediction with bidirectional encoder representations in Japanese text-to-speech synthesis," in *Proc. of Interspeech 2021*, 2021.

[42] K. Kurihara, N. Seiyama, and T. Kumano, "Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS," *IEICE Transactions on Information and Systems*, vol. 104, no. 2, pp. 302–311, 2021.