Language Independent Speech-to-Singing-Voice Conversion

Akinori Ito[a,b]

[a]Graduate School of Engineering, Tohoku University, Sendai, Japan
[b]Advanced Institute of So-Go-Chi Informatics, Tohoku University, Sendai, Japan

ABSTRACT

This research addresses the challenge of converting spoken voice into singing voice in a language-independent manner. Traditional speech-to-singing systems often rely on language-specific phoneme alignment or require parallel singing datasets, which limits their applicability across languages and speakers. To overcome these constraints, the authors propose a novel framework that utilizes voiced/unvoiced (V/UV) classification and music state modeling to align speech with musical scores without relying on linguistic content. The approach begins by extracting a V/UV state sequence from input speech using a convolutional layer built on top of a pretrained HuBERT model. Simultaneously, a music state sequence is generated from a monophonic musical score using a decay function that models note intensity over time. These two sequences are then aligned using Dynamic Time Warping (DTW), allowing the system to synchronize speech features with musical timing and pitch. After alignment, the World vocoder is employed to analyze and synthesize the singing voice. The spectral and aperiodic components of speech are aligned to the music sequence, while pitch is replaced with musical pitch to produce the final singing output. Experimental results demonstrate the effectiveness of the proposed V/UV classification using the ATR speech database. The system could generate singing voices from spoken input without requiring phoneme-level annotations or parallel singing data, but still there is a room for quality improvement.

Keywords: Singing voice, Voiced/unvoiced classification, HuBERT

1. INTRODUCTION

There are thousands of languages in the world, but more than 3,000 languages are reportedly at risk of extinction.[1] People have made a significant effort to avoid language extinction. Language revitalization[2,3] is an effort to increase the number of speakers of minority languages. There are several steps in the language revitalization. Development of educational material[4,5] is an important step to reproduce the speakers of that language. Additionally, it is crucial to draw the public's attention to the target language through festivals or entertainment content.[6]

Music plays a crucial role in the education and cultural promotion of those languages.[7–9] Not only folk music, but also pop music is an important source of promoting Indigenous language,[10] especially among young people and children. However, it is not always easy to find instrumental players and singers for that language. To develop such musical content for minority languages, Sleeper developed a singing synthesizer of Cherokee using UTAU singing synthesizer system.[11]

When developing music content using singing synthesizers for a new language, we first need to develop a model of singing voice synthesis for that language. It requires a deep understanding of language and speech technology, as well as effort to collect language and singing voice data. Therefore, if we have a system that generates a singing voice without developing complicated language-dependent synthesis models, it would be beneficial for the activity of language revitalization.

In this paper, we propose a language-independent speech-to-singing conversion system. If we provide any speech and musical score to the system, it converts the speech into the singing voice associated with the given score. Therefore, this system can be used for generating any musical content, including both folk songs and pop music.

Further author information: (Send correspondence to A.I.)
A.I.: E-mail: aito.spcom@tohoku.ac.jp

2. RELATED STUDIES

Saitou et al. proposed a system named *SingBySpeaking*, the first system to convert a talking speech into a singing voice.[12,13] This system assumes that the user speaks the lyrics of the song to be generated, and the score of the target song is given. The system first converts the lyrics into a phoneme sequence and calculates the phoneme-speech alignment using the Viterbi algorithm. At this time, the system knows which phoneme is associated with which note using the score. Then the system changes the duration and pitch of phonemes according to the duration and pitch of the corresponding note. The STRAIGHT[14] is used to manipulate the speech. Additionally, the system converts the speech spectrum to make it sound more like a singing voice by adding the singing formant and vibrato.

Vijayan et al.[15] proposed another method, the template-based speech-to-singing conversion, which assumes that the user has a singing voice of the target song. By aligning the input speech with the template singing voice, we can determine how the input speech should be converted, including its duration and pitch. As a result, we can generate a singing voice of the source speaker mimicking the song of the template singer.

Recently, deep-learning-based models have been used for the speech-to-singing task. Perach et al. reformulated the speech-to-singing task as a style transfer problem.[16] Their system receives the source speech and melody contour (piano roll image) using separate encoders. It generates the converted spectrogram that sounds like a singing voice using a U-net-like network. They incorporated multi-task learning to recognize the phoneme sequence of the input speech, thereby enhancing the linguistic clarity of the generated singing voice. Li et al.[17] proposed a speech-to-singing conversion method based on a self-supervised pre-trained model. They utilized the acoustic token, which is commonly employed in speech synthesis or translation models. When generating the token sequence, they use the pitch information to create the singing voice of the designated pitch.

All of these systems assume that the language of the input speech (and thus the output singing voice) is known. Additionally, the machine-learning-based model assumes a sufficient amount of linguistic resources, such as speech and singing voice databases. However, in the case of low-resource languages, it is difficult, or impossible, to gather such data or develop a system for phoneme alignment. Therefore, we need to establish a language-independent framework for speech-to-singing conversion.

3. PROPOSED METHOD

3.1 Overview

Figure 1 shows the overview of the system. Since the system is language-independent, we only use voiced and unvoiced speech classification, which determines three-class classification (voiced, unvoiced, and silence) for each 20 ms frame. Then the sequence of voiced/unvoiced/silence states ("V/UV state sequence" in the figure) is generated. Next, frame-by-frame pitch and velocity information ("Music state sequence" in the figure) is generated from the given musical score. These two state sequences are aligned using the dynamic time warping (DTW) technique. In this process, the musical state sequence is kept unchanged, and only the V/UV state sequence is aligned. From the DTW block, the alignment information is generated.

The input speech is analyzed using the World vocoder system.[18] Separately, the pitch information generated from the score is combined with the aligned spectrum and aperiodicity, and then those features are used to re-synthesize the speech.

3.2 Voiced/Unvoiced classification

Since the proposed system is language-independent, any specific linguistic knowledge, such as phoneme inventory, cannot be utilized. Therefore, we decided to use only information of voiced, unvoiced, and silence. The voiced part corresponds to the notes, and the unvoiced and silence parts are the release parts of notes or rests.

We used HuBERT,[19] a pre-trained transformer model developed by self-supervised learning, as the feature extractor. HuBERT is known to be efficient for extracting speech features of multiple languages.[20] We attached one 1-D convolutional layer with 256 hidden units to summarize the output of the transformers into three classes (voiced, unvoiced, and silence).

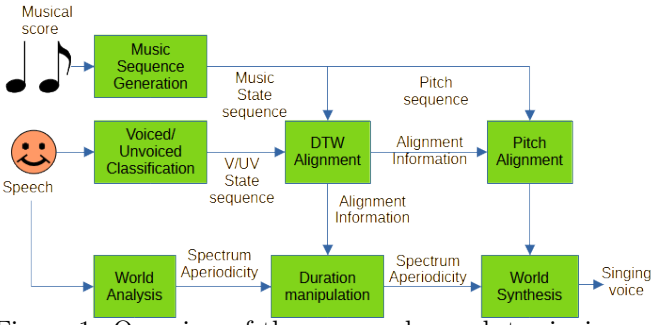


Figure 1: Overview of the proposed speech-to-singing system.

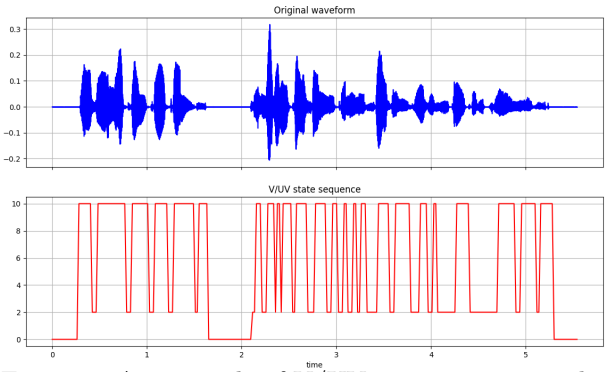


Figure 2: An example of V/UV state sequence. The upper pane is the waveform, and the lower pane is the corresponding V/UV state sequence.

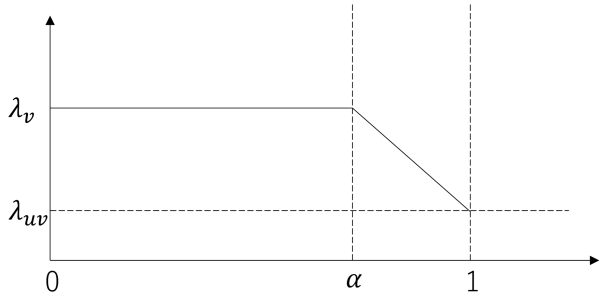


Figure 3: The decay function $D(x)$.

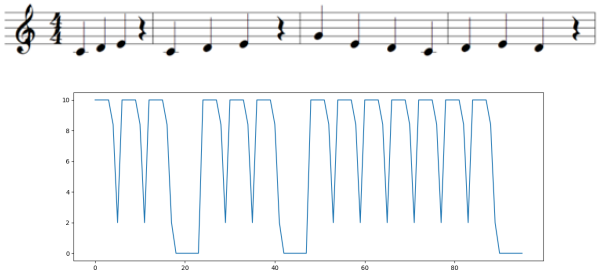


Figure 4: An example of music state sequence.

Finally, the V/UV state sequence is generated. Figure 2 shows an example of V/UV state sequence. Here, the sequence is $s_1, \ldots, s_T$, where $s_i \in \{0, \lambda_{uv}, \lambda_v\}$ is a state value of frame i. $0 < \lambda_{uv} < \lambda_v$ are the state values corresponding to the silence, unvoiced, and voiced frames. In the experiment, we employed $\lambda_{uv} = 2$ and $\lambda_v = 10$.

3.3 Music state sequence generation

Next, we define the music state sequence. As we assume that the musical score is monophonic, the state of frame t is either a particular time after the previous onset of a note, or within the duration of a rest.

Let $N_1, \ldots, N_K$ be a sequence of notes, where $N_i = (h_i, L_i)$. Here, h_i is the pitch or rest and L_i is the duration (number of frames) of the note. Then we define the music state sequence as follows. Let the duration (in frames) until the note N_i be

$$M_0 = 0, \ M_i = M_{i-1} + L_i = \sum_{j=1}^{i} L_i. \tag{1}$$

Then, the music state sequence $q_1, \ldots, q_N$ is defined as

$$q_{M_i+j} = \begin{cases} 0 & h_i \text{ is a rest} \\ D(j/L_i) & \text{otherwise} \end{cases} \tag{2}$$

Here, $D(x)$ is a decay function, defined as

$$D(x) = \begin{cases} \lambda_v & x < \alpha \\ \lambda_v + \frac{(\lambda_{nv} - \lambda_v)(x-\alpha)}{1-\alpha} & x \geq \alpha \end{cases} \tag{3}$$

for $0 \leq x \leq 1$. $0 < \alpha < 1$ is a parameter controlling the decay; $\alpha = 0.75$ was employed in the experiment. Figure 3 shows the overview of the decay function, and Figure 4 shows an example of music state sequence.

3.4 Alignment using the Dynamic Time Warping

The V/UV state sequence $s_1, \ldots, s_T$ and the music state sequence $q_1, \ldots, q_N$ are aligned using the Dynamic Time Warping (DTW) technique.

Let $g(i, j)$ be the accumulated distance and $b(i, j)$ be the backpointer. Then the values of $g(i, j)$ are calculated using the following recursive formula:

$$g(i,j) = \infty \ (i \leq 0 \text{ or } j \leq 0) \tag{4}$$

$$g(1,1) = |s_1 - q_1| \tag{5}$$

$$g(1,j) = |s_1 - q_j| + g(1, j-1) + P(1, j-1) \ (j > 1) \tag{6}$$

$$g(i,j) = |s_i - q_j| + \min \begin{cases} g(i, j-1) + P(i, j-1) \\ g(i-1, j-1) \\ g(i-2.j-1) + P(i-2, j-1) \end{cases} \tag{7}$$

$$g(i,j) = |s_i - q_j| + g(i-1, j-1) \tag{8}$$

Eq. (7) is used when s_i is a voiced or silent frame, while Eq. (8) is used when s_i is an unvoiced frame. When calculating the minimum operation in Eq. (7), the selection is recorded in $b(i, j)$ for use in the backtrace. $P(i, j)$ is a penalty term to penalize uneven alignment.

3.5 Analysis and synthesis using World vocoder

When analyzing the input speech using the World vocoder, it converts the input into three components: the smoothed spectrum, aperiodicity, and pitch (f0). Then only the spectrum and aperiodicity are aligned to the music state sequence according to the alignment information. Note that the frame shift of U/UV classification is 20 ms, while that of the World vocoder is 5 ms. Therefore, the alignment sequence is stretched four times to align the frame shift to 5 ms.

After aligning the spectrum and aperiodicity, the pitch sequence from the music state sequence is combined to generate the singing voice waveform.

4. EXPERIMENT

4.1 V/UV classification

The V/UV classification model is based on hubert-base-ls960 downloaded from HuggingFace. ATR speech database set B was used for training, validation, and evaluation of the model. Six speakers' (four males and two females) 1,800 utterances (300 utterances/speaker) were used for training the model. The database has phoneme labels; we regarded the vowels, semivowels, and the moraic nasal as voiced, and all the other consonants as unvoiced. Although there are voiced consonants (such as /m/ or /g/), we regard them as unvoiced because these consonants are not stretched or shrunk according to the note duration. We used 200 utterances from two speakers for verification and 200 utterances from the other two speakers for evaluation. No same speakers and sentences are included in different sets.

For comparison, we employed two relatively simple V/UV classifiers:

1. **Simple**: WebRTC voice activity detector[21] combined with librosa pitch tracker based on pYIN.[22] First, silence frames are determined using the VAD, and then we regarded the frames as voiced when pitch is estimated in that frame.

2. **World**: WebRTC and the World's aperiodicity. The World calculated the aperiodicity vector for each frame, which has higher value for unvoiced frame. Therefore, we calculated the average of the aperiodicity vector and compared with a threshold. If the average value is larger than the threshold, we regard that frame as unvoiced. The threshold is optimized on the evaluation set.

Table 1 shows the frame-by-frame accuracy result. This result clearly shows that the proposed U/UV classifier outperforms other conventional methods. Figure 5 shows a visualized example of V/UV classification. From this example, it is clear that the proposed method gives almost exact classification result compared to the annotation.

Table 1: Frame-by-frame accuracy of U/UV classification

Method	Accuracy
Proposed	95.7
Simple	55.2
World	77.3

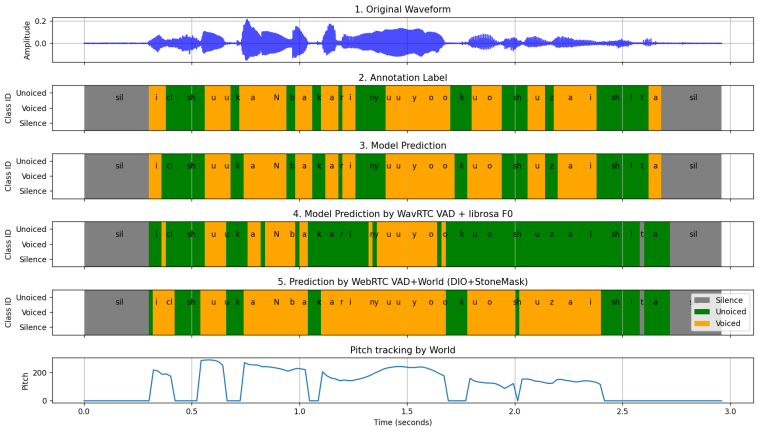
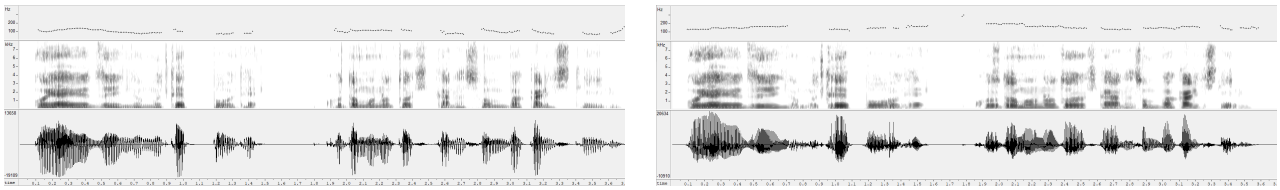


Figure 5: An example of V/UV classification



(a) Input voice (b) Generated singing voice

Figure 6: Examples of the original and generated voice signals

4.2 Generation of singing voice

Next, we examined generation of singing voice from speech. We used a male's Japanese voice, "*Oyayuzurino muteppoōde kodomonokorokara sonbakari siteiru.*" The voice is 3.8 s long. For the music, we used the score shown in Figure 4. The length of the generated singing voice was set to be almost the same with the original speech.

Figure 6 shows the pitch contour, spectrogram, and waveform of the original and processed signals. We can see that the syllable intervals in (b) are more uniform than that in (a), and the pitch contour of (b) becomes the melody line of Figure 4. Although we have not completed subjective evaluation of speech quality, the quality of (b) is not high. Part of the reason of the degradation is the use of World vocoder, which makes the voice degraded on the large manipulation. However, we can still perceive the original words in the speech and the pitch sounds accurate.

5. CONCLUSIONS

This paper proposes a language-independent speech-to-singing conversion system. This system is based on three fundamental techniques. First, highly accurate voiced/unvoiced classification is achived using a model based on HuBERT pretrained model. Second, DTW-based alignment on the V/UV state sequence and the music state sequence enables the language-free alignment of the input speech and the musical score. Third, analysis and synthesis of the speech are based on the World vocoder system. The system could convert the input speech into singing voice, but the quality of generated speech is not satisfactory.

As future work, the quality of the generated speech should be improved using neural vocoders. Nevertheless, we need to add singing-specific expressions to the output voice, such as vibrato and singing formant.

REFERENCES

[1] Campbell, L. and Belew, A., [*Cataloguing the world's endangered languages*], vol. 711, Routledge London and New York (2018).

[2] Hinton, L., "Language revitalization: An overview," *The green book of language revitalization in practice* **1**, 18 (2001).

[3] Guerrettaz, A. M. and Engman, M., "Indigenous language revitalization," in [*Oxford encyclopedia of race and education*], Oxford University Press (2023).

[4] Hinton, L., "Language revitalization and language pedagogy: New teaching and learning strategies," in [*Applied linguists needed*], 41–52, Routledge (2014).

[5] Nee, J., "Creating books for use in language revitalization classrooms: considerations and outcomes," *L2 Journal: An Open Access Refereed Journal for World Language Educators* **12**(1) (2020).

[6] Hara, K. and Heinrich, P., "27. linguistic and cultural revitalization," *Handbook of the Ryukyuan Languages* (2015).

[7] Vallejo, J. M., "Revitalising language through music: a case study of music and culturally grounded pedagogy in two Kanien'ke: ha (Mohawk) language immersion programmes," in [*Ethnomusicology Forum*], **28**(1), 89–117, Taylor & Francis (2019).

[8] Ansah, M. A., Agyeman, N. A., and Adjei, G., "Revitalizing minority languages using music: Three South-Guan languages of Ghana in focus," *Research Journal in Advanced Humanities* **3**(1), 19–34 (2022).

[9] Dembling, J., "Instrumental music and Gaelic revitalization in Scotland and Nova Scotia," *International Journal of the Sociology of Language* **2010**(206), 245–254 (2010).

[10] Huang, K., "'we are indigenous people, not primitive people.': the role of popular music in indigenous language revitalization in taiwan," *Current Issues in Language Planning* **24**(4), 440–459 (2023).

[11] Sleeper, M., "Singing synthesizers: Musical language revitalization through UTAUloid," *Canadian Journal of Applied Linguistics* **27**(2), 52–84 (2024).

[12] Saitou, T., Goto, M., Unoki, M., and Akagi, M., "Speech-to-singing synthesis: Converting speaking voices to singing voices by controlling acoustic features unique to singing voices," in [*2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*], 215–218, IEEE (2007).

[13] Saitou, T., Goto, M., Unoki, M., and Akagi, M., "Speech-to-singing synthesis system: Vocal conversion from speaking voices to singing voices by controlling acoustic features unique to singing voices," in [*National Conference on Man-Machine Speech Communication (NCMMSC2009)*], (2009).

[14] Kawahara, H., Masuda-Katsuse, I., and De Cheveigne, A., "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication* **27**(3-4), 187–207 (1999).

[15] Vijayan, K., Dong, M., and Li, H., "A dual alignment scheme for improved speech-to-singing voice conversion," in [*2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*], 1547–1555, IEEE (2017).

[16] Parekh, J., Rao, P., and Yang, Y.-H., "Speech-to-singing conversion in an encoder-decoder framework," in [*ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*], 261–265, IEEE (2020).

[17] Li, R., Huang, R., Wang, Y., Hong, Z., and Zhao, Z., "Self-supervised singing voice pre-training towards speech-to-singing conversion," *arXiv preprint arXiv:2406.02429* (2024).

[18] Morise, M., Yokomori, F., and Ozawa, K., "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems* **99**(7), 1877–1884 (2016).

[19] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing* **29**, 3451–3460 (2021).

[20] Koshikawa, T., Ito, A., and Nose, T., "Fast and speaker-independent utterance selection for ASR-free CALL systems of minority languages," in [*APSIPA Annual Summit and Conference*], (2025).

[21] Sredojev, B., Samardzija, D., and Posarac, D., "WebRTC technology overview and signaling solution design and implementation," in [*2015 38th international convention on information and communication technology, electronics and microelectronics (MIPRO)*], 1006–1009, IEEE (2015).

[22] Mauch, M. and Dixon, S., "pYIN: A fundamental frequency estimator using probabilistic threshold distributions," in [*2014 ieee international conference on acoustics, speech and signal processing (icassp)*], 659–663, IEEE (2014).