Human-in-the-loop speech-design system and its evaluation

Daichi Kondo* and Masanori Morise† ‡
* Integrated Graduate School of Medicine, Engineering, and Agricultural Sciences,
University of Yamanashi, Japan
E-mail: g18tk007@yamanashi.ac.jp
† School of Interdisciplinary Mathematical Sciences, Meiji University, Japan
E-mail: mmorise@meiji.ac.jp Tel/Fax: +81-3-5343-8332
‡ JST, PRESTO, Japan

***Abstract*—We propose human-in-the-loop (HITL) speech-design system with an interface. General text-to-speech (TTS) systems generate the speech waveform from the input text without the need for manual modification. In particular, end-to-end TTS systems can synthesize speech as naturally as human speech. However, it is difficult for users to modify the speech parameters without degrading sound quality. The purpose of this study was to enable collaboration between the user and a deep neural network (DNN) to develop a system with which a user can control the speech parameters without sound-quality degradation. The main problem to be solved is to improve the quality of the speech-parameters generated from the speech parameters designed by the user. We developed several acoustic models with DNNs to meet the purpose of this study. We carried out a subjective evaluation to determine the effectiveness of the proposed system. The subjective score regarding Muffledness improved by using the proposed system compared with speech processed using a TTS system that involves signal-processing without a DNN.**

I. Introduction

Various methods, such as formant speech synthesis [1] and concatenative speech synthesis [2], have been proposed for text-to-speech (TTS) synthesis. In particular, statistical parametric speech synthesis (SPSS) [3] has been widely used, and a hidden Markov model (HMM) [4] has been used as the fundamental technique for supporting such systems. Speech synthesized with such systems is inferior to that uttered by humans regarding sound quality. In 2013, an SPSS system that using a DNN was proposed [5], and the sound quality approached that of human speech.

These systems require a high-quality vocoder [6]. The speech waveform is generated from the speech parameters generated using an HMM or DNN. A recent study on SPSS proposed a system called WaveNet [7], [8] that requires no vocoder. This system enables a higher level of speech synthesis than a conventional speech-synthesis system. End-to-end systems, such as Tacotron [9], were then proposed and are becoming mainstream.

The purpose of general SPSS studies is developing an automatic speech synthesizer without the need for user manipulation. By inputting text, an SPSS system generates ideal speech that requires no modification. However, the user would want to manipulate the speech even if the system can output speech similar to that produced by humans. For example, a user often manipulates the speech parameters locally by post-processing, which degrades the manipulated speech.

Degradation in manipulated speech has been addressed in previous studies. Voice morphing [10], [11] and voice conversion [12] have been proposed as techniques that include speech-parameter manipulation. These studies have shown that it is difficult to manipulate speech parameters without sound-quality degradation. The main cause of such degradation is the mismatch between pitch and timbre in the speech parameters. A vocoder decomposes the speech waveform into the fundamental frequency (f_{o}), spectral envelope (Sp), and aperiodicity (Ap). Manipulation of the f_{o} should be carried out with other parameters related to timbre.

The purpose of this study was to develop the human-in-the-loop (HITL) speech-design system to prevent such degradation by using a DNN to generate speech parameters from those designed by the user. The novelty of this study is the collaboration between the user and DNN. In the proposed system, the user can control the f_{o}, and the DNN attempts to generate other speech parameters without sound-quality degradation. We call this system the human-in-the-loop (HITL) speech design system that includes an interface we also developed. We verified the effectiveness of the system through a subjective evaluation.

In Section 2, we explain the concept of the proposed system and discuss the interface of the proposed system in Section 3. In Section 4, we discuss the evaluation conditions and the results. In Section 5, we clarify the effectiveness of the acoustic models used in the evaluation. We conclude in Section 6 with a brief summary and mention future work.

II. Concept of proposed system

SPSS has been used to synthesize natural speech from input text. In current state-of-the-art systems [7], [8], [9], it is possible to generate the speech waveform as naturally as the human speech. Since these systems require no vocoder, the user cannot manipulate the speech parameters. Post-processing after generating the speech waveform is generally required to manipulate speech parameters, but sound quality degrades. This is because there is no interaction between the f_{o} and Sp. In cases in which intonation is manipulated naturally, the Sp

should also be manipulated along with the f_o to enable this interaction.

We attempted to solve this problem by constructing a neural network to enable interaction between the f_o and Sp. The synthesis component of current TTS systems consist of two. One part outputs the phoneme boundaries and the f_o contour from the text. The other part outputs other parameters from the phoneme boundaries and f_o contour. We adopted the approach that outputs the acoustic parameters including the f_o contour again from the phoneme boundary and the f_o contour manipulated by the user. Sound-quality degradation is expected when the f_o contour is highly manipulated by the user. In cases in which the manipulated f_o contour slightly changes, it is expected that the output with the timbre reflects the interaction.

A problem in constructing such a neural network is with an accent-dependent model. When outputting the phoneme boundaries and f_o contour from the text, an accent-dependent model is indispensable. On the other hand, in the neural network with the accent-dependent model, sound-quality degradation is expected because of the mismatch between the manipulated f_o contour and accent. We therefore trained the neural network by adopting a non-accent-dependent model.

As mentioned above, we also developed an interface called TalkingHead for the proposed system. Similar interfaces handling speech design are VOCALOID [13] using diphone speech synthesis [14] and v.morish [15] using real-time voice morphing. The purpose of TalkingHead is to support the collaboration between the user and DNN.

III. TalkingHead: Interface for HITL Speech-Design System

We developed TalkingHead with TTS, speech-design, and speech-waveform-generation functions from the designed parameters. Fig. 1 shows a snapshot of TalkingHead. The top part of the figure represents the TTS component. The user can input text into the text box. The text is first analyzed, then the phoneme boundaries and f_o contour are displayed at the bottom.

The bottom part of the figure represents the phoneme-duration/f_o-contour-design component. This component has two tabs: a tab for manipulating the phoneme duration and one for manipulating the f_o contour. The user can directly describe the f_o contour by using a mouse.

A. TTS component

In the TTS component, the user can synthesize speech from a text and reproduce the synthesized speech waveform. When the user enters a text in the text box then clicks the Synthesize button, the speech waveform is synthesized. At the same time, the phoneme duration and f_o contour of the synthesized speech are displayed in the phoneme-duration/f_o-contour-design component. By clicking the Play button, synthesized speech is reproduced.

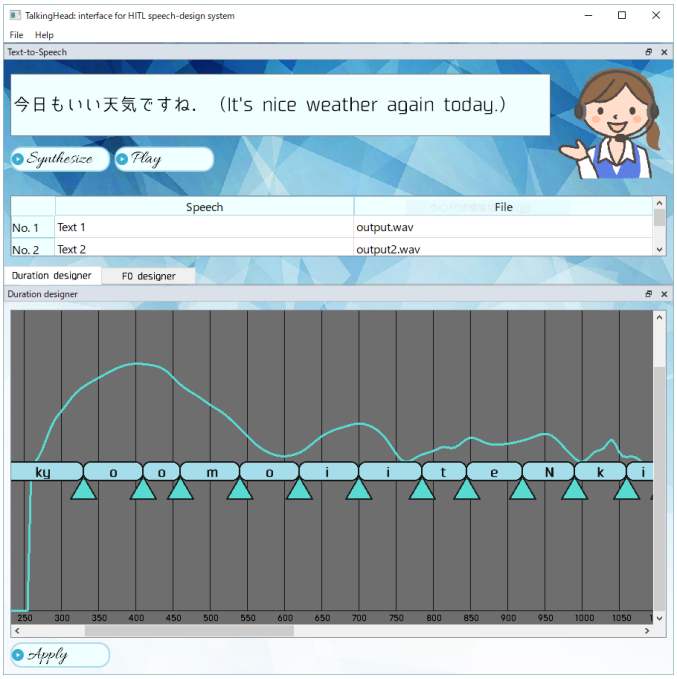

Fig. 1. Snapshot of TalkingHead

B. Phoneme-duration-design tab

In the phoneme-duration-design tab, the duration of each phoneme and f_o contour are displayed. The horizontal and vertical axes represent the time and frequency, respectively. The user can manipulate the phoneme duration by dragging the cursor related to the start and end times of each phoneme in the horizontal direction. When the Apply button is clicked, the speech waveform is generated from the trained acoustic model using the manipulated phoneme duration as the input. The speech waveform reflecting the design can be confirmed by clicking the Play button in the TTS component.

C. f_o-design tab

In the f_o-design tab, f_o contour is displayed as well as the phoneme duration. By dragging the displayed f_o contour with the mouse, the user can draw the f_o contour. When the Apply button is clicked, the speech waveform is generated from the trained acoustic model using the drawn f_o contour as the input. The speech waveform reflecting the design can be confirmed by clicking the Play button in the TTS component. By clicking each tab, the user can switch between phoneme duration and f_o design.

IV. Evaluation

We conducted an evaluation to ascertain the improvement in sound quality of speech synthesis with the proposed system for manually manipulated f_o contour.

A. Acoustic models used in evaluation

1) Features used for training acoustic models: Fig. 2 illustrates the DNN-based acoustic models used in the evaluation. We extracted linguistic features from labels in the speech

database by using algorithms in Merlin [16], which converts labels to its numerical representation in the text analysis. We also removed accent-dependent linguistic features to address mismatch by using the manually manipulated f_{o} as one of the input features. We used WORLD [17] as the high-quality vocoder for speech-parameter extraction. Since it has several estimators for each parameter, we used Harvest [18], CheapTrick [19], [20], and D4C [21] to estimate the f_{o}, Sp, Ap, and voiced/unvoiced features, respectively. The frame shift was set to 5 ms, and other parameters were set to their defaults. The log f_{o} was used instead of the linear f_{o}. Then, Sp and Ap were encoded into 60 mel-cepstrum and 5-band aperiodicity (BAP), respectively. These values were determined by the conventional result [22]. The delta and delta-delta features of log f_{o}, mel-cepstrum, and BAP were also used for training the acoustic models.

2) Concepts of acoustic models: Fig. 2 also illustrates the relationships between the input and output in each acoustic model. Each model has different input and output features. In Model A, non-accent-dependent linguistic features were used as input features, and Sp and Ap were obtained as output features. In Model B, non-accent-dependent linguistic features and f_{o} were used as input features, and Sp and Ap were obtained as output features. In Model C, non-accent-dependent linguistic features and f_{o} were used as input features, and f_{o}, Sp, and Ap were obtained as output features. These acoustic models were prepared to investigate the effective combinations of features by determining which features positively affect the sound quality.

Feed-forward neural networks that have twelve hidden layers with 600 nodes in each layer were used. The tanh activation function and Adam [23] optimization were also used in the training. We used the Japanese speech corpus of Saruwatari Lab, University of Tokyo (JSUT) [24], which has 7,696 utterances, as the training data.

3) Advantage of Model C: By using non-accent-dependent linguistic features and f_{o}, we assumed that we can model the change in f_{o}, which cannot be expressed with an accent. The sound quality of Model C would be the best of all models because it outputs a more appropriate f_{o} contour. It is superior to Models A and B, which does not use f_{o} as input/output or output features, respectively.

B. Speech synthesis used for each condition

Fig. 3 illustrates the speech-synthesis procedure in the evaluation. Condition 1 involved the baseline speech stimuli synthesized using DNN-based SPSS. The acoustic model trained with this DNN used accent-dependent linguistic features. Additionally, f_{o}, Sp, and Ap were used for other conditions. Similarly to the proposed system, feed-forward neural networks that have twelve hidden layers with 1024 nodes in each layer were used. The tanh activation function and Adam [23] optimization were used in the training. Condition 2 involved speech stimuli that manipulated the intonation under Condition 1. We then synthesized speech stimuli using the manipulated f_{o} and original Sp and Ap.

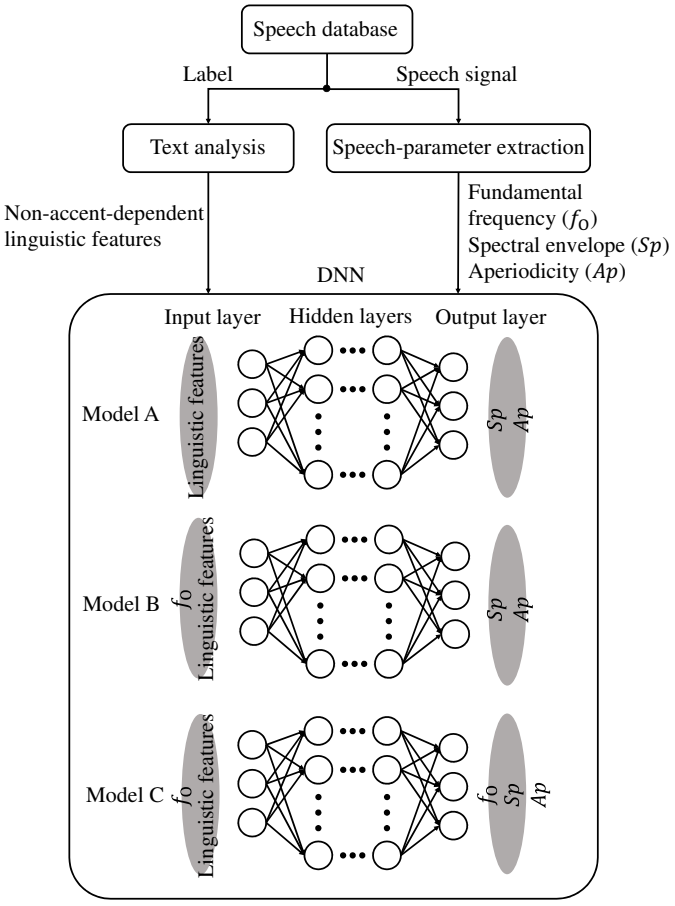

Fig. 2. Acoustic models used in evaluation

Conditions 3, 4, and 5 were speech stimuli generated using Models A, B, and C, respectively. Linguistic features used for the input to the models were non-accent-dependent and extracted from the label under Condition 2. Speech stimuli generated under Conditions 3 and 4 were synthesized using the manipulated f_{o} and output parameters (Sp and Ap). Speech stimuli generated under Condition 5 were synthesized using output parameters (f_{o}, Sp, and Ap). The difference between the conditions was whether the manipulated f_{o} was used.

C. Evaluation conditions

Table I lists the evaluation conditions. The evaluation was carried out based on the comparison mean opinion score (CMOS) defined by ITU-T recommendation P.800 Annex E [25] . Twenty participants with normal hearing ability listened to two speech stimuli and scored the second speech stimulus compared with the first on a 7-point scale (Much Worse: -3 to Much Better: 3). The CMOS can generally be used to evaluate the smaller sound-quality difference than the mean opinion score (MOS). We used a sound-proof room with an A-weighted sound pressure level (SPL) of 18 dB was used, and a set of headphones (Sennheiser HD650).

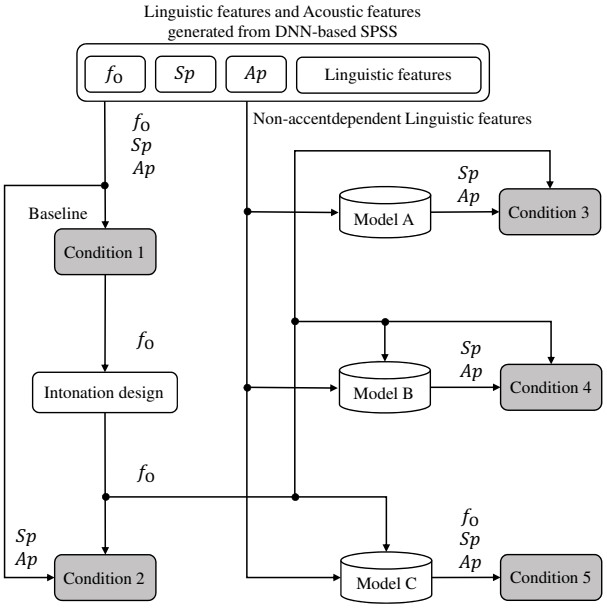

Fig. 3. Speech synthesis used for each condition

TABLE I
EVALUATION CONDITIONS

Evaluation protocol	
Method	Comparison mean opinion score evaluation (CMOS)
Number of participants	20
Number of stimuli	50 (10 utterances per 5 conditions)
Evaluation items	Naturalness, Muffledness, and Human-ness
Environment and equipment for reproduction	
Environment	Soundproof room
Background noise	18 dB (A-weighted SPL)
Headphones	Sennheiser HD650
Audio I/O	Roland QUAD-CAPTURE

Ten utterances not included in the training data were used as texts of each speech stimulus from ATR503 [26], and the total number of speech stimuli was 50. We manually designed intonation by drawing the f_0 contour using TalkingHead, the same as in the f_0-design tab in Fig. 1. In Fig. 3, Condition 1 was the baseline and used as the reference in the CMOS evaluation. Conditions 2 to 5 were used as the evaluation targets. Evaluation items was determined to verify not only the sound quality (Naturalness) but also the Muffledness that is one of the problems in SPSS. Human-ness was also used to confirmed whether synthesized speech was similar to the real speech. The speech stimuli were randomized and reproduced to the participants. To suppress the effect of intonation of the manipulated f_0 on the evaluation, we instructed participants to evaluate the stimulus regardless of the intonation.

D. Results

Fig. 4 shows the results. The vertical axis represents the average scores under each condition. The error bar represents the 95% confidence interval. We carried out a statistical

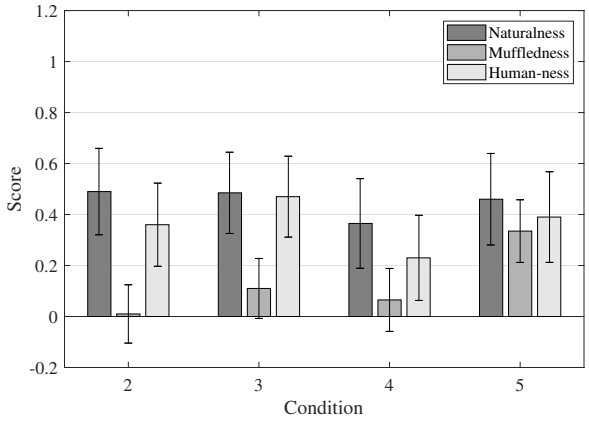

Fig. 4. Results of CMOS evaluation

analysis on the results, which showed that the scores under all conditions were superior to those from the baseline regarding Naturalness and Human-ness. Condition 3 was superior to Condition 2 regarding Muffledness and Human-ness. Condition 4 was inferior to Condition 2 regarding Naturalness and Human-ness. Additionally, the score of Muffledness in Condition 5 was significantly greater than those of the others ($p < 0.001$). The results indicate that our hypothesis stating that Model C is the best was supported in the Muffledness. The results from the other items, however, did not show the effectiveness of Model C.

V. DISCUSSION

In this section, we discuss the effectiveness of the proposed system based on the evaluation results. The scores under all conditions were superior to those from the baseline regarding Naturalness and Human-ness. Although we instructed the participants regarding intonation in advance, it seems that these two evaluation items were affected by intonation. Condition 3 was superior to Condition 2 regarding Muffledness and Human-ness. The reason for this might be due to accent errors between the label and speech of the training data used for the acoustic model of the baseline. Japanese accents often change under conditions such as region, era, and speaker. In Model A used in Condition 3, there was no accent dependency in the labels. Since there were no accent errors in the training data, the score was considered high.

The score of Muffledness under Condition 5 was the highest of all conditions. Model C used f_0 as both input and output, so we analyzed the f_0 contour as the input and output of Model C. Fig. 5 shows an example of the f_0 contour as the input and output of Model C. The vertical axis represents the f_0 under each condition. The f_0 contour was modified, but maintaining the intonation of the f_0 was also observed.

By using f_0 as the input, interaction among acoustic features that can not be expressed with accent was obtained. By suppressing the smoothing of the acoustic features, the Muffledness score can be improved. On the other hand, Condition 4 was inferior to Condition 2 in Naturalness and Human-ness.

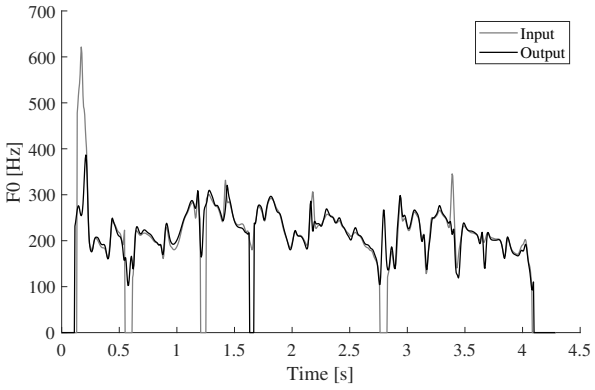

Fig. 5. Example of f_{o} contour in Model C

Model B used under Condition 4 used f_{o} as the input. These results indicate that the proposed system is effective.

VI. Conclusion

We proposed the HITL speech-design system with an interface we developed called TalkingHead. Compared with conventional SPSS systems, we trained acoustic models for speech design. To address sound-quality degradation by speech design, we used non-accent-dependent linguistic features and manipulated f_{o} for the trained acoustic model. As a result, the score of Muffledness significantly improved.

The next step of the study is to expand TalkingHead and evaluate its usability. We implemented it with feed-forward neural networks in this study, but it is a future task to implement it with a DNN that has higher performance and confirm the effect of different types of DNNs on the sound quality in our system.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP16H01734 and JP16H05899, and JST PRESTO Grant Number JPMJPR18J8, Japan.

References

[1] D. H. Klatt, "Software for a cascade/parallel formant synthesizer," *Journal of the Acoustical Society of America*, vol. 67, no. 3, pp. 971–995, 1980.

[2] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," *in Proc. ICASSP 1996*, vol. 1, pp. 373–376, 1996.

[3] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, pp. 1039–1064, 2009.

[4] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in hmm-based speech synthesis," *in Proc.Eurospeech*, 1999.

[5] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," *in Proc. ICASSP 2013*, pp. 7962–7966, 2013.

[6] H. Dudley, "Remaking speech," *J. Acoust. Soc. Am.*, vol. 11, no. 2, pp. 169–177, 1939.

[7] 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 preprint arXiv:1609.03499*, 2016.

[8] A. van den Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. v. d. Driessche, E. Lockhart, L. C. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," *arXiv preprint arXiv:1711.10433*, 2017.

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[10] H. Ye and S. Young, "High quality voice morphing," *in Proc. ICASSP 2004*, 2004.

[11] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and H. Banno, "Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown," *in Proc. ICASSP 2009*, pp. 3905–3908, 2009.

[12] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Journal of the Acoustical Society of Japan (E)*, vol. 11, no. 2, pp. 71–76, 1990.

[13] H. Kenmochi and H. Ohshita, "Vocaloid—commercial singing synthesizer based on sample concatenation," *in Proc. INTERSPEECH 2007*, pp. 4011–4010, 2007.

[14] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5-6, pp. 453–467, 1990.

[15] M. Morise, M. Onishi, H. Kawahara, and H. Katayose, "v. morish ' 09: A morphing-based singing design interface for vocal melodies," *Lecture Notes in Computer Science*, vol. LNCS 5709, pp. 185–190, 2009.

[16] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," *in Proc. SSW9 2016*, pp. 218–223, 2016.

[17] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. Inf. & Syst.*, vol. E99-D, pp. 1877–1884, 2016.

[18] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," *in Proc. INTERSPEECH2017*, pp. 2321–2325, 2017.

[19] ——, "CheapTrick, a spectral envelope estimator for high-quality speech synthesis," *Speech Communication*, vol. 67, pp. 1–7, 2015.

[20] ——, "Error evaluation of an f0-adaptive spectral envelope estimator in robustness against the additive noise and f0 error," *IEICE Trans. Inf. & Syst.*, vol. E98-D, no. 7, pp. 1405–1408, 2015.

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

[22] M. Morise, G. Miyashita, and K. Ozawa, "Low-dimensional representation of specral envelope without deterioration for full-band speech analysis/synthesis system," *in Proc. INTERSPEECH 2017*, pp. 409–413, 2017.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[24] R. Sonobe, S. Takamichi, and H. Saruwatari, "Jsut corpus: free large-scale japanese speech corpus for end-to-end speech synthesis," *arXiv preprint arXiv:1711.00354*, 2017.

[25] I. Rec, "P. 800: Methods for subjective determination of transmission quality," *International Telecommunication Union, Geneva*, 1996.

[26] 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 Communication*, vol. 9, no. 4, pp. 357–363, 1990.