JVS corpus: free Japanese multi-speaker voice corpus

Shinnosuke Takamichi, Kentaro Mitsui, Yuki Saito,
Tomoki Koriyama, Naoko Tanji, and Hiroshi Saruwatari

Graduate School of Information Science and Technology, The University of Tokyo,
7-3-1 Hongo Bunkyo-ku, Tokyo 133–8656, Japan.

`shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp`

Abstract

Thanks to improvements in machine learning techniques, including deep learning, speech synthesis is becoming a machine learning task. To accelerate speech synthesis research, we are developing Japanese voice corpora reasonably accessible from not only academic institutions but also commercial companies. In 2017, we released the JSUT corpus, which contains 10 hours of reading-style speech uttered by a single speaker, for end-to-end text-to-speech synthesis. For more general use in speech synthesis research, e.g., voice conversion and multi-speaker modeling, in this paper, we construct the JVS corpus, which contains voice data of 100 speakers in three styles (normal, whisper, and falsetto). The corpus contains 30 hours of voice data including 22 hours of parallel normal voices. This paper describes how we designed the corpus and summarizes the specifications. The corpus is available at our project page[1].
Index Terms: voice corpus, Japanese, speech synthesis

1. Introduction

Thanks to developments in deep learning techniques, studies on speech have been targeted actively [1, 2, 3, 4]. Nowadays, speech synthesis, e.g., text-to-speech, singing voice synthesis, voice conversion, and speech coding, is becoming a machine learning task. Easily accessible voice corpora help to not only accelerate speech-related research but also improve the reproductivity of a study. In 2017, we released a large-scaled Japanese speech corpus, named the JSUT corpus [5], for end-to-end text-to-speech synthesis. The corpus included 10 hours of reading-style speech data uttered by a single native Japanese speaker and all pronunciations of daily-use characters and individual readings in Japanese [6]. Since Oct. 2017, the project page [7] was accessed more than 6,000 times (75% from Japan and 25% from foreign countries) from more than 60 countries. We believe that the JSUT corpus has become one of the most used Japanese corpora for modern speech synthesis research [8, 9].

Towards more general purposes of speech-related research, this paper introduces a new Japanese voice corpus, named the *JVS* (Japanese versatile speech) corpus. The corpus is designed to have many benefits for many types of users as follows.

High-quality format: The audio files are sampled at 24 kHz, encoded at 16 bit, and formatted in RIFF WAV.

High-quality recording: The recordings were controlled by a professional sound director and done in a recording studio.

Many speakers: The corpus includes 100 native Japanese speakers, and all of the speakers are professional, e.g., voice actor/actress.

Many styles: Each speaker utters not only normal speech but also whisper and falsetto voices.

Large in scale: In total, the corpus contains 30 hours of voice data.

Parallel/non-parallel utterances: Each speaker utters parallel, i.e., common among speakers, and non-parallel, i.e., completely different among speakers, utterances.

Many tags: The corpus includes not only voice data but also transcriptions, gender information, F_0 ranges, speaker similarity, and phoneme alignments.

Free for research: The corpus is free to use for research in academic institutions and commercial companies.

Easily accessible: The corpus is freely downloadable online.

The next section describes how we designed the corpus.

2. Corpus design

The corpus consists of the following four sub-corpora. Their names are formatted as *[NAME][NUM_UTT]*. *[NUM_UTT]* indicates the number of utterances per speaker.

parallel100: 100 parallel normal (reading-style) utterances
nonpara30: 30 non-parallel normal utterances
whisper10: 10 whisper utterances
Falsetto10: 10 falsetto utterances

The directory structures of the corpus are listed below. The speaker name is formatted as *jvs[SPKR_ID]*. *[SPKR_ID]* indicates the speaker ID with the range of 1 through 100.

```
□ jvs001
  └─□ parallel100
      ├─□ wav24kHz16bit
      ├─□ lab
      └─▤ transcripts_utf8.txt
  └─□ nonpara30
      ├─□ wav24kHz16bit
      ├─□ lab
      └─▤ transcripts_utf8.txt
  └─□ whisper10
      ├─□ wav24kHz16bit
      └─▤ transcripts_utf8.txt
  └─□ falsetto10
      ├─□ wav24kHz16bit
      └─▤ transcripts_utf8.txt
□ jvs002
□ ...
□ jvs100
▤ speaker_similarity_male.csv
▤ speaker_similarity_female.csv
▤ duration.txt
▤ gender_f0range.txt
```

2.1. Sub-corpora

This section describes how we designed the four sub-corpora.

2.1.1. parallel100

Parallel voices, i.e., utterances that are common among speakers, are used for voice conversion [10, 11], speaker factorization [12], multi-speaker modeling [8], and so on. We used 100 phonetically-balanced sentences of the sub-corpus "voiceactress100"[2] of the JSUT corpus [5], and we let speakers utter the sentences. This corpus contains not only the audio files but also the transcriptions (stored in "parallel100/transcript_utf8.txt") and phoneme alignment (stored in "parallel100/lab").

2.1.2. nonpara30

The use of non-parallel voices, i.e., utterances that are completely different among speakers, is a challenging but more realistic situation than that of parallel voices. Sentences to be uttered are randomly selected from the JSUT corpus excluding its sub-corpus "voiceactress100." Each speaker uttered 30 utterances that are different among speakers. This sub-corpus also includes transcriptions and phoneme alignments. Note that, the sentences are not phonetically balanced unlike the sub-corpora "parallel100."

2.1.3. whisper10

Whispering is used to quietly communicate, i.e., convey secret information without being overheard. Analysis [14], synthesis [15], recognition [16] and conversion [17] of whispered voices have the potential to augment our silent-speech communication. The first five sentences of this sub-corpus are the same as those of the sub-corpus "parallel100," and they are parallel among speakers. The remaining five sentences are the same to those of the sub-corpus "nonpara30," and they are non-parallel among speakers. Namely, ten utterances per speaker are parallel between whispered voices and normal voices.

2.1.4. Falsetto10

Falsetto is a vocal register occupying the F_0 range that is higher than normal voices. The physiology of falsetto is different from that of normal voices [18], and the analysis and synthesis of falsetto are remaining tasks for signal processing-based vocoders. The first five sentences of this sub-corpus are the same as those of the sub-corpus "parallel100." The remaining five sentences are the same as those of the sub-corpus "nonpara30" but different to those of the sub-corpus "whisper10." Namely, five utterances are parallel among speakers, ten are parallel between normal voice and falsetto, and five are parallel between whisper and falsetto.

2.2. Tags

This section describes some of the annotation results.

- F_0 **range (gender_f0range.txt)**: Typical pitch extractors, e.g., [19, 20, 21], have a range for F_0 search, and the setting is critical for the results ultimately obtained for the voices. This corpus contains manually annotated

F_0 ranges per speaker for his/her normal voices.

- **Speaker similarity (speaker_similarity_*.csv)**: Perceptual similarity between speakers is useful for selecting speakers (or models) [22] and modeling speaker space [23]. This corpus contains perceptual similarity scores between all pairs of speakers of each gender.

- **Duration (duration.txt)**: Duration, i.e., data size, and speech rate are also included. Phoneme-level duration is calculated from the results of phoneme alignments.

3. Results of data collection

3.1. Corpus specs

We hired 100 native Japanese professional speakers, which included 49 male and 51 female speakers. Their voices were recorded in a recording studio. Recording for each speaker was done within one day. The recordings were controlled by a professional sound director. The voices were originally sampled at 48 kHz and downsampled to 24 kHz by SPTK [24]. The 16-bit/sample RIFF WAV format was used. Sentences (transcriptions) were encoded in UTF-8. The full context and monophone labels were automatically generated by Open JTalk [25]. The phoneme alignments were automatically generated by Julius [26]. F_0 ranges were manually annotated in accordance with hands-on voice conversion [27]. The WORLD vocoder [20, 28] extracted F_0. Commas were added between breath groups. For annotating perceptual similarity scores, we followed Saito et al.'s study [23] and used a crowdsourcing service, Lancers [29], which is a famous crowdsourcing service in Japan. Each listener scored the perceptual similarity for each pair of speakers from -3 (completely different) and $+3$ (very similar). A final score for each speaker pair was obtained by averaging listeners' scores. Ten different listeners scored each speaker pair, and 1,000 listeners participated in total.

3.2. Analysis

3.2.1. Duration

Table 1 lists the statistics for speaker-wise duration. This corpus contains 26 hours of normal voices and 4 hours of other-style voices. Each speaker uttered approximately 15.7 minutes of normal voices, 1.24 minutes of whispered voices, and 1.18 minutes of falsetto. In the sub-corpus "parallel100," the transcription was common among speakers, but the duration was very different; speaker "jvs084" uttered 1.8 times slower than speaker "jvs020."

3.2.2. Perceptual speaker similarity

Fig. 1 shows matrices of perceptual similarity scores. For example, the most similar pair was "jvs019" and "jvs096." Also, a speaker that was most dissimilar from the other speakers was "jvs010."

4. Conclusion

In this paper, we constructed a corpus named the JVS corpus. The corpus was designed for speech-related research using multi-speaker and multi-style voices. Text data of the corpus is licensed as shown in the LICENCE file in the JSUT corpus [7]. The tags are licensed with CC BY-SA 4.0. The audio data may be used for

- Research by academic institutions

---

[2]The original sentences are included in the Voice Actress Corpus [13], and the one included in the JSUT corpus had commas added at the phrase break positions.

Table 1: *Speaker-wise duration statistics. Silence parts were included to calculate these values.*

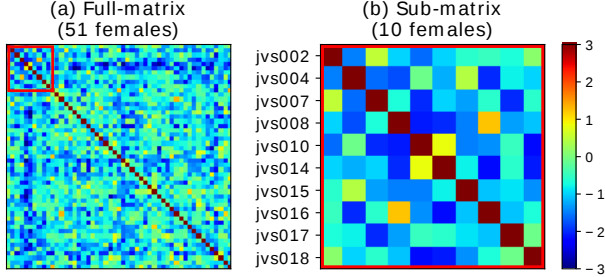|	Mininum [min.]	Average [min.]	Maximum [min.]	Total (100 speaker) [hour]
parallel100 (100 utterances)	10.11 (jvs020)	13.11	18.24 (jvs084)	22
nonpara30 (30 utterances)	2.12 (jvs099)	2.62	3.86 (jvs036)	4.4
whisper10 (10 utterances)	0.95 (jvs045)	1.24	1.69 (jvs018)	2.0
falsetto10 (10 utterances)	0.90 (jvs045)	1.18	1.61 (jvs035)	2.0
Total	-	-	-	30.4

Figure 1: *Speaker similarity matrix of 51 Japanese females and (b) its sub-matrix obtained by large-scale subjective scoring.*

- Non-commercial research, including research conducted within commercial organizations
- Personal use, including blog posts.

Our project page at `https://sites.google.com/site/shinnosuketakamichi/research-topics/jvs_corpus` describes the terms for commercial use.

5. Acknowledgements

6. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine of IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[2] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," vol. abs/1609.03499, 2016.

[3] S. Takamichi, K. Tomoki, and H. Saruwatari, "Sampling-based speech parameter generation using moment-matching network," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017.

[4] Y. Saito, S. Takamichi, and H. Saruwatari, "Statistical parametric speech synthesis incorporating generative adversarial networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 755–767, Jun. 2018.

[5] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," vol. arXiv abs/1711.00354, 2017.

[6] Governments of Japan Agency for Cultural Affairs, "List of daily-use kanjis `http://www.bunka.go.jp/kokugo_nihongo/sisaku/joho/joho/kijun/naikaku/kanji/index.html`," 2010.

[7] "JSUT: Japanese speech corpus of Saruwatari Lab, the University of Tokyo corpus," `https://sites.google.com/site/shinnosuketakamichi/publication/jsut`.

[8] S. Ueno, M. Mimura, S. Sakai, and Tatsuya Kawahara, "Multi-speaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition," in *Proc. ICASSP*, Brighton, United Kingdom, May 2019, pp. 6161–6165.

[9] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using dual supervised adversarial networks with continuous wavelet transform F0 features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 10, pp. 1535–1548, Oct. 2019.

[10] 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.

[11] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, Mar. 1998.

[12] H. Lu and S. King, "Factorized context modeling for Text-to-Speech synthesis," in *Proc. ICASSP*. Vancouver, Canada, May 2013.

[13] y_benjo and MagnesiumRibbon, "Voice-actress corpus," `http://voice-statistics.github.io/`.

[14] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.

[15] V. A. Petrushin, L. I. Tsirulnik, and V. Makarova, "Whispered speech prosody modeling for TTS synthesis," Chicago, U.S.A., May 2010.

[16] S.-C. Jou, T. Schultz, and A. Waibel, "Whispery speech recognition using adapted articulatory features," in *Proc. ICASSP*, Philadelphia, U.S.A., Mar. 2005, vol. 1, pp. 1009–1012.

[17] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, Nov. 2012.

[18] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception," *The Journal of the Acoustical Society of America*, vol. 90, no. 5, pp. 2394–2410, Nov. 1991.

[19] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigne, "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*, vol. 27, no. 3–4, pp. 187–207, 1999.

[20] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[21] D. Talkin, "REAPER: Robust Epoch And Pitch EstimatoR," `https://github.com/google/REAPER`.

[22] P. Lanchantin, Mark J.F. Gales, S. King, and J. Yamagishi, "Multiple-average-voice-based speech synthesis," in *Proc. ICASSP*, Florence, Italy, May 2014, pp. 285–289.

[23] Y. Saito, S. Takamichi, and H. Saruwatari, "DNN-based speaker embedding using subjective inter-speaker similarity for multi-speaker modeling in speech synthesis," in *Proc. SSW10*. Vienna, Austria, Sep. 2019.

[24] "Speech signal processing toolkit (SPTK)," `http://sp-tk.sourceforge.net/`.

[25] "Open jtalk," `http://open-jtalk.sourceforge.net/`.

[26] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine," in *Proc. EU-ROSPEECH*, Aalborg, Denmark, Sep. 2001, pp. 1691–1694.

[27] T. Toda, "Hands on voice conversion," `https://www.slideshare.net/NU_I_TODALAB/hands-on-voice-conversion`, 2018.

[28] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[29] "Lancers," `http://www.lancers.jp`.