Prosody-aware subword embedding considering Japanese intonation systems and its application to DNN-based multi-dialect speech synthesis

Takanori Akiyama, Shinnosuke Takamichi and Hiroshi Saruwatari
Graduate School of Information Science and Technology, The University of Tokyo, Japan

***Abstract*—This paper presents prosody-aware subword embedding considering Japanese intonation systems and its application to DNN (deep neural network)-based multi-dialect speech synthesis. In accordance with recent improvements of speech synthesis in rich-resourced languages, the research trend is shifting to more challenging languages such as Japanese dialects that still have undefined prosodic contexts. Conventional prosody-aware word embedding can unsupervisedly extract the contexts in a data-driven manner using words and F_0 sequences. However, accurate contexts for unknown words are difficult to generate. To solve this problem, we propose prosody-aware subword embedding considering Japanese intonation systems. The unsupervised subword model, which is trained considering language and acoustic characteristics, can tokenize an unknown word into known subwords suitable for prosody-aware embedding. We also propose a modulation filtering method considering intra-subword moras to improve the embedding accuracies. We apply the methods to not only Japanese but also Japanese multi-dialect speech synthesis. In the multi-dialect case, we propose subword models shared among dialects and embedding models conditioned by dialect information. The experimental evaluation demonstrates that the proposed multi-dialect methods can improve speech quality in some Japanese dialects.**

Index Terms: deep neural network, DNN-based speech synthesis, prosody-aware subword embedding, Japanese intonation systems, multi dialect

I. Introduction

Statistical parametric speech synthesis is a method of synthesizing speech using statistical models [1]. Deep neural network (DNN)-based ones [2]–[5] have especially attracted a lot of attention and remarkably improved synthetic speech quality in rich-resourced languages such as English and Japanese. Thanks to this, the research trend is shifting to more challenging languages such as dialect speech synthesis that enables diversity in speech communication augmented by speech synthesis.

We aim to build the multi-dialect speech synthesis shown in Fig. 1, which can synthesize a single speaker's voice in the preferred dialect. This system is similar to multi-lingual speech synthesis in rich-resourced languages [6], but there are two significant differences. First, speech characteristics among dialects within a target country (e.g., Japan) change more continuously than those among rich-resourced languages spoken in different countries (e.g., Japan and the U.S.A). Therefore, we expect that the characteristics of synthesized speech will be intuitively and continuously controlled, and quickly adapted to the unseen dialects. Another factor, which we address in this paper, is that contextual features used for speech synthesis are not yet defined in most Japanese dialects. In the case of rich-resourced languages, the dictionary-based or rule-based approaches (e.g., flite [7] or open-jtalk [8]) are often adopted to generate the contexts. However, applying such approaches to many kinds of dialects is unrealistic and time-consuming. In this paper, we address automatic retrieval of prosodic contexts, which is one of the main factors of Japanese dialects.

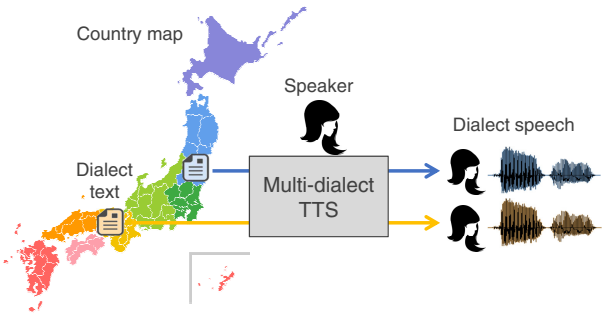

Fig. 1. Concept of multi-dialect speech synthesis. A single speaker's speech is synthesized from a text of a preferred dialect. This paper addresses automatic prosodic context generation for multi-dialect speech synthesis.

To tackle the automatic prosodic context generation, Ijima *et al*. [9] proposed prosody-aware word embedding for English speech synthesis. This method inspired by word embedding [10] that unsupervisedly extracts prosodic contexts by training embedding models from word and F_0 sequence pairs. However, in the case of languages that have the enormous numbers of words, such as Japanese and multiple Japanese dialects, training of the embedding models becomes inaccurate because the number of model parameters is significantly increased. Also, accurate contexts for unknown words are difficult to generate.

In this paper, we propose prosody-aware subword embedding considering Japanese intonation systems. The proposed subword model, which is unsupervisedly trained considering language models (i.e., frequency counts of subwords) and accent phrase boundaries, can tokenize an unknown word into known subwords suitable for prosody-aware embedding. The

embedding models are trained using pairs of the tokenized subword and corresponding F_0 sequence, and the subword-level prosody contexts are gained as bottleneck features of the embedding models. Also, we propose a modulation filtering method considering intra-subword moras to improve the embedding accuracies. We apply the proposed methods to not only Japanese but also Japanese multi-dialect speech synthesis. In the multi-dialect case, we propose mixing-dialect subword models shared among dialects and multi-dialect embedding models conditioned by dialect information. The experimental evaluation demonstrates that (1) the proposed method outperforms conventional prosody-aware word embedding in terms of the naturalness of speech in Japanese speech synthesis, and (2) proposed multi-dialect models can improve naturalness of speech in some Japanese dialects compared with Japanese-common-language models.

II. DNN-BASED SPEECH SYNTHESIS USING PROSODY-AWARE WORD EMBEDDING

In DNN-based speech synthesis, many types of contextual features are extracted from the input text, e.g., pronunciation, prosody (e.g., accent type or stress), and duration (e.g, frame position in the current phoneme) contexts. Prosody-aware word embedding [9] is an unsupervised method of extracting the prosodic contexts, and the DNN-based embedding models are trained from a speech corpus including pairs of a word and a continuous F_0 sequence (obtained from a F_0 sequence and spline interpolation [11]). The embedding models predict a continuous F_0 sequence from a corresponding word vector (given as a one-hot vector). Before using the continuous F_0 sequence in the model training process, the sequence corresponding to one word is first resampled to be a fixed-length vector. Then, lower-order components of the discrete cosine transform (DCT) of the vector are used for predicting. The prosodic context is gained as bottleneck features of this neural network. This method suffers from two problems. First, it is difficult to robustly train the models when the number of vocabulary increases and to extract the contexts for words not included in the training data. Second, the use of the fixed-order DCT components that ignore the word's complexity (e.g., the number of syllables) may lead to the modeling of unnecessary F_0 information or disregarding necessary F_0 information. For instance, it is unnatural to use the same extent of F_0 information for 'a' as for 'linguistic.'

III. PROSODY-AWARE SUBWORD EMBEDDING CONSIDERING JAPANESE INTONATION SYSTEMS

This section proposes prosody-aware subword embedding considering Japanese intonation systems. Fig. 2 shows the process. The input text is tokenized into not words but subwords. The proposed accent-phrase-informed subword model is unsupervisedly trained using text corpora on the basis of frequency counts of subwords and accent phrase boundaries. After resampling a continuous F_0 sequence corresponding to one subword, the proposed mora-informed modulation filtering method is applied to the resampled sequence.

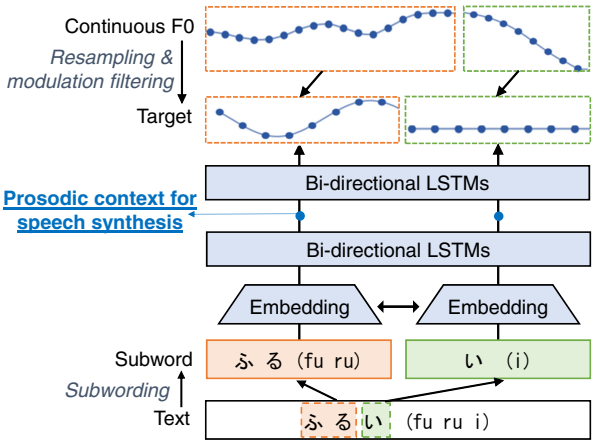

Fig. 2. Architecture of prosody-aware subword embedding.

A. Accent-phrase-informed subword model

Subword segmentation is a method of tokenizing rare words into subword units [12] and can alleviate an out-of-vocabulary problem by tokenizing an unknown word (e.g., 'linguistic') into a known subword sequence (e.g., 'lin,' 'gui,' and 'stic'). This paper uses a substring-level language model-based unsupervised segmentation method [13] that tokenizes a raw text into subwords considering frequency counts of subwords in the training data.

Also, we propose a subword segmentation method considering not only language models but also accent phrase boundaries of Japanese. The language model-based segmentation is not appropriate for prosody-aware embedding because accent types are basically independent among Japanese accent phrases (language units consisting of some words) and a one subword corresponding to parts of multiple accent phrases degrades the prediction accuracy of F_0. Therefore, assuming that the accent phrase boundaries of raw texts are given in the training data, we build a language model excluding subwords corresponding to multiple accent phrases. For example, we assume that 'aaabbbccc' is a raw text and the phrase boundaries are between 'a' and 'b,' and 'b' and 'c.' We calculate frequency counts of subwords 'aa,' 'bb,' and 'cc' but not those of 'ab' and 'bc.' This calculation makes likelihoods of 'ab' lower, and finally the built subword model splits them into 'aaa,' 'bbb,' and 'ccc.'. Table 1 lists an example of a raw text and its subwords. We can see that considering accent phrase boundaries ('Subwords (acc)') can avoid making subwords corresponding to multiple accent phrases.

B. Mora-informed modulation filtering

Japanese is a mora-timed language, which means it has mora (sub-syllable) isochrony, and has two types ("high" and "low") of mora-level accents. Consequently, when using a F_0 sequence for embedding, detailed structures except for high and low tones at even temporal intervals

TABLE I
EXAMPLE OF SUBWORD SEGMENTATION. '/' INDICATES AN ACCENT PHRASE BOUNDARY. CONSIDERING ACCENT PHRASE BOUNDARIES CAN AVOID MAKING SUBWORDS CORRESPONDING TO MULTIPLE ACCENT PHRASES.

Raw text	本当な / のかも / しれない
Subword (only language models)	本当　なの　かもしれない
Subword (w/ accent phrases)	本当　な　の　かも　しれない

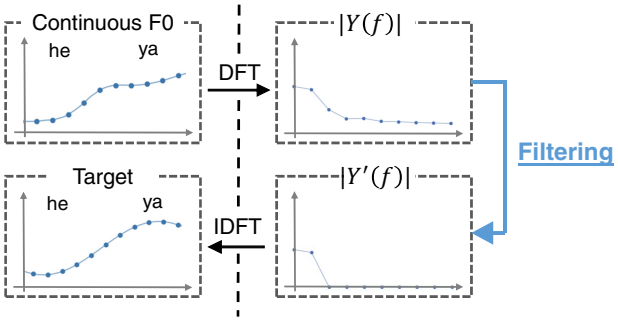

Fig. 3. The example of modulation filtering ($m = 2$). DFT indicates discrete Fourier transform.

do not need to be predicted. This removal process is implemented by modulation filtering (filtering in modulation spectrum [14] domain) to the resampled continuous F_0 sequence. The modulation spectrum is defined as the log-scaled power spectrum of the speech parameters (e.g., continuous F_0 sequence). Here, obtaining the number of moras m by grapheme-to-phoneme conversion, we suppose a T-frame resampled continuous F_0 sequence corresponding to the m-mora subword. Let $[Y(0), \cdots, Y(f), \cdots, Y(T-1)]^\top$ be its modulation spectra. $Y(f)$ is a modulation frequency component at modulation frequency index f. We construct a filter $C = [C(0), \cdots, C(f), \cdots, C(T-1)]^\top$ that removes unnecessary components on the basis of the number of intra-subword moras. $C(f)$ is given as follows:

$$C(f) = \begin{cases} 1 & (f \le f_{\text{th}} \text{ or } f \ge T - f_{\text{th}}) \\ 0 & (\text{otherwise}) \end{cases}$$
$$f_{\text{th}} = \begin{cases} 0 & (m = 1) \\ \frac{m+1}{2} & (\text{otherwise}) \end{cases} \quad (1)$$

Given the filtered modulation spectrum $Y(f)' = Y(f)\,C(f)$, the continuous F_0 sequence for embedding is gained as inverse discrete Fourier transform of $Y(f)'$. This filtering preserves at least m peaks/valleys (i.e., "high" and "low") of the F_0 contour for the m-mora subword. Fig. 3 shows an example of the two-mora case. Removing second and the higher modulation frequency components can remove the detailed changes of continuous F_0 while holding high or low tones at even temporal intervals.

IV. DNN-BASED MULTI-DIALECT SPEECH SYNTHESIS

This section applies methods proposed in **Section III** to multi-dialect speech synthesis and proposes mixing-dialect subword models and dialect-conditioned embedding models.

A. Mixing dialect subword models

If subword models trained by only common language corpora are applied to a dialect text, phrases frequently used in the dialect are finely segmented and the embedding accuracy is decreased. However, amounts of texts of one dialect are often limited to build the dialect-dependent language models. Therefore, we propose mixing-dialect subword models shared among dialects. The models are trained using texts of multiple dialects. We expect that this model can avoid unnatural subword segmentation caused by the models of the common language.

B. Dialect-conditioned subword embedding

Furthermore, we propose an embedding model conditioned by dialect information for modeling multiple dialect prosodies. The additional vector $\boldsymbol{q}_d$ for the d-th dialect is fed to the first hidden layer (first Bi-directional long short-term memories (LSTMs) in Fig. 2) of subword embedding models, and the model is trained using multi-dialect corpora. We used two types of dialect information.

Dialect codes (discrete representation): Inspired by DNN-based speech synthesis using speaker codes [15], $\boldsymbol{q}_d$ is given as a D-dimensional one-hot vector, which means 1 stands at the d-th vector component. D is the number of dialects in the training data.

Geography (continuous representation): Inspired by DNN-based speech synthesis using *i-vector* or *d-vector* [16]–[19], a continuous-valued vector is used for conditioning. Here, geographic coordinates of the dialect-speaking area are utilized as dialect information. $\boldsymbol{q}_d$ is given as $[a_d, o_d]^\top$, where a_d and o_d are geographic latitude and longitude of the central city of the d-th dialect. Because accents and pronunciations are strongly related to geographic relationships of dialects, we expect this representation to be suitable for multi-dialect speech synthesis.

V. EXPERIMENTAL EVALUATION

A. Experimental conditions

We used 15,676 utterances of the JNAS corpus [20] and 5,390 utterances of the JSUT corpus [21] for subword segmentation and embedding in common-language Japanese. The utterances contained 24,324 and 14,680 different words, respectively. We used sentencepiece [13] for language model-based subword segmentation, and set the number of subwords to 4,000 involving an unknown tag. This number was experimentally optimized in our preliminary evaluation. In the conventional method [9], word embedding strongly degraded speech quality in our evaluation. Therefore, we used subword embedding for not only the proposed method but also the conventional method. The WORLD [22] (D4C edition [23]) analysis-synthesis system was used to extract

the speech parameters and synthesize the waveform. Speech signals were sampled at a rate of 16 kHz, and the shift length was set to 5 ms. To ignore speaker differences in the corpus, continuous F_0 sequences were normalized to have zero-mean unit-variance. The resampled continuous F_0 sequence length was 64. To obtain alignments between the F_0 sequence and the subword sequence for subword embedding, we independently calculated alignments between the word sequence and the phoneme sequence using fast_align [24] and between the phoneme sequence and the continuous F_0 sequence using Julius [25]. The architecture of DNNs for embedding was Feed-Forward networks that include two bi-directional LSTM hidden layers and a bottleneck layer connecting two LSTMs. We used a Rectified Linear Unit (ReLU) [26] as the activate function of the bottleneck layer. The size of the bottleneck layer was 64.

We used a single speaker's 5,390 utterances in the JSUT corpus for training acoustic models. Contextual features include 190-dimensional quinphones, 3-dimensional within-phoneme duration vectors, prosodic contexts of previous, current and subsequent subwords (total 192 dimensions), and 9-dimensional within-subword duration vectors. The acoustic models were Feed-Forward networks that include 3×512-unit ReLU hidden layers. Predicted speech features include 0th-through-39th mel-cepstral coefficients, 5-band aperiodicity values [27], [28], continuous F0, their delta features, and voiced/unvoiced flags. In training, the speech features were normalized to have zero-mean unit-variance.

B. Evaluation of prosody-aware subword embedding in speech synthesis of a common language in Japanese

First, we evaluate the effectiveness of the prosody-aware subword embedding (**Section III**) in speech synthesis of a common language in Japanese. The test set is 600 Japanese sentences randomly selected from the JSUT corpus. The set was not included in the training data.

We evaluated synthetic speech of three systems:

1) **Conventional:** conventional method [9] with 1st-through-10th DCT components
2) **Proposed:** proposed mora-informed modulation filtering
3) **Proposed (acc):** proposed mora-informed modulation filtering and accent-phrase-informed subword segmentation

1) Objective evaluation: We calculated the root mean squared error (RMSE) of normalized continuous log-scaled F_0 sequences of synthetic and natural speech. Fig. 4 shows the result. We can see that the proposed modulation filtering improves RMSE compared with the conventional method. In addition, the proposed accent-phrase-informed subword segmentation further improves the RMSE.

2) Subjective evaluation: We conducted preference AB tests to evaluate the naturalness of the synthetic speech. We presented every pair of generated utterances of the systems in random order. Fifty listeners participated in each evaluation. Fig. 5 shows the results. There is no significant difference between **Conventional** and **Proposed**, but we can see that

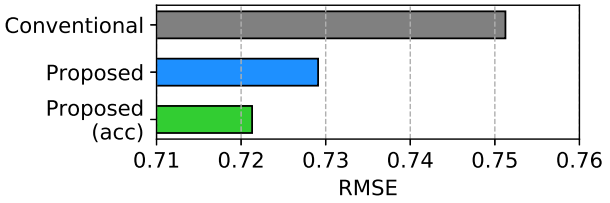

Fig. 4. RMSE of continuous log F0.

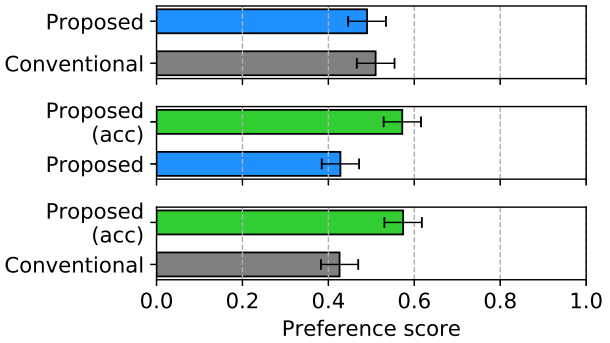

Fig. 5. Subjective evaluation results (Error bars indicate 95 % confidential interval).

the proposed accent-phrase-informed subword segmentation (**Proposed (acc)**) improves the naturalness of synthetic speech. These results demonstrate the effectiveness of the proposed method in a common language in Japanese and also demonstrate that modifying language units (i.e., subword segmentation) is more effective for improving perceptual naturalness rather than modifying F_0 (i.e., modulation filtering).

C. Performance evaluation in multi-dialect speech synthesis

Next, we evaluate the proposed multi-dialect speech synthesis (**Section IV**). The training data of subword segmentation and embedding is composed of the JNAS, JSUT, and CPJD corpora [29]. The CPJD corpus consists of text, speech and area information of 20 Japanese dialects, and 4,114 of its utterances were used for training. The JNAS and JSUT corpora were used as standard Japanese (i.e, Tokyo dialect). Twenty sentences per dialect, which are not included in the training data were used for evaluation. We evaluated synthetic speech of the following systems:

1) **Common:** subword/embedding models trained using common-language corpora (JNAS and JSUT)
2) **Dialect Code:** proposed method with dialect codes
3) **Geography:** proposed method with geographic contexts

Because estimating accent phrase boundaries of dialect texts is an unsolved problem in Japanese (this is our future work), we apply only modulation filtering to three systems.

We conducted preference AB tests on the naturalness of the synthesized dialect speech. We recruited native listeners who had lived in the dialect-speaking area for more than three

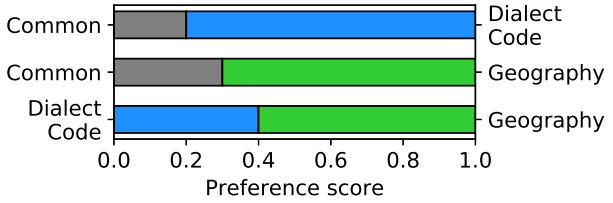

Fig. 6. Subjective evaluation results of Miyazaki-ben.

TABLE II
SUMMARY OF EVALUATION IN MULTI-DIALECT SPEECH SYNTHESIS. THE NUMBERS IN THE CENTER CELLS ARE THE COUNTS OF SELECTED SYSTEMS. IN THE CASE OF "COMMON" VS. "DIALECT CODE," NATIVE LISTENERS OF FOUR DIALECTS PREFERRED SPEECH SAMPLES OF "DIALECT CODE" MORE THAN THOSE OF "COMMON."

Method A			Method B
Common	8	4	Dialect Code
Common	7	5	Geography
Dialect Code	5	6	Geography

years. It was difficult to collect native listeners for all of the 20 dialects, but we finally found one listener for each of 12 dialects (Hokkaido-ben, Toshuu-ben, Kyo-kotoba, Osaka-ben, Nara-ben, Okayama-ben, Hiroshima-ben, Tosa-ben, Iyo-ben, Awa-ben, Fukuoka-ben, and Miyazaki-ben). Each native listener evaluated 20 randomly presented synthetic speech samples of his/her dialect. Table II summarizes the results, and Fig. 6 shows the result for one dialect (Miyazaki-ben). We can see that proposed methods were preferred in four (**Dialect Code**) or five (**Geography**) dialects more than common-language models (**Common**). Also, **Geography** is slightly better than **Dialect Code**. These results suggest the proposed multi-dialect speech synthesis is effective in some dialects.

VI. CONCLUSION

This paper presented prosody-aware subword embedding considering Japanese intonation systems and its application to multi-dialect speech synthesis. To alleviate the out-of-vocabulary problem and inaccurate prosody information, we proposed prosody-aware subword embedding with mora-informed modulation filtering and accent-phrase-informed subword segmentation. We further extended the proposed method to multi-dialect speech synthesis using mixing-direct subword models and dialect-conditioned subword embedding. The proposed methods were found to be effective for not only Japanese but also multi-dialect speech synthesis. For future work, we will investigate the use of other types of dialect information and compare proposed methods with a dictionary-based approach [8] in rich-resourced languages.

Acknowledgements Part of this work was supported by SECOM Science and Technology Foundation, and JSPS KAKENHI Grant Number JP17H06101.

REFERENCES

[1] H. Zen, K. Tokuda, and A. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

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

[3] 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.

[4] Y. Saito, S. Takamichi, and H. Saruwatari, "Training algorithm to deceive anti-spoofing verification for DNN-based speech synthesis," in *Proc. ICASSP*, Orleans, U.S.A., Mar. 2017, pp. 4900–4904.

[5] S. Takamichi, K. Tomoki, and H. Saruwatari, "Sampling-based speech parameter generation using moment-matching network," Stockholm, Sweden, Aug. 2017, pp. 3961–3965.

[6] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *Proc. ICASSP*, Shanghai, China, Mar. 2016, pp. 5540–5544.

[7] "CMU flite," http://www.festvox.org/flite/.

[8] "Open jtalk," http://open-jtalk.sp.nitech.ac.jp/.

[9] Y. Ijima, H. Nobukatsu, R. Matsumura, and T. Asami, "Prosody aware word-level encoder based on BLSTM-RNNs for DNN-based speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 764–768.

[10] T. Mikolov, M. karafiat, L. Burget, J. Cernocky, and S. Khudanpur, "Recurent neural network based language model," in *Proc. INTERSPEECH*, Chiba, Japan, Sep. 2010, pp. 1045–1048.

[11] J. Latorre, M. J. F. Gales, S. Buchholz, K. Knill, M. Tamura, Y. Ohtani, and M. Akamine, "Continuous f0 in the source-excitation generation for HMM-based TTS: Do we need voiced/unvoiced classification?" in *Proc. ICASSP*, Prague, Czech Republic, 2011, pp. 4724–4727.

[12] R. Senrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proc. ACL*, Berlin, Germany, 2016, pp. 1715–1725.

[13] "sentencepiece," https://github.com/google/sentencepiece.

[14] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti, and S. Nakamura, "Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 755–767, 2016.

[15] N. Hojo, Y. Ijima, and H. Mizuno, "An investigation of DNN-based speech synthesis using speaker codes," in *Proc. INTERSPEECH*, San Francisco, U.S.A., Sep. 2016, pp. 2278–2282.

[16] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 788–798, 2011.

[17] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. G. Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, Florence, Italy, May. 2014, pp. 4080–4084.

[18] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 879–883.

[19] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in DNN-based speech synthesis using d-vectors," in *Proc. INTERSPEECH*, Aug. 2017, pp. 3404–3408.

[20] K. Itou, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," *Journal of the Acoustical Society of Japan (E)*, vol. 20, no. 3, pp. 199–206, 1999.

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

[22] 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.

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

[24] C. Dyer, V. Chahuneau, and N. A. Smith, "A simple, fast, and effective reparameterization of IBM model 2," in *Proc. NAACL*, Atlanta, U.S.A., Jun. 2013, pp. 644–648.

[25] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine julius," in *Proc. APSIPA ASC*, Sapporo, Tokyo, Oct. 2009, pp. 131–137.

[26] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. AISTATS*, Lauderdale, U.S.A., Apr. 2011, pp. 315–323.
[27] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *MAVEBA 2001*, Firentze, Italy, Sep. 2001, pp. 1–6.
[28] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano, "Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation," in *Proc. INTERSPEECH*, Pittsburgh, U.S.A., Sep. 2006, pp. 2266–2269.
[29] S. Takamichi and H. Saruwatari, "CPJD corpus: Crowdsourced parallel speech corpus of japanese dialects," Miyazaki, Japan, Feb 2018, to appear.