TECHNICAL REPORT

Tohoku Kiritan singing database: A singing database for statistical parametric singing synthesis using Japanese pop songs

Itsuki Ogawa^{1,*} and Masanori Morise^{2,†}

¹*Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences,
University of Yamanashi,
4-3-11, Takeda, Kofu, 400-8511 Japan*

²*School of Interdisciplinary Mathematical Sciences, Meiji University,
4-21-1, Nakano, Nakano-ku, Tokyo, 164-8525 Japan*

(Received 19 October 2020, Accepted for publication 19 January 2021)

Abstract: We have built a singing database that can be used for research purposes. Since recent songs are protected by copyright law, researchers typically use songs that can be used without copyright. With changes to the copyright law in Japan in 2019, we can now release a singing database consisting of songs protected by the law under several restrictions. Our database mainly consists of Japanese pop songs by a professional singer. We collected a total of 50 songs with around 57 minutes of vocals recorded in a studio. After recording, we labeled the phoneme boundaries and converted the songs into the MusicXML format required for the study of statistical parametric singing synthesis. Statistical analysis of the database was then carried out. First, we counted the number of phonemes to clarify their distribution. Second, we performed acoustical analysis on the distribution of pitch, the interval between notes, and duration. Results showed that although the information is biased, the amount of singing is sufficient in light of the findings of a prior study on singing synthesis. The corpus is freely available at our website, <https://zunko.jp/kiridev/login.php> [1].

Keywords: Singing database, Singing synthesis, Statistical analysis, Copyright law of Japan

1. INTRODUCTION

Significant progress has been made with statistical parametric speech synthesis (SPSS) [2] due to the availability of vast databases and deep neural networks (DNNs). Conventionally, hidden Markov model (HMM) speech synthesis [2] outputs acoustic features representing speech waveforms from linguistic features obtained from text. The acoustic features are converted into a speech waveform by using a vocoder [3]. The initial speech database used for SPSS was around one hour long [4], which made it possible to synthesize at a lower cost than with concatenation synthesis. However, the sound quality of synthesized speech was inferior to that of human speech. The recent use of DNNs [5] has enabled advanced modeling, and while the sound quality of synthesized speech has consequently improved, it is still different from natural speech. WaveNet [6] was proposed as a DNN-based speech synthesizer that uses no vocoder, and its sound quality is

indistinguishable from human speech. It is supported by a huge amount of speech data (around 44 hours) compared to HMM speech synthesis. Furthermore, end-to-end speech synthesis techniques, such as Tacotron [7,8] and Deep Voice 3 [9], no longer require text analysis. The size of the speech database is over 100 hours. If we could prepare a large amount of text and speech data, latest speech synthesis techniques could be used to synthesize high-quality speech that is indistinguishable from a human speech. The quality of the database is crucial in terms of determining the sound quality of the synthesized speech.

Recently, statistical parametric singing synthesis has been utilized to synthesize high-quality singing voices by using HMM [10] and DNNs [11–14]. Unlike speech, singing voices include long tones and vibrato, and the intonation is different. SPSS requires a language model for intonation in word units, while singing synthesis does not. However, it is difficult to use a speech database as a singing database because singing synthesis requires information about music scores. A singing database must therefore consist of both singing voice and labeled data such as lyrics and musical scores. One of the main difficulties in creating singing databases for research in Japan is the copyright

*e-mail: g19tk006@yamanashi.ac.jp

†e-mail: mmorise@meiji.ac.jp

[doi:10.1250/ast.42.140]

issue, so currently few singing databases are publically available. In this study, we endeavored to build a singing database that can be used freely for research purposes. To this end, we have constructed a singing database composed of the singing voice of a voice actor.

Our main objective is to construct a singing database as a common resource not only for researchers but also for developers. In the past, such singing databases have generally consisted of out-of-copyright songs. Since modern pop songs could not be used due to the copyright law, children's songs have typically been used for singing synthesis studies [11,12]. In 2019, the copyright law in Japan changed to allow more freedom for the statistical analysis of data. In this study, we therefore focus on the new copyright laws and have built a singing database consisting of Japanese pop (J-pop) songs.

In Sect. 2 of this paper, we introduce similar databases used for speech/singing synthesis studies and explain the concept of our database. In Sect. 3, we analyze the database we built from several aspects and discuss its effectiveness. We conclude in Sect. 4 with a brief summary and mention of future work.

2. CONCEPT OF THE DATABASE

In this section, we first introduce related databases used for statistical parametric speech/singing synthesis and describe the difficulty such databases have faced with the copyright laws of Japan. We then clarify the problem and propose how to solve it.

2.1. Related Works

Various speech/singing databases are used for SPSS and singing synthesis. We list several examples in Table 1. The size of the database affects the sound quality of synthesized speech in SPSS and singing synthesis; specifically, to achieve sufficient sound quality, it can take anywhere from ten hours to several hundred hours for the end-to-end method and around one hour for the statistical parametric method. Also, a multi-speaker/singer database is required for a multi-speaker text-to-speech (TTS) system and voice conversion [15]. Furthermore, label data is required for learning. If there is no label data, users are required to create their own labels, which means that the quality will be different for each user. Therefore, when constructing a database that can be used as a baseline, the necessary label information should be provided. In the case of singing synthesis, information on the score is also required. Since the databases used for SPSS and statistical parametric singing synthesis typically contain about one hour of data, the database we constructed in this work is also about one hour.

Table 1 Databases for speech/singing synthesis.

Speech database				
Name	No. of speakers	Hour	Label	Language
NIT ATR503 M001 [16]	1	0.8	Yes	English
VCTK [17]	109	44	No	English
LibriSpeech [18]	2,484	820	No	English
LJSpeech [19]	1	24	No	English
JSUT [20,21]	1	10	Yes	Japanese
JVS [21,22]	100	30	No	Japanese
LibriTTS [23]	2,456	585	No	English
Singing database				
Name	No. of singers	Hour	Label	Language
NUS-48E [24]	12	2.8	Yes	English
NIT SONG070 F001 [25]	1	1.2	Yes	Japanese
VocalSet [26]	20	10.1	No	English
JSUT-song [27]	1	0.4	Yes	Japanese
JVS-MuSiC [28]	100	2.3	No	Japanese

2.2. Copyright Law of Japan

It is difficult to distribute a singing database compared to a speech database because of copyright law. Specifically, while we can cite a sentence by taking the appropriate steps, we cannot cite a song, including lyrics and audio. In particular, releasing singing voices violates the *performance clause* that is proscribed by the law. One simple solution is for researchers to use a singing database for their own research but not distribute it.

Another solution is to use out-of-copyright songs and thereby distribute a singing database [29] the same as with a speech database. In Japan, many children's songs are out-of-copyright, so several singing databases made up of such songs have been provided [25,27,28]. Since the singing style depends on the genre, conventional singing databases can feature vocalists using children's songs quite naturally. However, it would be difficult to sing songs of other genres quite so naturally. To expand the scope of singing synthesis studies, a singing database in other genres should be distributed. Such a database would also be useful as a baseline database.

In 2019, the copyright law in Japan changed to allow more freedom for statistical analysis. A particularly important change is the amendment of Article 30-4 of the Copyright Act 2019, which is partially quoted as follows [30].

It is permissible to exploit a work, in any way and to the extent considered necessary, in any of the following cases, or in any other case in which it is not a person's purpose to personally enjoy or cause another person to enjoy the thoughts or sentiments expressed in that work; provided, however, that this does not apply if the action would

unreasonably prejudice the interests of the copyright owner in light of the nature or purpose of the work or the circumstances of its exploitation:

- (i) *if it is done for use in testing to develop or put into practical use technology that is connected with the recording of sounds or visuals of a work or other such exploitation;*
- (ii) *if it is done for use in data analysis (meaning the extraction, comparison, classification, or other statistical analysis of the constituent language, sounds, images, or other elemental data from a large number of works or a large volume of other such data; the same applies in Article 47-5, paragraph (1), item (ii));*
- (iii) *if it is exploited in the course of computer data processing or otherwise exploited in a way that does not involve what is expressed in the work being perceived by the human senses (for works of computer programming, such exploitation excludes the execution of the work on a computer), beyond as set forth in the preceding two items.*

If a singing database is distributed to function as training data for statistical parametric singing synthesis, we can use the database without actually listening to the song. This suggests that we can distribute a singing database including songs protected by the copyright law for the purpose of statistical research. To guarantee the appropriate usage of the database, a certificate of consent to avoid violation of the purpose is required. We have therefore created a website on which users sign a letter of intent electronically.

2.3. Specifications of the Database

The singing database we built is for the study of singing synthesis. As another genre of the database, we focus on the J-pop songs sung by many users in Japan. Since the singing skill of the vocalist is an important factor in this genre, we hired a professional singer to record the songs. Ms. Himika Akaneya is a member of the idol group Iris and has experience recording with VOICEROID and UTAU, Japanese singing synthesizer applications. As a voice actor, she has played a character named *Tohoku Kiritan* in the Tohoku Zunko Project. We asked her to sing songs for our database by mimicking this character.

As of March 2020, Iris has released 24 CDs containing over 50 songs, mostly in the Japanese anime and J-pop genres. We recorded 50 of these songs under the conditions listed in Table 2 at a recording studio in Tokyo. We mainly recorded the first verses only as part of the agreement with Ms. Akaneya's management. The total recorded duration was around 3 hours, 31 minutes, and the total singing duration was around 57 minutes. This value was calculated from the phoneme label data.

Table 2 Overview of Kiritan singing database.

Singer	Ms. Himika Akaneya
Genre	Japanese anime song and J-pop
No. of songs	50
Total song duration	3 hours 31 minutes
Total phoneme duration	Around 57 minutes
Recording environment	Studio
Microphone	NEUMANN U 87 Ai
Sampling/Quantization	96 kHz/24 bit

2.4. Alignment of Phoneme and Score from Recorded Singing

To maintain the sound quality of the singing synthesis for all users, the singing database requires not only audio but also label data. In our singing database, the label data consist of a lyric and musical score. The boundaries of all phonemes were manually aligned by an expert and the basic label data were generated using McCab [31,32], NEologd [33,34] and Julius [35]. The boundaries were then aurally checked and manipulated by the expert.

For the lyric labels, we distributed both musical instrument digital interface (MIDI) and MusicXML files. MusicXML is a score in the XML format and can be converted from MIDI files. The basic labels were generated by Melodyne [36] and then manipulated by another expert.

3. STATISTICAL ANALYSIS OF THE DATABASE

After constructing the database, we carried out statistical analysis. The concept of this analysis is to support the user in using the database. We used the phoneme balance, pitch, and duration of each note and then illustrated their histograms. We then clarified the effectiveness of the database on basis of the results.

3.1. Phoneme

In our database, phoneme information is based on Sinsy [10,11,14]. Table 3 lists the number of vowels and single vowels. For example, /a/ in /ka/ is counted as all vowels, while a single /a/ is counted as a single vowel. For all vowels, the highest frequency vowel was /a/. In the single vowels category, /i/ was the highest.

Table 4 shows the number of phonemes in the Japanese mora unit. Each cell relates to Japanese *kana*, which are the syllabaries. Rows represent consonants used in Sinsy, and columns represent the five Japanese vowels. A dash indicates that there is no *kana*. In addition to the

Table 3 Number of all vowels and single vowels.

	a	i	u	e	o
All vowels	3,058	2,500	1,414	1,460	2,235
Single vowels	223	750	129	164	134

Table 4 Number of phonemes for a combination of vowels and consonants.

		Vowel				
		a	i	u	e	o
Consonant	k	361	382	241	140	255
	s	131	0	166	108	111
	t	343	7	5	342	307
	n	453	229	15	76	315
	h	81	90	—	9	42
	m	197	230	71	163	268
	y	65	—	84	0	155
	r	341	222	236	186	77
	w	196	5	—	2	240
	g	230	22	26	37	25
	z	21	0	80	31	10
	d	220	7	4	165	141
	b	87	59	48	32	47
	p	25	14	28	1	14
	ky	4	—	8	0	14
	sh	8	273	7	0	45
	ch	32	104	5	3	9
	ny	0	—	1	0	0
	hy	1	—	0	0	1
	my	0	—	3	0	0
	ry	1	—	1	0	10
	gy	0	—	2	0	1
	j	32	92	8	0	10
	py	0	—	3	0	0
	v	0	0	1	0	0
	ts	0	0	178	0	0
	f	6	14	64	1	4

information provided in the table, two other phonemes were observed: the syllabic nasal /N/ and choked sounds /cl/. These had 361 and 261 phonemes, respectively.

Since we used songs that had been released on CDs, the phoneme balance was ignored. As a result, there are zeros in several cells. However, several cells that indicate zero have no kana; for example, the number of phonemes /kye/ is zero, for which there is no related kana. One concern is that the frequencies of several consonants are low, which suggests that the user is required not to use such phonemes when composing the music in singing synthesis.

3.2. Score

We then analyzed the musical score information in terms of the pitch, interval between notes, and duration. These parameters were calculated from the MusicXML files. The analysis results are shown in the histogram, where the horizontal axis represents each analyzed parameter.

Figure 1 shows the histogram of the pitch. The minimum was F3 and the maximum was E5. The range was about 2 octaves. The mode was A4, and the median and average were G4. Figure 2 shows the histogram of the interval. The interval is a semitone, minus the pitch before

Fig. 1 Histogram of pitch. Number below the key represents octave number. **Fig. 2** Histogram of interval.

the pitch. All intervals above 10 or below -10 were combined into one. Since ± 6 was included in the dissonance, they are thought not to appear much in this database.

Figure 3 shows the histogram of the BPM. The minimum was 84 and the maximum was 190. Figure 4 shows the histogram of the duration. The step size was 0.1 seconds. Durations over 1.4 seconds were combined into one bar in the histogram. The maximum length was 3.7 seconds. The mode was 0.1 seconds, the median was 0.2 seconds, and the average was about 0.33 seconds. Most of the notes were between 0.1 and 0.4 seconds, i.e., quite short. There were few notes longer than 1.0 seconds.

3.3. Discussion

Our database consists of songs totaling around 57 minutes. This volume is presumably sufficient for natural

Fig. 3 Histogram of BPM.

Fig. 4 Histogram of duration. Durations above 1.4 seconds were combined into 1.4.

singing synthesis because the neural parametric singing synthesizer [12,13] has achieved natural singing synthesis from just Japanese children's songs totaling around 31 minutes. In a TTS system, phoneme-balanced sentences [37] are often used to cover every phoneme efficiently. In singing synthesis, it is difficult to make such phoneme-balanced sentences because the sentences are used as lyrics.

Changing the lyrics of a song would solve this problem, but this raises the issue of copyright. Composing original songs with phoneme-balanced lyrics is one solution. Since the singing style affects the singing output synthesized by a system, the composer needs to create songs that are easy to sing. While composing such songs is outside the scope of the present work, it is an important task for the future study of statistical parametric singing synthesis.

The genre recorded in the database is also important. Conventional singing synthesizers in Japan have mainly

used Japanese children's songs due to copyright law. The database built in this study provides another genre, which suggests that genre dependence in statistical parametric singing synthesis should be examined. While there are several issues to resolve in the database, such as the phoneme balance, we believe that showing how to build and release the singing database in compliance with the copyright law of Japan is an important contribution of this paper.

4. CONCLUSION

In this paper, we presented a singing database we built for research purposes and carried out statistical analysis on it. This database consists of 50 J-pop songs recorded by a female vocalist. We investigated several aspects of the constructed database. In the speech databases for TTS synthesis, phoneme-balanced sentences have been used, but since we utilized songs that have already been released on a CD, the phoneme-balance was ignored in this work. Our database has already been used for a singing synthesizer called NEUTRINO [38].

Our next step is to develop a statistical parametric singing synthesis using this database. We aim to construct DNNs that can use not only the music score and lyrics but also the F0 as the input to support the manipulation of F0, such as for human-in-the-loop speech design [39]. The implementation of a singing synthesizer including such an editor and DNNs is an important topic for future study.

ACKNOWLEDGMENTS

This work was supported by a JST PRESTO Grant Number JPMJPR18J8. We thank Mr. Y. Oda for coordinating the recording.

REFERENCES

- [1] "Tohoku Kiritan singing database for researchers (in Japanese)," <https://zunko.jp/kiridev/login.php> (accessed 11 Oct. 2020).
- [2] H. Zen, K. Tokuda and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, **51**, 1039–1064 (2009).
- [3] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, **11**, 169–177 (1939).
- [4] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," *Proc. ISCA SSW6*, pp. 294–299 (2007).
- [5] H. Zen, A. Senior and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *Proc. ICASSP*, pp. 7962–7966 (2013).
- [6] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv 1609.03499* (2016).
- [7] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. 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," *Proc. Interspeech*,

- pp. 4006–4010 (2017).
- [8] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyriannakis and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” *Proc. ICASSP*, pp. 4779–4783 (2018).
 - [9] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman and J. Miller, “Deep Voice 3: 2000-speaker neural text-to-speech,” *Proc. ICLR*, pp. 1–16 (2018).
 - [10] K. Oura, A. Mase, T. Yamada, S. Muto, Y. Nankaku and K. Tokuda, “Recent development of the HMM-based singing voice synthesis system — Sinsy,” *Proc. ISCA SSW7*, pp. 211–216 (2010).
 - [11] M. Nishimura, K. Hashimoto, K. Oura, Y. Nankaku and K. Tokuda, “Singing voice synthesis based on deep neural networks,” *Proc. Interspeech*, pp. 2478–2482 (2016).
 - [12] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer modeling timbre and expression from natural songs,” *Appl. Sci.*, **7**(12), pp. 1–23 (2017).
 - [13] M. Blaauw and J. Bonada, “A neural parametric singing synthesizer,” *Proc. Interspeech*, pp. 4001–4005 (2017).
 - [14] Y. Hono, S. Murata, K. Nakamura, K. Hashimoto, K. Oura, Y. Nankaku and K. Tokuda, “Recent development of the DNN-based singing voice synthesis system — Sinsy,” *Proc. APSIPA ASC*, pp. 1003–1009 (2018).
 - [15] Y. Stylianou, O. Cappe and E. Moulines, “Continuous probabilistic transform for voice conversion,” *IEEE Trans. Speech Audio Process.*, **6**, 131–142 (1998).
 - [16] HTS Working Group, “The NITech Japanese speech database NIT ATR503 M001,” <http://hts.sp.nitech.ac.jp/archives/2.3/HTS-demo.NIT-ATR503-M001.tar.bz2> (accessed 11 Oct. 2020).
 - [17] J. Yamagishi, “English multi-speaker corpus for CSTR voice cloning toolkit,” <https://doi.org/10.7488/ds/2645> (accessed 11 Oct. 2020).
 - [18] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, “LibriSpeech: An ASR corpus based on public domain audio books,” *Proc. ICASSP*, pp. 5206–5210 (2015).
 - [19] K. Ito, “The LJ speech dataset,” <https://keithito.com/LJ-Speech-Dataset> (accessed 11 Oct. 2020).
 - [20] R. Sonobe, S. Takamichi and H. Saruwatari, “JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis,” *arXiv 1711.00354* (2017).
 - [21] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari, “JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research,” *Acoust. Sci. & Tech.*, **41**, 761–768 (2020).
 - [22] S. Takamichi, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji and H. Saruwatari, “JVS corpus: Free Japanese multi-speaker voice corpus,” *arXiv 1908.06248* (2019).
 - [23] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen and Y. Wu, “LibriTTS: A corpus derived from LibriSpeech for text-to-speech,” *Proc. Interspeech*, pp. 1526–1530 (2019).
 - [24] Z. Duan, H. Fang, B. Li, K. Sim and Y. Wang, “The NUS sung and spoken lyrics corpus: A quantitative comparison of singing and speech,” *Proc. APSIPA ASC*, pp. 1–9 (2013).
 - [25] HTS Working Group, “The NITech Japanese speech database NIT SONG070 F001,” <http://hts.sp.nitech.ac.jp/archives/2.3/HTS-demo.NIT-SONG070-F001.tar.bz2> (accessed 11 Oct. 2020).
 - [26] J. Wilkins, P. Seetharaman, A. Wahl and B. Pardo, “VocalSet: A singing voice dataset,” *Proc. ISMIR*, pp. 468–474 (2018).
 - [27] S. Takamichi, N. Tanji and H. Saruwatari, “JSUT collection ver. 1: Voice corpora of reading-style speech, singing voice and vocal imitation uttered by a single Japanese speaker,” <https://sites.google.com/site/shinnosuketakamichi/publication/jsut-song> (accessed 11 Oct. 2020).
 - [28] H. Tamaru, S. Takamichi, N. Tanji and H. Saruwatari, “JVS-MuSiC: Japanese multispeaker singing-voice corpus,” *arXiv 2001.07044* (2020).
 - [29] M. Goto, H. Hashiguchi, T. Nishimura and R. Oka, “RWC music database: Popular, classical, and jazz music databases,” *Proc. ISMIR*, pp. 287–288 (2002).
 - [30] “Japanese law translation database system,” Ministry of Justice, Amendment of Act No. 30 of 2018 (2018). <http://www.japaneselawtranslation.go.jp/law/detail/?id=3379&vm=04&re=02&new=1> (accessed 11 Oct. 2020).
 - [31] T. Kudo, K. Yamamoto and Y. Matsumoto, “Applying conditional random fields to Japanese morphological analysis,” *Proc. EMNLP*, pp. 230–237 (2004).
 - [32] T. Kudo, “MeCab: Yet another part-of-speech and morphological analyzer,” <https://taku910.github.io/mecab> (accessed 11 Oct. 2020).
 - [33] T. Sato, T. Hashimoto and M. Okumura, “Implementation of a word segmentation dictionary called mecab-ipadic-NEologd and study on how to use it effectively for information retrieval (in Japanese),” *Proc. 23rd Annu. Meet. Assoc. Natural Language Processing*, pp. NLP2017–B6–1 (2017).
 - [34] T. Sato, “Neologism dictionary based on the language resources on the web for MeCab,” <https://github.com/neologd/mecab-ipadic-neologd> (accessed 11 Oct. 2020).
 - [35] A. Lee, T. Kawahara and K. Shikano, “Julius—an open source real-time large vocabulary recognition engine,” *Proc. Eurospeech*, pp. 1691–1694 (2001).
 - [36] “Celemony|What is Melodyne?,” <https://www.celemony.com/en/melodyne/what-is-melodyne> (accessed 11 Oct. 2020).
 - [37] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara and K. Shikano, “ATR Japanese speech database as a tool of speech recognition and synthesis,” *Speech Commun.*, **9**, 357–363 (1990).
 - [38] “NEUTRINO —Neural singing synthesizer— (in Japanese),” <https://n3utrino.work> (accessed 11 Oct. 2020).
 - [39] D. Kondo and M. Morise, “Human-in-the-loop speech-design system and its evaluation,” *Proc. APSIPA ASC*, pp. 608–612 (2019).

Itsuki Ogawa received the B.S. degree from University of Yamanashi, Yamanashi, Japan, in 2019. He is currently an M.E. student of Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi. His research interests include speech analysis/synthesis, singing voice synthesis.

Masanori Morise received the Ph.D. degree in engineering from Wakayama University in 2008. He was a JSPS Research Fellow (DC1) in 2006–2008, a postdoctoral researcher at Kwansei Gakuin University in 2008–2009, an Assistant Professor at Ritsumeikan University in 2009–2013, a Project Assistant Professor at University of Yamanashi in 2013–2017, and an Associate Professor at University of Yamanashi in 2017–2019. He is currently an Associate Professor at Meiji University. His research interests include speech analysis/synthesis and speech design interface.