JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research

Shinnosuke Takamichi*, Ryosuke Sonobe†, 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

(Received 27 August 2019, Accepted for publication 8 January 2020)

Abstract: In this paper, we develop two corpora for speech synthesis research. Thanks to improvements in machine learning techniques, including deep learning, speech synthesis is becoming a machine learning task. To accelerate speech synthesis research, we aim at developing Japanese voice corpora reasonably accessible from not only academic institutions but also commercial companies. In this paper, we construct the JSUT and JVS corpora. They are designed mainly for text-to-speech synthesis and voice conversion, respectively. The JSUT corpus contains 10 hours of reading-style speech uttered by a single speaker, and the JVS corpus contains 30 hours containing three styles of speech uttered by 100 speakers. This paper describes how we designed the corpora and summarizes the specifications. The corpora are available at our project pages.

Keywords: Voice corpus, Speech synthesis, Text-to-speech, Voice conversion

PACS number: 43.72.Ja [doi:10.1250/ast.41.761]

1. INTRODUCTION

Thanks to developments in deep learning techniques, studies on speech have been targeted actively [1–4]. Speech synthesis, e.g., text-to-speech, singing voice synthesis, voice conversion, and speech coding, is becoming a machine learning task and a popularized technology. Nowadays, many speech synthesis systems are developed by not only speech researchers but also machine learning researchers and non-researchers.

In Japan, since around 1990, a variety of Japanese voice corpora, e.g., ATR503 [5] and JNAS [6], have been developed and managed by national institutions. However, they are out of touch with modern speech synthesis research, as described below.

Low accessibility: Reproducibility of a study is an important perspective in machine learning. However, accessibility of the existing managed corpora is low for foreign researchers and machine learning researchers.

Limited specifications: Though traditional statistical approaches to text-to-speech, e.g., hidden Markov model-based ones [7], require voice data of less than one hour, deep learning-based approaches [2,8] require larger amounts of data, e.g., more than 10 hours [2]. However, there was no corpus that has such a larger size speech for Japanese speech synthesis. Also, voice data in the existing corpora is sampled at low sampling frequency (e.g., 16 kHz), but modern speech synthesis requires the higher sampling rate [9].

Limited use: Deep learning [1] and its toolkits (e.g., chainer [10]) benefit from not only an improved quality of synthetic speech but also accelerated research-production cycles. However, the use of the existing corpora is limited for research purposes.

To overcome the above limitations, we constructed two Japanese voice corpora, named the JSUT (Japanese speech corpus of Saruwatari lab of the University of Tokyo) corpus and the JVS (Japanese Versatile Speech) corpus. The corpora are designed to have many benefits as follows.

High accessibility: The corpora are freely downloadable online (JSUT [11] and JVS [12]).

Rich specifications: We employ a high sampling rate (24 or 48 kHz), an uncompressed audio format (RIFF WAV), and recording in quiet environments (anechoic room or recording studio). Also, a large amount of
voice data is included; 10 hours in the JSUT corpus, and 30 hours in the JVS corpus.

**Flexible use:** The corpora are free to use for research at academic institutions and commercial companies. Also, they are available for commercial use (see Sect. 5). The JSUT corpus designed in Sect. 2 is mainly for end-to-end Japanese text-to-speech synthesis [13]. The corpus includes 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 [14]. The JVS corpus designed in Sect. 3 is mainly for voice conversion [4, 15]. The corpus contains 30 hours of parallel/non-parallel multi-style speech uttered by 100 speakers. This paper describes how we designed each corpus. Sections 2 and 3 describe the designs of the JSUT and JVS corpora, respectively. Section 4 presents the results of data collection. We conclude in Sect. 5 with a brief summary.

## 2. DESIGNS OF JSUT CORPUS

For end-to-end text-to-speech synthesis, the main purpose of the JSUT corpus is to cover all of the main pronunciations of daily-use Japanese characters, not to cover intermediate representations such as phonemes as is used in existing corpora [5]. The corpus includes the following nine sub-corpora. Their names are formatted as \[ \text{[NAME]}\text{[NUM UTT]} \]. \text{[NUM UTT]} indicates the number of utterances of the sub-corpora.

- **basic5000:** utterances covering all of the main pronunciations of daily-use Japanese characters.
- **countersuffix26:** utterances including individual readings of counter suffixes.
- **loanword128:** utterances including loanwords, e.g., verbs or nouns.
- **utparaphrase512:** utterances for which a word or phrase of a piece of text is replaced with its paraphrase.
- **voiceactress100:** parallel speech for a free corpus of Japanese voice actresses [16].
- **onomatopee300:** utterances including famous Japanese onomatopee (onomatopoeia).
- **repeat500:** repeatedly spoken utterances.
- **travel1000:** travel-domain utterances.
- **precedent138:** precedent-domain utterances.

The directory structures of the corpus are listed below.

2.1. **Sub-corpora**

This section describes how we designed the nine sub-corpora. The transcription of the sub-corpuses is saved in transcripts_utf8.txt in each directory.

2.1.1. **basic5000**

This is the main sub-corpora of the JSUT corpus. In Japanese, 2136 kanji characters (kanji are the logographic characters used in the modern Japanese writing system) are officially defined as daily-use characters [14], and each character has individual pronunciations consisting of its individual onyomi (Chinese readings) and kunyomi (Japanese readings). For example, we pronounce “一” (one in English) as “ichi,” “itsu,” “hito,” and “hito (tsu).” We collected 5000 sentences from Wikipedia [17] and the TANAKA Corpus [18] so that all pronunciations of the daily-use kanji characters could be covered. The first 4,000 sentences of the subset BASIC5000 were randomly selected from the TANAKA corpus. Next, the greedy algorithm was used for making a minimum set of sentences from Wikipedia sentences to cover readings not covered by the 4,000 sentences. Some of the pronunciations cannot be found in these corpora, therefore, we manually made additional sentences to cover the remaining readings. Examples of the sentences and phonemes are as follows.

1) どこで邦貨をドルに変えられるか。
   (dokode hokkaodoro nikaihara reku)

2) これを付けている人間は、激しい寂しさに襲われる。
   (koreotsuketeruningen wa puchage shiiseki yokokann sosowaru)

2.1.2. **countersuffix26**

In Japanese, numerals cannot quantify nouns by themselves, and the pronunciation of the numerals changes depending on the suffix. For example, “二” (two in English) is pronounced “ni” with “個” (ko) as the suffix and “ふた” with “つ” (tsu). We crowdsourced 26 sentences including such counter suffixes. Examples of the sentences are as follows.

1) 一つ、二つ、三つ、四つ、五つ、六つ、七つ、八つ、九つ、とお、いくつ。
   (hitotsu tsu pautatsu pamu micltsu pauyou cltsu pautuitsutus pau mucltsu pauanatatsu pauyacltsu paukokonotsu pauotto pau i kutsu)

2.1.3. **loanword128**

Japanese sentences spoken daily have many loanwords,
e.g., verbs and nouns. For example, “ググる (guguru)” is a verb meaning to Google, and “ディズニー (dyizunii)” means Disney. The pronunciations and accents of loan-words are a curious task in spoken language processing [19]. We crowdsourced such words and sentences. We first instructed crowdworkers to answer a loan word that they know, and we listed loan words. Then, we instructed other crowdworkers to choose one from the list and write a sentence containing the word. Also, we collected sentences from Wikipedia that included pronunciations not included in the modern Japanese system, for example, sentences that had a Japanese-accented foreign proper name. We searched such pronunciations written by katakana, using the search engine of Wikipedia. Examples of the sentences are as follows.

1) 好きだった子に急に振られて，ガククリ来た。
   (sukidacltakonikyuunifuraretepau
gacakurikita)
2) もう終わった事に，いつまでもゴチャゴチャとどうるさ
   (mooowaclatkotoinpaitsumadem
gochagachatourusai)

2.1.4. utparaphrase512
Paraphrasing, e.g., lexical simplification, is a technique that substitutes a word or phrase into another sentence [20,21]. It can support the reading comprehension of a wide range of readers in speech communication. The SNOW E4 corpus [21,22] includes sentences and a list of their paraphrased words. We chose one paraphrased word per sentence, and constructed 256 sentences and paraphrased sentences. The total number of sentences was 512. Examples of the sentences are as follows.

1) 専門には，深いんだから。
   (senmonniwapautoinnakara)
2) 専門には，詳しくないんだから。
   (senmonniwapaukuswashikunainaka
ara)

2.1.5. voiceactress100
The Voice Actress Corpus [16] is a free speech corpus of professional Japanese voice actresses and includes not only neutral but also emotional voices. Collecting parallel speech for this speech corpus is very helpful in building attractive and emotional speech synthesis systems. We used sentences from this corpus and manually modified the pause positions. Examples of the sentences are as follows.

1) 一方で，漁業と商業で，りャネス港は繁栄していた。
   (ilocdepaugyogootshoogyode
pauryanesukooawahaneeshiteita)
2) ギレスピーは，マッギーを通じて，イネスと知り合っ
   (greesupiiwapuamadgiotsuujite
paunesutohiracita)

2.1.6. Onomatoope300
Onomatoope (onomatopoeia) has an important role in connecting speech and non-speech sounds in nature, and Japanese is rich in onomatoopeia words. We crowdsourced 300 sentences having individual onomatoopeia words. Examples of the sentences are as follows.

1) 好きだった子に急に振られて，ガククリ来た。
   (sukidacltakonikyuunifuraretepau
gacakurikita)
2) もう終わった事に，いつまでもゴチャゴチャとどうるさ
   (mooowaclatkotoinpaitsumadem
gochagachatourusai)

2.1.7. repeat500
Human speech production is not deterministic, i.e., speech waveforms always differ even if we try to reproduce the same linguistic and para-linguistic information. Some studies [3,23] are addressing speech synthesis that can reproduce natural inter-utterance variation. To quantify the variation, we recorded utterances spoken repeatedly by a single speaker. The speaker made utterances 5 times for each of the 100 sentences of the Voice Actress Corpus [16].

2.1.8. travel1000 and precedent138
We further constructed sentences whose domain differed from the above corpora. One thousand travel-domain sentences were collected from English-Japanese Translation Alignment Data [24]. Also, 138 copyright-free precedent sentences were collected from [25]. The words and phrases of the precedent sentences were significantly different from the above corpora. Also, some sentences were too difficult to read. Therefore, we manually removed and modified these sentences to make reading easier. Examples of the sentences are as follows.

1) それは, 私の行くエルロイハウスの近くです。
   (sorewamawatashinoikueruoiha
usunochikakudesu)
2) 後掲証拠等によれば, 以下の事実が認められる。
   (kookeeshookotooniyorebapauika
nojjitsugamiteraru)

2.2. Tags
As described in Sect. 4, the voice data was recorded during a half year. Voice data recorded during such long periods causes there to be objective and subjective differences among recording days [26]. Therefore, “recording_info.txt” in each directory lists what day the voice data was recorded.

3. DESIGNS OF JVS CORPUS
For voice conversion, the main purpose of the JVS corpus is to collect parallel/non-parallel voice data of multiple speakers in multiple styles. The corpus consists of the following four sub-corpora. Their names are formatted
as [NAME][NUM_UTT].
paral100: 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.

3.1. Sub-corpora

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

3.1.1. parallel100

Parallel voices, i.e., utterances that are common among speakers, are used for voice conversion [15,27], speaker factorization [28], multi-speaker modeling [13], and so on. We used 100 phonetically balanced sentences of the sub-corpus “voiceactress100” of the JSUT corpus, 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”).

3.1.2. 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 were 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 “paral100.”

3.1.3. whisper10

Whispering is used to quietly communicate, i.e., convey secret information without being overheard. The analysis [29], synthesis [30], recognition [31] and conversion [32] of whispered voices have the potential to augment silent-speech communication. The first five sentences of this sub-corpus are the same as those of the sub-corpus “paral100,” and they are parallel among speakers. The remaining five sentences are the same as 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.

3.1.4. 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 [33], 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 “paral100.” The remaining five sentences are the same as those of the sub-corpus “nonpara30” but different from 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.

3.2. Tags

This section describes some of the annotation results.

$F_0$ range (gender_f0range.txt): Typical pitch extractors, e.g., [34–36], 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) [37] and modeling the speaker space [38]. 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.

4. RESULTS OF DATA COLLECTION

4.1. JSUT Corpus

4.1.1. Corpus specs

We hired a female native Japanese speaker and recorded her voice in our anechoic room. She was not a
professional speaker but had experience working with her voice. The recordings were made in February, March, September, and October of 2017 for a few hours each day. The speaker made the recordings herself with our recording system. The speech data was sampled at 48 kHz. We used Lancers [39] as the crowdsourcing service. The 16bit/sample RIFF WAV format was used. Commas were added between breath groups. The positions of the commas were manually annotated.

4.1.2. Analysis

Table 1 lists the duration of each sub-corpus. The JSUT corpus contains 10 hours of normal voices including 6.77 hours in the basic corpus (“basic5000”).

We analyzed the linguistic and speech information of the constructed corpus. Note that not all of the data was used for the analysis to shorten the computation time. First, we counted the number of moras (sub-syllables) and words within one utterance by using MeCab [40] and NEologd [41,42]. Utterance length is an important factor in speech synthesis using sequence-to-sequence mechanisms [43–45]. Figures 1 and 2 show histograms of the moras and words, respectively. As we can see, the corpus included a variety of lengths, from short utterances (a few words and moras) to long utterances (70 words and 140 moras).

Next, we analyzed the changes in speech statistics per recording day. Speech data recorded during long periods causes there to be objective and subjective differences among recording days [26]. The mean of log $F_0$ was calculated for each recording day. $F_0$ was extracted by using the WORLD analysis-synthesis system [35,46]. Figure 3 shows the result. There was no special tendency in the first half of the recordings, but we can see that the log $F_0$ increased for the days of the second half.

4.2. JVS Corpus

4.2.1. 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 [47]. The 16-bit/sample RIFF WAV format was used. Sentences (transcriptions) were encoded in UTF-8. Full context and monophone labels were automatically generated by Open JTalk [48]. The phoneme alignments were automatically generated by Julius [49]. $F_0$ ranges were manually annotated in

| Sub-corpus         | Duration [hour] |
|--------------------|-----------------|
| basic5000 (5,000 utterances) | 6.77            |
| countersuffix26 (26 utterances) | 0.08            |
| loanword128 (128 utterances) | 0.15            |
| utparaphrase512 (612 utterances) | 0.65            |
| voiceactress100 (100 utterances) | 0.17            |
| onomatopee138 (300 utterances) | 0.45            |
| repeat500 (500 utterances) | 0.84            |
| travel1000 (1,000 utterances) | 0.97            |
| precedent138 (138 utterances) | 0.15            |
| total              | 10.23           |
accordance with hands-on voice conversion [50]. Commas were added between breath groups. For annotating perceptual similarity scores, we followed Saito et al.’s study [38] and used a crowdsourcing service. Each listener scored the perceptual similarity for each pair of speakers from −3 (completely different) to +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.

4.2.2. Analysis

Table 2 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.”

Figure 4 shows matrices of perceptual similarity scores. For example, the most similar pair was “jvs019” and “jvs096.” Also, one speaker that was most dissimilar from the other speakers was “jvs010.”

Figure 5 shows Log-$F_0$ mean values of each speaker. We can see that the JVS corpus covers a wider $F_0$ range from exp(4.6) = 100 Hz to exp(5.6) = 270 Hz.

5. CONCLUSION

In this paper, we constructed two corpora named the JSUT corpus and JVS corpus. These corpora were designed mainly for text-to-speech synthesis and voice conversion, respectively. The text data of the corpora is licensed as shown in the LICENCE file in the JSUT corpus. The tags of the corpora are licensed with CC BY-SA 4.0. The audio data of the corpora may be used for:

- Research by academic institutions,
- Non-commercial research, including research conducted within commercial organizations,
- Personal use, including blog posts.

Our project pages [11,12] describe the terms for commercial use.

ACKNOWLEDGEMENTS

Part of this work was supported by the SECOM Science and Technology Foundation, the GAP foundation program of the University of Tokyo, and the MIC/SCOPE #182103104. We thank Dr. Masahiro Mizukami of the Nara Institute of Science and Technology for the fruitful discussion on the paraphrase corpus, Assistant Prof. Kazuhide Yamamoto of the Nagoya University of Technology and Tomoyuki Kajiwara of the Tokyo Metropolitan University for the use of the SNOW E4 corpus, and the person in charge of the Voice Actress Corpus for the use of their corpus.

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 Process. Mag., 29, 82–97 (2012).
S. TAKAMICHI et al.: JSUT AND JVS CORPORA

[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,” arXiv 1609.03499 (2016).

[3] S. Takamichi, K. Tomoki and H. Saruwatari, “Sampling-based speech parameter generation using moment-matching networks,” Proc. Interspeech, Stockholm, Sweden, Aug., pp. 3961–3965 (2017).

[4] Y. Saito, S. Takamichi and H. Saruwatari, “Statistical parametric speech synthesis incorporating generative adversarial networks,” IEEE/ACM Trans. Audio Speech Lang. Process., 26, 84–96 (2018).

[5] M. Abe, Y. Sagisaka, T. Umeda and H. Kuwabara, “Speech database user manual,” ATR Tech. Rep., no. TR-I-0166M (1990).

[6] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoaka, T. Kobayashi, K. Shikano and S. Itahashi, “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” J. Acoust. Soc. Jpn. (E), 20, 199–206 (1999).

[7] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi and K. Oura, “Speech synthesis based on hidden Markov models,” Proc. IEEE, 101, 1234–1252 (2013).

[8] H. Zen, A. Senior and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” Proc. ICASSP, Vancouver, Canada, May, pp. 7962–7966 (2013).

[9] A. v. d. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals and K. Kavukcuoglu, “Parallel WaveNet: Fast high-fidelity speech synthesis,” arXiv 1711.10433 (2017).

[10] S. Tokui, R. Okuta, T. Akiba, Y. Niitani, T. Ogawa, S. Saito, S. Suzuki, K. Uenishi, B. Vogel and H. Y. Vincent, “Chainer: A deep learning framework for accelerating the research cycle,” arXiv 1908.00213 (2019).

[11] S. Takamichi, “JSUT: Japanese speech corpus of Saruwatari Lab, the University of Tokyo corpus,” https://sites.google.com/site/shinnosuketakamichi/publication/jsut (accessed 1 Nov. 2017).

[12] S. Takamichi, “JVS: Japanese Versatile Speech corpus,” https://sites.google.com/site/shinnosuketakamichi/research-topics/jvs-corpus (accessed 1 Sep. 2019).

[13] S. Ueno, M. Mimura, S. Sakai and T. Kawahara, “Multispeaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition,” Proc. ICASSP, Brighton, United Kingdom, May, pp. 6161–6165 (2019).

[14] Governments of Japan Agency for Cultural Affairs, “List of daily-use kanjis http://www.bunka.go.jp/kokugo/nihongo/sisakuto/joho/joho/kanji/index.html” (accessed 1 Aug. 2017).

[15] T. Toda, A. W. Black and K. Tokuda, “Voice conversion based on maximum likelihood estimation of spectral parameter trajectory,” IEEE Trans. Audio Speech Lang. Process., 15, 2222–2235 (2007).

[16] ybenjo and MagnesiumRibbon, “Voice-actress corpus,” http://voice-statistics.github.io/ (accessed 1 Aug. 2017).

[17] Wikipedia, https://ja.wikipedia.org/ (accessed 1 Aug. 2017).

[18] Y. Tanaka, “Compilation of a multilingual parallel corpus,” Proc. Pacilng2001, Kitakyushu, Japan, Sep., pp. 265–268 (2001).

[19] H. Kubozono, “Where does loanword prosody come from?: A case study of Japanese loanword accent,” Lingua, 116, 1140–1170 (2006).

[20] M. Moku, K. Yamamoto and A. Makabi, “Automatic easy Japanese translation for information accessibility of foreigners,” Proc. Workshop Speech and Language Processing Tools in Education, Mumbai, India, Dec., pp. 85–90 (2012).

[21] T. Kajiwara and K. Yamamoto, “Evaluation dataset and system for Japanese lexical simplification,” Proc. ACL/IJCNLP 2015 Student Research Workshop, Beijing, China, July, pp. 35–40 (2015).

[22] “SNOW E4: Evaluation data set of Japanese lexical simplification,” http://www.jnlp.org/SNOW/E4 (2010) (accessed 1 Aug. 2017).

[23] H. Tamaru, Y. Saito, S. Takamichi, T. Koriyama and H. Saruwatari, “Generative moment matching network-based random modulation post-filter for DNN-based singing voice synthesis and neural double-tracking,” Proc. ICASSP, Brighton, United Kingdom, May, pp. 7070–7074 (2019).

[24] M. Uiyama and M. Takahashi, “English-Japanese translation alignment data,” http://www2.nict.go.jp/astrec-att/member/mutymuya/align/index.html (2003) (accessed 1 Aug. 2017).

[25] “COURTS IN JAPAN,” http://www.courts.go.jp/app/hanrei_jp/search1 (accessed 1 Aug. 2017).

[26] H. Kawai, T. Toda, J. Ni, M. Tsuzaki and K. Tokuda, “XIMERA: A new TTS from ATR based on corpus-based technologies,” Proc. SSW5, Pittsburgh, U.S.A., June, pp. 179–184 (2004).

[27] Y. Stylianou, O. Cappe and E. Moulines, “Continuous probabilistic transform for voice conversion,” IEEE Trans. Speech Audio Process., 6, 131–142 (1998).

[28] H. Lu and S. King, “Factorized context modeling for Text-to-Speech synthesis,” Proc. ICASSP, Vancouver, Canada, May, pp. 7849–7853 (2013).

[29] T. Ito, K. Takeda and F. Itakura, “Analysis and recognition of whispered speech,” Speech Commun., 45, 139–152 (2005).

[30] Y. A. Petrushin, L. I. Tsirulnik and V. Makarova, “Whispered speech prosody modeling for TTS synthesis,” Proc. Speech Prosody, Chicago, U.S.A., May (2010).

[31] S.-C. Jou, T. Schultz and A. Waibel, “Whisper speech recognition using adapted articulatory features,” Proc. ICASSP, Philadelphia, U.S.A., Mar., Vol. 1, pp. 1009–1012 (2005).

[32] T. Toda, M. Nakagiri and K. Shikano, “Statistical voice conversion techniques for body-conducted unvoiced speech enhancement,” IEEE Trans. Audio Speech Lang. Process., 20, 2505–2517 (2012).

[33] D. G. Childers and C. K. Lee, “Vocal quality factors: Analysis, synthesis, and perception,” J. Acoust. Soc. Am., 90, 2394–2410 (1991).

[34] 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 Commun., 27, 187–207 (1999).

[35] M. Morise, F. Yokomori and K. Ozawa, “WORLD: A vocoder-based high-quality speech synthesis system for real-time applications,” IEEE Trans. Inf. Syst., E99-D, 1877–1884 (2016).

[36] D. Talkin, “REAPER: Robust Epoch And Pitch EstimatoR,” https://github.com/google/REAPER.

[37] P. Lanchantin, M. J. F. Gales, S. King and J. Yamagishi, “Continuous probabilistic transform for voice conversion,” Proc. ICASSP, Florence, Italy, May, pp. 285–289 (2014).

[38] Y. Saito, S. Takamichi and H. Saruwatari, “DNN-based speaker embedding using subjective inter-speaker similarity for multi-speaker modeling in speech synthesis,” Proc. SSW10, Vienna, Austria, Sep., pp. 51–56 (2019).

[39] “Lancers,” http://www.lancers.jp (accessed 1 Aug. 2017).

[40] T. Kudo, K. Yamamoto and Y. Matsumoto, “Applying conditional random fields to Japanese morphological analysis,” Proc. EMNLP, Barcelona, Spain, Jul., pp. 230–237 (2004).
Shinnosuke Takamichi received the Ph.D. degree from the Graduate School of Information Science, Nara Institute of Science and Technology, Nishino, Japan, in 2016. He is currently an Assistant Professor at The University of Tokyo. He has received more than ten paper/achievement awards including the 3rd IEEE Signal Processing Society Japan Young Author Best Paper Award.

Naoko Tanji received the Ph.D. degree from Harvard University, Boston, MA, USA, in 2017. She joined the Department of Artificial Intelligence at the University of Tokyo, Tokyo, Japan, in 2017, where she is currently an Assistant Professor. Her research interests include statistical audio signal processing, machine learning, and natural language processing. She is a member of IEEE, ISCA, and ASJ.

Kentaro Mitsui received the B.E. degree in engineering from the University of Tokyo, Japan, in 2019. He is studying for the M.S. degree in information physics and computing at the University of Tokyo. His research interests include speech synthesis, signal processing, and machine learning.

Tomoki Koriyama received the B.E. degree in computer science, and the M.E. and Dr. Eng. degrees in information processing from Tokyo Institute of Technology, Tokyo, Japan, in 2009, 2010, and 2013, respectively. In 2013 he joined the Research Laboratory of Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology as a Japan Society for the Promotion of Science Research Fellow. He is currently an Assistant Professor at the Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan. Dr. Koriyama was a recipient of The Aways Prize Young Researcher Award from Acoustic Society of Japan. He is a member of IEEE, ISCA, ASJ, IEICE, and IPSJ.

Ryosuke Sonobe is currently an undergraduate student in the Chemistry department at the University of Tokyo, Japan. He was engaged in this research as an intern student. Since then, he has been engaged in speech synthesis research for enterprises.

Hiroshi Saruwatari received the B.E., M.E., and Ph.D. degrees from Nagoya University, Japan, in 1991, 1993, and 2000, respectively. He joined SECOM IS Laboratory, Japan, in 1993, and Nara Institute of Science and Technology, Japan, in 2000. From 2014, he is currently a Professor of The University of Tokyo, Japan. His research interests include statistical audio signal processing, blind source separation (BSS), and speech enhancement. He has put his research into the world’s first commercially available Independent-Component-Analysis based BSS microphone in 2007. He received paper awards from IEICE in 2001 and 2006, from TAP in 2004, 2009, 2012, and 2018, from IEEE IROS2005 in 2006, and from APSIPA in 2013 and 2018. He received DOCOMO Mobile Science Award in 2011, Ichimura Award in 2013, The Commendation for Science and Technology by the Ministry of Education in 2015. Achievement Award from IEICE in 2017, and Hako-Award in 2018. He has been professionally involved in various volunteer works for IEEE, EURASIP, IEICE, and ASJ. He is an APSIPA Distinguished Lecturer from 2018.