

PJS: phoneme-balanced Japanese singing voice corpus

Junya Koguchi¹ and Shinnosuke Takamichi²

¹ Meiji University, Japan.

² Graduate School of Information Science and Technology, The University of Tokyo, Japan.

cs202027@meiji.ac.jp, shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

Abstract

This paper presents a free Japanese singing voice corpus that can be used for highly applicable and reproducible singing voice synthesis research. A singing voice corpus helps develop singing voice synthesis, but existing corpora have two critical problems: data imbalance (singing voice corpora do not guarantee phoneme balance, unlike speaking-voice corpora) and copyright issues (cannot legally share data). As a way to avoid these problems, we constructed a PJS (phoneme-balanced Japanese singing voice) 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. This paper describes how we built the corpus.

Index Terms: Singing voice corpus, singing voice synthesis, music information processing, phoneme balance

1. 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], and auto-regressive types [3]), and many products have been launched (e.g., Sinsy [4] and NEUTRINO [5]).

Freely available singing voice corpora contribute to applicable and reproducible singing voice synthesis research. Corpora are being developed in many languages (e.g., Chinese [6], English [7], etc. [8]). The leading Japanese corpus, the large RWC Music Database [9, 10], 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 [11] was designed for singing voice synthesis. The corpus contains a selection of 50 songs made up of childrens songs and anime songs. By comparing these corpora, we aim to develop a smaller corpus for easy-to-train machine learning. The HTS demo [12] and JVS-MuSiC [13] examples never guarantee phoneme balance, which is an important factor in creating a smaller corpus. Phoneme imbalance typically results in phonetic lack in synthesized singing voices.

This paper describes the construction of a phoneme-balanced singing voice corpus named the phoneme-balanced Japanese singing voice (*PJS*). Using the Voice Actress Corpus [14], a phoneme-balanced speaking voice corpus, we composed melodies for 100 sentences. Additionally, our corpus contributes the following:

Singing and speaking voices: We recorded both singing voices and parallel speaking voices. This paired data contributes to speaking-singing research (e.g., [15]).

Descriptions of compositions: We noted descriptions of melody compositions. These descriptions contribute to natural-language-based music information research.

CC BY-SA 4.0 license: All the data in our corpus is licensed with CC BY-SA 4.0. Therefore, our corpus is available for both research and commercial use, unlike existing corpora [6, 7, 8, 9, 10, 11, 12].

Availability online: Our corpus can be freely downloaded from our project page [16].

The following sections describe the details of the corpus.

2. Corpus design

2.1. 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 [14].

```
├── 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 [14] as the lyric
- $PJS100_ [SENTENCE_ID]_speech.wav$: speaking voice that utters a sentence from the phoneme-balanced speaking-voice corpus [14]
- $PJS100_ [SENTENCE_ID].mid$: 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 [14]. The following sections describe the composition and recording conditions.

2.2. 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. 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.

2.3. 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 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) [17] microphone, a JZ MICROPHONES Pop Filter [18] windscreen, and an RME Fireface UCX [19] audio interface.

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

3. Corpus specifications

3.1. 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, and the recording of the speaking voice was 12.09 minutes long. Therefore, texts are shared between singing and speaking voices, but the duration of the singing voice is longer than that of speaking voice. This is consistent with existing work [15].

Figure 1 and **Figure 2** show histograms of the keys and tempos of our corpus, respectively. 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.

3.2. Music score analysis

Japanese songs typically use one musical note per Japanese syllable but not always. **Figure 3** 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 [20]) are needed to train singing voice synthesizers.

4. 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. Therefore, our corpus can contribute to research areas beyond singing voice synthesis. In our future work, we will add a variety of singing styles, such as falsetto and growl voices.

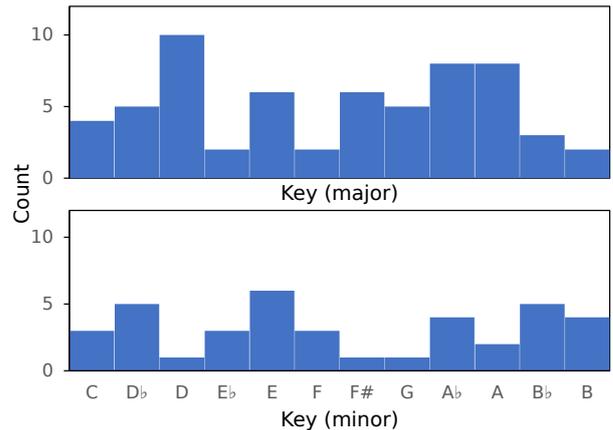


Figure 1: *Key histogram of our corpus. There are fewer songs in minor keys than in major keys.*

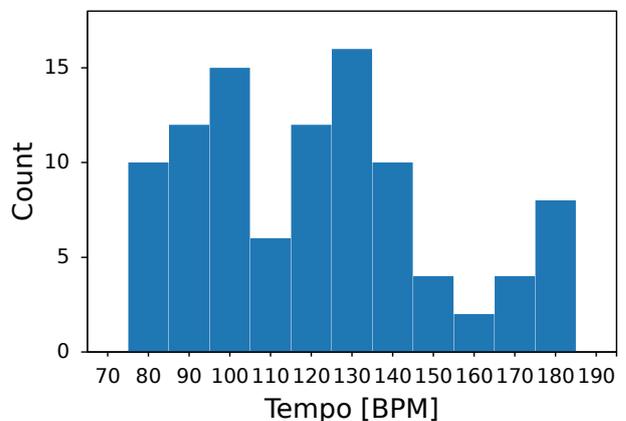


Figure 2: *Tempo histogram of our corpus. The songs only range from 80 to 160 beats per minute (BPM).*

The PJS corpus is available on our project page [16]. All the data is licensed with the CC BY-SA 4.0 license.

Acknowledgements: Part of this research was supported by the GAP Foundation Program of the University of Tokyo.

5. 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] 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.
- [4] “Sinsy,” <http://www.sinsy.jp/>.
- [5] “NEUTRINO,” <https://n3utrino.work/>.

♩ = 90

またとー じのよおにごだい みよおおー とよばれる しゅよあー なみよーあー の

4
お ちゅー おあー に はいされることもあー いい

Figure 3: 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.

- [6] 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.
- [7] 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.
- [8] 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*, Cham, Aug. 2017, pp. 323–332, Springer International Publishing.
- [9] 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.
- [10] 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.
- [11] M. Morise, “Tohoku Kiritan singing voice corpus,” <https://zunko.jp/kiridev/login.php>.
- [12] “HMM-based speech synthesis system (HTS),” <http://hts.sp.nitech.ac.jp/>.
- [13] H. Tamaru, S. Takamichi, N. Tanji, and H. Saruwatari, “JVS-MuSiC: free Japanese multispeaker singing-voice corpus,” *arXiv preprint 2001.07044*, Jan. 2020.
- [14] y_benjo and MagnesiumRibbon, “Voice actress corpus,” <http://voice-statistics.github.io>.
- [15] 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.
- [16] “PJS: Phoneme-balanced japanese singing voice corpus,” https://sites.google.com/site/shinnosuketakamichi/research-topics/pjs_corpus.
- [17] Lewitt, “440 FLEX,” <https://www.lewitt-audio.com/microphones/lct-recording/lct-441-flex>.
- [18] JZ MICROPHONE, “Pop filter,” <https://intshop.jzmic.com/collections/accessories/products/pop-filter>.
- [19] RME, “Fireface UCX,” <https://www.rme-audio.de/fireface-ucx.html>.
- [20] 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.