PJS: phoneme-balanced Japanese singing-voice corpus

Junya Koguchi* and Shinnosuke Takamichi† and Masanori Morise*
* Meiji University, Tokyo, Japan
E-mail: cs202027@meiji.ac.jp
† The University of Tokyo, Tokyo, Japan
E-mail: shonnosuke_takamichi@ipc.i.u-tokyo.ac.jp
‡ Meiji University, Tokyo, Japan and JST, PRESTO, Japan
E-mail: mmorise@meiji.ac.jp

***Abstract*—This paper presents a free Japanese singing-voice corpus that can be used for highly applicable singing-voice synthesis research. A singing-voice corpus helps develop singing-voice synthesis, but existing corpora have two critical problems: data imbalance (i.e., singing-voice corpora do not guarantee phoneme balance, unlike speaking-voice corpora) and copyright issues (i.e., cannot legally share data). To avoid these problems, we constructed a phoneme-balanced Japanese singing-voice (PJS) corpus that guarantees phoneme balance and is licensed with CC BY-SA 4.0, and we composed melodies using a phoneme-balanced speaking-voice corpus. Furthermore, to temporally align phoneme sequences with speech feature sequences, we compare three alignment methods: Viterbi alignment of hidden Markov models, dynamic time warping using a synthesized voice, and statistical voice conversion. Experimental results demonstrate that 1) our corpus contains more unique monophones and diphones than an existing corpus, and 2) the voice-conversion-based method provides the most accurate alignment.**

I. Introduction

With the recent developments in deep learning and signal processing, we can now synthesize high-quality singing voices. Various deep learning architectures have been utilized (e.g., feed-forward [1], recurrent [2], sequence-to-sequence [3] and auto-regressive types [4]), and many products have been launched (e.g., Sinsy [5], NEUTRINO [6], and Synthesizer V [7]).

Freely available singing-voice corpora contribute to applicable and reproducible singing-voice synthesis research. Corpora are being developed in many languages (e.g., Chinese [8], English [9], etc. [10]). The leading Japanese corpus, the large RWC Music Database [11], [12], was developed 15 years ago. While the RWC corpus was designed for more general use in music information research, the recently developed Tohoku Kiritan database (hereinafter called "Kiritan") [13], HTS demo [14], and JVS-MuSiC [15] were designed for singing-voice synthesis. Some open-source codes already support building singing-voice synthesis using these corpora, and we expect these corpora to continue to contribute. However, existing corpora have two critical problems: data imbalance and copyright issues. The former means that existing corpora do not guarantee phoneme balance (unlike speaking-voice corpora); since phonemes are discrete symbols of linguistic content, synthesizers trained using a phoneme-imbalance corpus lose significant intelligibility of the synthesized singing voice. The latter means that we cannot legally share data. Namely, a third party cannot distribute expanded and modified corpus data. This paper describes the construction of a phoneme-balanced singing-voice corpus named the "phoneme-balanced Japanese singing-voice corpus" (*PJS*). Our corpus contributes the following:

Phoneme-balanced: Using the Voice Actress Corpus [16], a phoneme-balanced speaking-voice corpus, we composed melodies for 100 sentences. As demonstrated in **Section III**, our corpus contains rare diphones that are not included in the existing corpora.
CC BY-SA 4.0 license: All the data in our corpus is licensed with CC BY-SA 4.0. Therefore, redistribution, remixing, and transformation of our corpus is permitted for any purpose (e.g., research and commercial use), unlike existing corpora [8], [9], [10], [11], [12], [13], [14].
Singing and speaking voices: We recorded both singing voices and parallel speaking voices. This paired data contributes to speaking-singing research (e.g., [17]).
Descriptions of compositions: We noted descriptions of melody compositions. These descriptions contribute to natural-language–based music information research.
Availability online: Our corpus can be freely downloaded from our project page [18].

Furthermore, we explore a method of obtaining phoneme-voice alignment. Since singing-voice systems often use the phoneme-voice alignment [19], [20], better alignment helps build better-quality systems. We compare three methods: hidden Markov models (HMMs), dynamic time warping (DTW) using a synthesized voice, and statistical voice conversion. The first one is typically performed for speaking voices [21], the second one is ours we propose in this paper, and the third one is a combination of our method and Kotani's work [22]. Experimental results demonstrate that 1) our corpus contains more unique monophones and diphones than an existing corpus, and 2) the voice-conversion–based method provides the most accurate alignment.

II. Corpus construction

This section describes how we design, compose, and record for the PJS corpus.

A. Directory structure

Here, we list the directory structure of our corpus. *[SENTENCE_ID]* in directory PJS100_*[SENTENCE_ID]* is the sentence ID of the original speaking-voice corpus [16].

```
PJS100_001
  PJS100_001_song.wav
  PJS100_001_speech.wav
  PJS100_001.mid
  PJS100_001.xml
  PJS100_001.lab
  PJS100_001.txt
PJS100_002
...
PJS100_100
```

The directory PJS100_*[SENTENCE_ID]* consists of the following files:

- PJS100_*[SENTENCE_ID]*_song.wav: singing voice we composed using a sentence from the phoneme-balanced speaking-voice corpus [16] as the lyric
- PJS100_*[SENTENCE_ID]*_speech.wav: speaking voice that utters a sentence from the phoneme-balanced speaking-voice corpus [16]
- PJS100_*[SENTENCE_ID]*.mid: MIDI file we used as the guide melody during recording
- PJS100_*[SENTENCE_ID]*.xml: musicXML file that describes musical note information
- PJS100_*[SENTENCE_ID]*.txt: musical information that songs use (e.g., genre, scale, artist, etc.)

We composed and recorded 100 phoneme-balanced sentences [16]. The following sections describe the composition and recording conditions.

B. Composition conditions

A native Japanese male in his twenties composed all the songs. He is not a professional composer but has work experience using his singing, composing, and recording skills. He composed melodies within his range using each of the phoneme-balanced sentences. In addition, he composed the melodies to follow the pitch accents of the lyrics as much as possible in his compositions. Japanese is a pitch-accent language that changes the meaning of words by accents, and the accents should be considered for composing natural Japanese songs [23]. The musical notes he composed were written in PJS100_*[SENTENCE_ID]*.xml. He composed a variety of melodies (based on genre, scale, etc.). Descriptions of the compositions were written in PJS100_*[SENTENCE_ID]*.txt. He also made a MIDI file (PJS100_*[SENTENCE_ID]*.xml) of the composed melody to guide the recording described below.

C. Recording conditions

The composer was also the singer. While listening to the guide melody generated from the MIDI file, he recorded his singing voice so that his pitch and tempo would be as in sync with the guide as possible. To avoid the proximity effect of the microphone, we let him maintain $15\,\mathrm{cm}$ between the microphone and his mouth. The recording environment was a simple soundproof room in which we attached sound-absorbing materials to the walls. The recording environment was not an anechoic chamber, so we recorded 15-second background noise each recording day for noise reduction after the recording. We used a Lewitt LCT 441 FLEX (cardioid mode) [24] microphone, a JZ MICROPHONES Pop Filter [25] windscreen, and an RME Fireface UCX [26] audio interface.

We also let him record his speaking voice in the same manner. We saved the singing and speaking voices in the $48\,\mathrm{kHz}$-sampled, 24 bit-encoded RIFF WAV format.

D. Phoneme-voice alignment

Because singing-voice synthesis typically uses explicit temporal alignments (unlike end-to-end text-to-speech synthesis), we also made a temporal alignment between a singing voice and a phoneme sequence [27], [28]. While manual annotation is promising, it requires human annotators experienced in finding phoneme boundaries. A well-known automatic way for speaking voices is the Viterbi alignment based on phoneme-dependent HMMs [29]. The HMMs are trained using pairs of phoneme sequences and speech features, and the Viterbi path becomes a temporal alignment. While it works for speaking voices, it does not work for singing voices because singing voices have a strong dependency on monophones and other contexts [30].

To solve the above problem, we propose two alternative methods. The first method is using an existing singing-voice synthesis system (e.g., NEUTRINO [6]) to synthesize a singing voice from the musical notes (including phoneme labels) of our PJS corpus. Since the phoneme labels and the synthesized voice are temporally aligned, we align the phoneme labels and the recorded voices with the DTW between the synthesized voice and recorded singing voice. The second method is an extended version of the first. Since the speakers differ between the synthesized voices and the recorded voices, both the phonetic information and the speaker identities affect the alignment obtained by DTW. To reduce the speaker effect, we introduce statistical voice conversion during DTW [31]. First, we obtain the temporal alignment between the synthesized and the recorded voices as described in the above. Then, we use the aligned voices to train a statistical voice conversion model (e.g., Gaussian mixture model [32], [33]) and convert the speaker identity of the synthesized voice to that of the recorded voice. Finally, we take the DTW between the converted voice and recorded voice. Since this method solves the speaker identity problem, we expect the alignment to be more accurate. We can iterate DTW and voice conversion to improve the accuracy further, but we drive them only once, following Kotani's result [22].

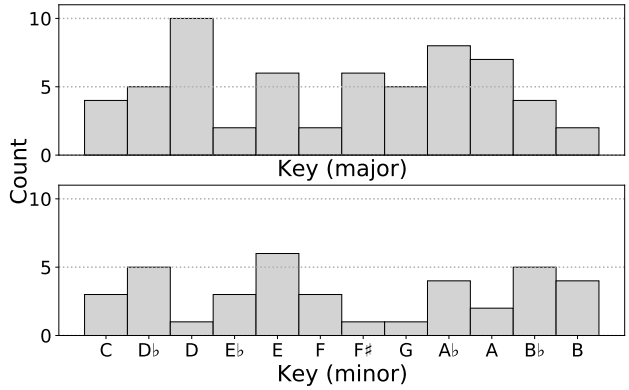

Fig. 1. Key histograms of our corpus. There are fewer songs in minor keys than in major keys.

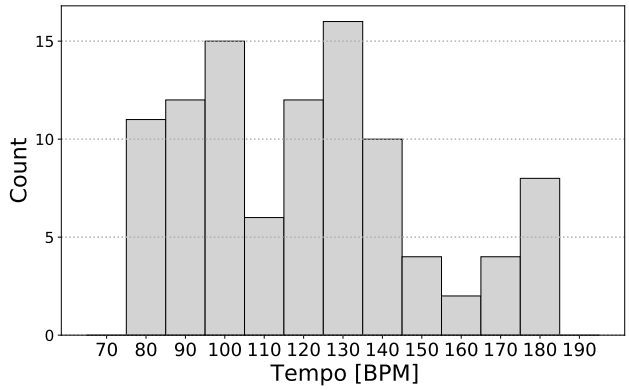

Fig. 2. Tempo histograms of our corpus. The songs range from 80 to 160 beats per minute (BPM).

III. Corpus specifications

A. Data statistics

The data size of the singing voice was larger than that of the speaking voice. The recording of the singing voice was 27.20 minutes long (18.68 minutes in voiced segments), and the recording of the speaking voice was 12.09 minutes long. Therefore, texts were shared between the singing voices and the speaking voices, but the duration of the singing voice was longer than that of speaking voice. This is consistent with existing work [17].

Figure 1 shows histograms of the keys of our corpus, and **Figure 2** shows histograms of the tempos of our corpus. As **Figure 1** shows, the tonics are well-balanced, while there are fewer songs in minor keys than in major keys. Moreover, as **Figure 2** shows, the tempos are distributed in a range between 80 to 160 beats per minute (BPM), indicating that this corpus may be unsuitable for synthesizing songs with extremely slow or fast tempos or in a minor key.

B. Phoneme balance and note intervals

In this section, we evaluate phonemes in our PJS corpus. Furthermore, we evaluate note intervals. For the comparison, we used Kiritan [13], one of the largest Japanese singing-voice corpora.

TABLE I
The number of unique monophones and diphones of our PJS corpus and the Kiritan database. Our PJS corpus contains more monophones and diphones than the Kiritan database.

Corpus	Monophone	Diphone
PJS (ours)	**34**	**349**
Kiritan	33	306

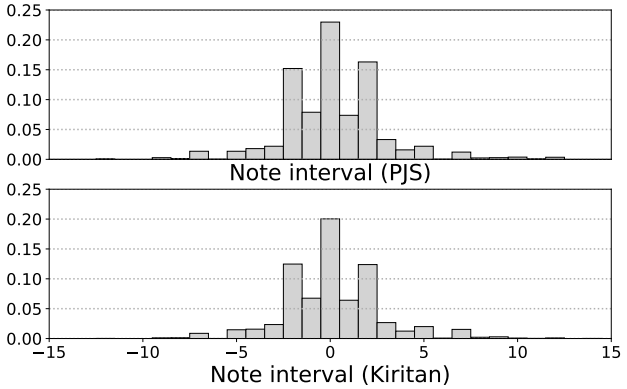

Fig. 3. Note-interval histograms of our PJS corpus and the Kiritan database. The histograms were normalized (i.e., the sum of values equals one in each histogram). Our PJS corpus has heavier tails than the Kiritan database.

Table I lists the number of unique monophones and diphones in the corpora. Our corpus contains more unique phonemes and diphones than the Kiritan database does. **Figure 3** shows note-interval histograms of two corpora. The PJS has heavier tails than the Kiritan, indicating that the melodies in our corpus have more variable note intervals than the Kiritan does.

C. Music score analysis: case study

Japanese songs typically use one musical note per Japanese syllable but not always. **Figure 4** is an example of such an exception, PJS100_001.xml. The multisyllabic notes *to-o*-ji and myo-o-*o-o* can be found on the first and second musical bars, respectively, where "-" indicates the syllable boundary. This means special processes (e.g., copying notes to each syllable [30]) are needed to train singing-voice synthesis systems.

D. Comparison of alignment methods

We evaluate the alignment accuracy of the following methods.

- **HMM**: Viterbi alignment of monophone-dependent five-state HMMs (1st paragraph in **Section II-D**)
- **DTW**: DTW between NEUTRINO-synthesized voices and recorded voices (1st half of 2nd paragraph in **Section II-D**)
- **DTW+VC**: DTW with voice conversion from a NEUTRINO-synthesized voice to a recorded voice (2nd

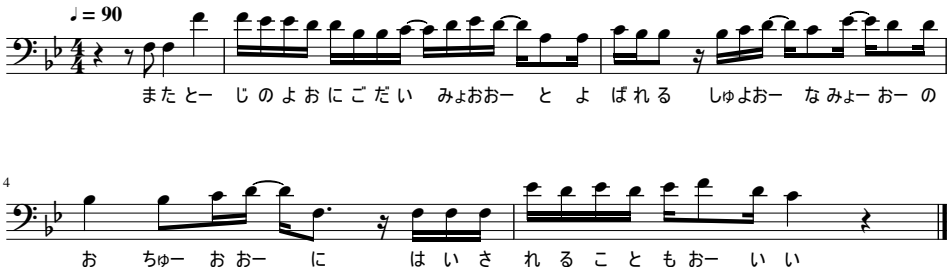

Fig. 4. Score of PJS100_001.xml. The lyrics are "mata tooji no yoo ni godai myoooo to yobareru shuyoo na myoooo no chuuoo ni haisareru koto mo ooi." Most (but not all) individual notes correspond to a single syllable. Some notes correspond to multiple syllables, such as *to-o*-ji on the first musical bar and myo-o-*o-o* on the second musical bar, where "-" indicates the syllable boundary.

TABLE II
SEPARATION METRIC OF ALIGNMENT METHODS. HIGHER IS BETTER.

Method	Separation metric
HMM	36.87
DTW (ours)	46.45
DTW-VC (ours)	**49.72**

half of 2nd paragraph in **Section II-D**)

Low-dimensional mel-cepstral coefficients were used as speech features. Since the reference alignment was unobservable, we could not calculate the distance between the reference and the resulting alignments (e.g., mean squared error between phoneme boundaries). Therefore, first, we segmented speech features following the resulting alignment, and we calculated a separation metric R defined as

$$R = \sum_{d=1}^{D} \frac{\sum_p \omega_p^{(d)} \left(\mu_p^{(d)} - \mu^{(d)}\right)^2}{\sum_p \omega_p^{(d)} {\sigma_p^{(d)}}^2}. \quad (1)$$

The subscript p indicates a phoneme label. d is a dimension index of a D-dimensional speech feature. $\mu_p^{(d)}$ is the mean, and $\sigma_p^{(d)}$ is the standard deviation. $\omega_p^{(d)}$ is the amount ratio of phoneme p: the number of frames of phoneme p divided by the total number of frames. These values are calculated from d-th dimension of speech features segmented following the resulting alignment. $\mu^{(d)}$ is the global mean calculated from the whole of speech features. The R can quantify the degree of separation of phonemes. When the resulting alignment can segment speech features for each phoneme accurately, intra-phoneme standard deviation (i.e., $\sigma_p^{(d)}$) becomes smaller, and R becomes larger.

Table II lists the results. "HMM" has the worst score, our proposed "DTW" has a better score, and "DTW+VC" has the best score. These results indicate that, 1) for singing voices, our DTW-based alignment is superior to HMM-based alignment (typically used for speaking voices) and 2) solving a speaker identity problem with voice conversion can enhance the performance.

IV. CONCLUSION

This paper presented the PJS corpus, a freely available phoneme-balanced Japanese singing-voice corpus. We confirmed the phoneme balance in our corpus by composing music based on a phoneme-balanced speaking-voice corpus. Our corpus consists of singing-voice data, parallel speaking-voice data, and the musical information that songs use. The result of the analysis indicated that our corpus was musically and phonetically well-balanced. Therefore, our corpus can contribute to research areas beyond singing-voice synthesis. Furthermore, we explored phoneme alignment methods and demonstrated that our method combined with DTW and voice conversion achieved the best accuracy. In our future work, we will add a variety of singing styles, such as falsetto and growl voices.

The PJS corpus is available on our project page [18]. All the data is licensed with the CC BY-SA 4.0 license.

ACKNOWLEDGMENT

Part of this research was supported by the GAP Foundation Program of the University of Tokyo and a JST PRESTO Grant Number JPMJPR18J8.

REFERENCES

[1] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, "Singing voice synthesis based on deep neural networks," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2478–2482.

[2] J. Kim, H. Choi, J. Park, M. Hahn, S. Kim, and J.-J. Kim, "Korean singing voice synthesis system based on an LSTM recurrent neural network," in *Proc. INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 1551–1555.

[3] Yu Gu, Xiang Yin, Yonghui Rao, Yuan Wan, Benlai Tang, Yang Zhang, Jitong Chen, Yuxuan Wang, and Zejun Ma, "ByteSing: A Chinese Singing Voice Synthesis System Using Duration Allocated Encoder-Decoder Acoustic Models and WaveRNN Vocoders," *arXiv preprint arXiv:2004.11012*, 2020.

[4] M. Blaauw and J. Bonada, "A neural parametric singing synthesizer modeling timbre and expression from natural songs," *Applied Sciences*, vol. 7, no. 12, Dec. 2017.

[5] "Sinsy," http://www.sinsy.jp/.

[6] "NEUTRINO," https://n3utrino.work/.

[7] "Synthesizer V," https://dreamtonics.com/en/synthesizerv/.

[8] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the MIR-1K dataset," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 310–319, Feb. 2010.

[9] Z. Duan, H. Fang, B. Li, K. C. Sim, and Y. Wang, "The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech," in *Proc. APSIPA ASC*, Kaohsiung, Taiwan, Oct. 2013, pp. 1–8.

[10] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *Proc. LVA/ICA*, Liberec, Czech Republic, Aug. 2017, pp. 323–332, Springer International Publishing.
[11] M. Goto and T. Nishimura, "AIST Humming Database: Music database for singing research," *The Special Interest Group Notes of IPSJ (MUS)*, vol. 82, pp. 7–12, Aug. 2005 (in Japanese).
[12] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Popular, classical, and jazz music databases," in *Proc. ISMIR*, Paris, France, Oct. 2002, vol. 2, pp. 287–288.
[13] M. Morise, "Tohoku Kiritan singing voice corpus," https://zunko.jp/kiridev/login.php.
[14] "HMM-based speech synthesis system (HTS)," http://hts.sp.nitech.ac.jp/.
[15] H. Tamaru, S. Takamichi, N. Tanji, and H. Saruwatari, "JVS-MuSiC: free Japanese multispeaker singing-voice corpus," *arXiv preprint 2001.07044*, Jan. 2020.
[16] y_benjo and MagnesiumRibbon, "Voice actress corpus," http://voice-statistics.github.io.
[17] Y. Ohishi, M. Goto, K. Itou, and K. Takeda, "Discrimination between singing and speaking voices," in *Proc. EUROSPEECH*, Lisbon, Portugal, Sep. 2005, pp. 1141–1144.
[18] "PJS: Phoneme-balanced japanese singing voice corpus," https://sites.google.com/site/shinnosuketakamichi/research-topics/pjs_corpus.
[19] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, "DeepS]inger: Singing Voice Synthesis with Data Mined From the web," *arXiv preprint arXiv:2007.04590*, 2020.
[20] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*, 2020, pp. 6189–6193.
[21] K. Tokuda, Y Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.
[22] G. Kotani, H. Suda, D. Saito, and N. Minematsu, "Experimental investigation on the efficacy of affine-dtw in the quality of voice conversion," in *Proc. APSIPA ASC*, 2019, pp. 119–124.
[23] S. Fukayama, K. Nakatsuma, S. Sako, T. Nishimoto, and S. Sagayama, "Automatic song composition from the lyrics exploiting prosody of japanese language," in *Proc. SMC*, 01 2010, pp. 299–302.
[24] Lewitt, "440 FLEX," https://www.lewitt-audio.com/microphones/lct-recording/lct-441-flex.
[25] JZ MICROPHONE, "Pop filter," https://intshop.jzmic.com/collections/accesories/products/pop-filter.
[26] RME, "Fireface UCX," https://www.rme-audio.de/fireface-ucx.html.
[27] Y. Wang, R. J. S.-Ryan, D. Stanton, Y. Wu, Ron J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 4006–4010.
[28] S. Jose, M. Soroush, K. Kundan, S. João F., K. Kyle, C. Aaron, and B. Yoshua, "Char2Wav: End-to-end speech synthesis," in *International Conference on Learning Representations (Workshop Track)*, April 2017.
[29] F. Brugnara, D. Falavigna, and M. Omologo, "Automatic segmentation and labeling of speech based on Hidden Markov Models," *Speech Communication*, vol. 1, no. 4, pp. 357–370, 1993.
[30] K. Nakamura, K. Oura, Y. Nankaku, and K. Tokuda, "HMM-based singing voice synthesis and its application to Japanese and English," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 265–269.
[31] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Proc. ICASSP*, 1988, vol. 1, pp. 655–658.
[32] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.
[33] K. Kobayashi and T. Toda, "sprocket: Open-Source Voice Conversion Software," in *Proc. Odyssey*, June 2018, pp. 203–210.