10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

DNN-based Speech Synthesis Considering Dialogue-Act Information and Its Evaluation with Respect to Illocutionary Act Naturalness

Nobukatsu Hojo[1], *Yusuke Ijima*[2], *Hiroaki Sugiyama*[1],
Noboru Miyazaki[1], *Takahito Kawanishi*[1], *Kunio Kashino*[1]

[1] NTT Communication Science Laboratories
[2] NTT Media Intelligence Laboratories
nobukatsu.hojo.cd@hco.ntt.co.jp

Abstract

This study aimed at improving synthesized speech generated by a text-to-speech (TTS) system used for a spoken dialogue system in regard to how naturally the synthesized speech conveys the system's intention to the hearer. We call the measure of naturalness in this case "illocutionary act naturalness". To achieve our aim, we utilized dialogue-act (DA) information as an auxiliary feature for a deep neural network (DNN)-based speech synthesis system. First, we constructed a speech database with DA tags. Second, we used the database to build the speech synthesis system. Third, we evaluated the method by comparing its performance with a DNN making use of conventional linguistic features and hidden Markov models (HMMs) supplemented with DAs. We conducted a listening test designed to evaluate illocutionary act naturalness. The results show that the proposed method improves the illocutionary act naturalness compared with the conventional method. We also found that the illocutionary act naturalness score depended on certain features of the test sentence as well as the DA and speech synthesis method. The results shows that a test set designed by considering these features will improve the reproducibility of the illocutionary act naturalness evaluation.
Index Terms: speech synthesis, spoken dialogue systems, illocutionary act naturalness

1. Introduction

This paper describes deep neural network (DNN)-based speech synthesis considering dialogue-act (DA) information in order to improve the naturalness of synthesized speech generated by a spoken dialogue system. In communication, hearers infer a speaker's intention from every utterance [1, 2]. If a spoken dialogue system generates an utterance in an unnatural way in an attempt to express its intention, it imposes an unnecessary cognitive load for inference upon its users. Here, the intention is related to prosody [3, 4] as well as a sentence. For example, different prosodies of "Excuse me" can convey different intentions, e.g., criticism or apology. Therefore, text-to-speech (TTS) synthesis systems play a role in natural expression of intentions. On the basis of these considerations, we attempted to improve TTS in regard to how naturally the synthesized speech conveys the system's intention to the hearer.

In the field of speech synthesis, naturalness is often defined as the quality of speech samples separated from contexts [5]. On the other hand, speech needs to be natural as a way to express intention, as described above. Based on the classification by the speech act theory [6, 7], we can rephrase the conventional naturalness, naturalness of an act to say something, as "locutionary act naturalness (LAN)". We can also rephrase naturalness in this work, naturalness of an act to convey intentions, as "illocutionary act naturalness (IAN)". Our study thus aims at improving the IAN of synthesized speech.

A promising approach to improving IAN is to reproduce the prosodic features of intentions. A related field of study is emotional speech synthesis [8, 9, 10, 11, 12, 13, 14]. The common point is that para-linguistic information [15] is expressed by TTS. However, emotions and intentions have different prosodic features. Emotional speech has salient features for the whole utterance. For example, "sad" speech generally has a lower F0 and a slower speech rate [16]. On the other hand, to express intentions, for example, the sentence final tone is also important in Japanese speech [17, 18, 19]. This feature is contrastive to emotional speech in that it appears locally in time. This difference shows the necessity of determining whether emotional speech synthesis methods are also effective for expressing intentions.

Previous studies have shown the effectiveness of utilizing DAs for TTS as a way of considering intentions [20, 21]. A DA is an abstract expression of a speaker's intention [22]. However, the previous studies have not revealed two points. First, they have investigated TTS based on concatenative synthesis [20] and hidden Markov models (HMMs) [21], but not DNNs. DNNs have been shown effective for emotional speech synthesis [14], so the question here is whether they are effective at expressing intentions as well. Second, they evaluated LAN, but not IAN. We believe IAN is an important attribute of synthesized speech.

Here, we propose DNN-based TTS considering DAs and report an evaluation with respect to IAN. In particular, we used DA as an auxiliary feature for feed-forward DNNs. Although sequence modeling with neural networks is effective for emotional speech synthesis, the limited size of the speech corpuses have prevented it from being used [12, 13, 14, 23]. The performance of the proposed method was compared with a DNN making use of only conventional linguistic features and hidden Markov models (HMMs) supplemented with DAs.

We also describe the design of a test set that is useful for deriving results with sufficient reproducibility. The problem is that evaluations of IAN may have low reproducibility because they depend on the selection of test-set sentences. The cause of this dependency is that IAN is likely to be affected by the sentences as well as the TTS methods, because different sentences may need different prosodic features of an intention (e.g. sentence final tones depend on sentence final particles in Japanese [17, 18, 19]). One way to alleviate this dependency is to use a larger test set, though this increases the cost of any evaluation experiment. Another approach is to make an assumption about which features of a sentence affect IAN and then design the test set by controlling the frequencies of those features in it. The evaluation experiment in this study is based on the latter approach. The validity of the design of the test set was examined by conducting a two-way analysis of variance (ANOVA) on the results of a subjective evaluation.

2. DNN-based TTS Considering DA

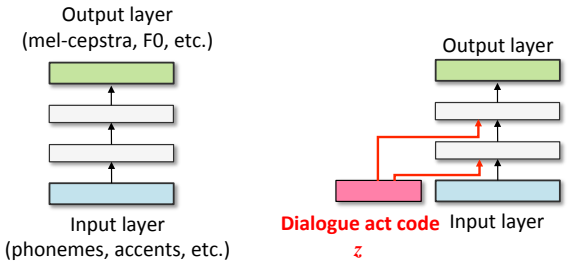
2.1. Speech Database with DA Tags [24]

To build a TTS system that can consider DAs, first, we constructed a speech database with DA tags. As a DA set for the experiments, we used the one proposed in [25]. This set is designed to cover a wide range of utterances of non-task oriented

10.21437/SpeechProsody.2020-199

Speaker	DA	Utterance
A		Is there a convenience store in your neighborhood?
B		There are 3 7-Elevens.
A	ADMIRATION	So many!

Figure 1: *Example of recording manuscript used to make the speech database. Voice for gray colored utterances was recorded. Preceding utterances were displayed to show conversational context. The dialogue was translated from Japanese by the authors.*



(a) *Conventional method.* (b) *Proposed method.*

Figure 2: *Schematic diagram of conventional and proposed methods. The same model architectures are used for the duration models and acoustic models.*

open-domain conversation and consists of 30 DAs. The sentences for the speech recording were extracted from a text chat database in Japanese. This database was originally gathered by Higashinaka et al. [26] and contains 3680 conversations (with 134K sentences). The utterances have been manually tagged with DAs by two experts.

The sentences for the recording were extracted by considering the balance of the frequency of phonemes and DAs on an entropy basis [27]. For the speech recording, we used the manuscripts illustrated in Fig. 1. The manuscript shows the sentences for recording, their DAs, and several preceding utterances as context. We recorded the speech of a Japanese female professional voice actor. We instructed her to read the manuscript silently first so that she would understand the conversational context before each utterance. She spoke each utterance in a natural conversational speaking style for the context and corresponding DAs. We derived 5177 sentences from these recordings. We excluded duplications (utterances with the same sentence and DA) from them and used 3410 utterances, about 140 minutes in total, as the speech database. All of the utterances were manually annotated with phonemes, accent types, and phonetic boundary information so that they can be used as training data for speech synthesis models.

2.2. DNN-based TTS considering DA

The straightforward method to generate speech considering DA is to train a model for each DA and to switch the model according to the DA when synthesizing. However, this method needs a considerable amount of training data for each DA to derive high-quality speech, which increases costs. The proposed method utilizes DAs as input for the DNN, which enables a single DNN to model the speech of all of the DAs. Conventional HMM-based methods have succeeded in generating speech from a small amount of speech data [8, 28, 29]. However, their performance is limited because their tree structure restricts the relationship between the inputs and outputs that can be modeled. Our DNN-based method does not have such a restriction in the model architecture, so we expected that it would be able to reproduce prosodic features of DAs more precisely than HMM-based methods.

Figure 2 shows a schematic diagram of the proposed method. The architecture is the same as the one of DNN-based speech synthesis using speaker codes [30], except that the auxiliary feature z indicates a DA, not the speaker ID. We used 1-hot code as the DA code z [30].

3. Evaluation Experiments

3.1. Speech Samples

We constructed five different TTS systems for the evaluation:

- •HMM-BASELINE: HMM making use only of conventional linguistic features [31],
- •HMM-MIXED: HMM-based style-mixed modeling method [8],
- •HMM-ADAPT: HMM-based average model and model adaptation method [28, 29],
- •DNN-BASELINE: DNN making use only of conventional linguistic features [32],
- •DNN-DACODE: The proposed DNN-based method using DA codes.

We compared the proposed method with conventional HMM-based methods, listed as HMM-MIXED, HMM-ADAPT and DNN-DACODE. We also compared DNN-BASELINE and DNN-DACODE to validate the effect of DAs derived by DNNs. We also evaluated HMM-BASELINE to validate the effect of DAs derived by HMMs.

For all five methods, we used acoustic features consisting of mel-cepstra, log F0, and band aperiodicities and their dynamic features, extracted at 5 ms intervals by using STRAIGHT [33]. For the DNN-based methods, voiced/unvoiced binary variables were added to the acoustic features. The number of dimensions of the acoustic features was 138 for the HMM-based methods and 139 for the DNN-based ones. The number of dimensions of the linguistic features for the DNNs was 486 for the duration models and 489 for the acoustic models. We used an affine-transform post filter [34] for the mel-cepstra sequence.

For DNN-BASELINE and DNN-DACODE, the duration models had two hidden layers with 256 sigmoid units each, while the acoustic models had four hidden layers with 256 sigmoid units each. 30-dimensional DA codes were input to all the hidden layers of DNN-DACODE. The parameters were updated to minimize the mean square error by using Adam algorithm [35]-based back-propagation. The learning rate was set to $\alpha = 0.00015$ for the acoustic models and 0.001 for the duration models.

For the HMM-based methods, we used five-state left-to-right multi-space probability distribution hidden semi-Markov models (MSD-HSMMs) without skip. The MDL parameter was set to $\alpha = 1.0$. For HMM-MIXED, we used questions with respect to DAs as well as linguistic information. We used 30 questions, each of which corresponded to a DA [8]. We added 8 questions, each of which corresponded to a DA class described in sec. 3.3. The total number of questions related to DAs was 38. For HMM-ADAPT, we trained an average voice model by using the speech data of all the DAs. Then we adapted the average voice model to each DA by using the speech data of the corresponding DA. We used a combination of constrained structural maximum a posteriori linear regression (CSMAPLR) and MAP adaptation [29].

We selected 120 utterances from the speech corpus as the test set, as described in sec. 3.3. The remaining utterances were used as training data for all five methods.

3.2. Evaluation Method of IAN

We conducted a subjective evaluation of IAN. Since conversational context is important in perceiving intentions, we displayed the context in a GUI (Fig. 3). We instructed the partic-

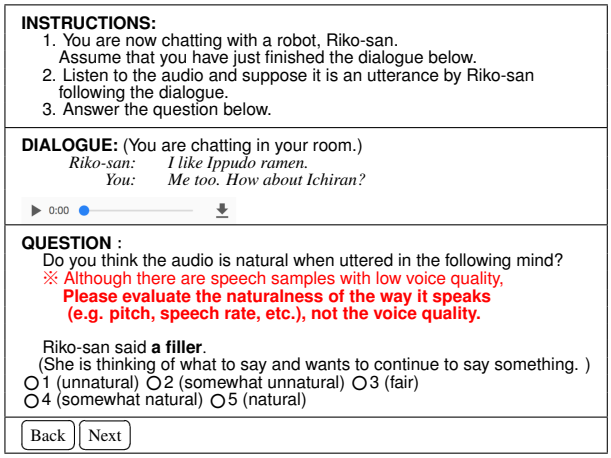


Figure 3: *Example of GUI used in the subjective evaluation experiments of IAN. The sentence of the speech sample is "so desune (Let's see)".*

Table 1: *The DA classes used in the evaluation experiments. The right column has 30 DAs in total because "self-disclosure" and "question" have 8 and 7 subclasses, respectively [25].*

DA Class	DA
INFORMATION	information, self-disclosure, sympathy, non-sympathy, approval
QUESTION	question (other than "question_self"), confirmation, proposal
QUESTION_SELF	question_self
REPEAT·PARAPHRASE	repeat, paraphrase
ACKNOWLEDGEMENT	acknowledgement
FILLER	filler
ADMIRATION	admiration
COURTESY	greeting, thanks, apology

ipants to assume that they were chatting with a robot and then to read the displayed context. They then listened to each audio sample while assuming that it was an utterance by the robot following the context. They evaluated how naturally the samples conveyed the reference intention. Describing the intention with a single word (e.g. "filler") would be hard for the participants to understand, and this might degrade the reliability of the evaluation's results [24]. Therefore, we displayed a description of the definition of the reference intention (e.g. "She is thinking what to say and wants to continue saying something") as well. We recruited 20 native Japanese-speaking participants (10 males and 10 females) from outside the authors' organization. The experiments were conducted using headphones and the GUIs illustrated in Fig. 3

3.3. Design of Evaluation Set

Conventional evaluations (e.g. [8]) often design a test set by randomly selecting sentences from a corpus. However, this procedure can produce results with low reproducibility for IAN because they depend on random selection of sentences. The cause of this dependency is that IAN is likely to be affected by the sentences as well as by the TTS method, as different sentences may need to reproduce different prosodic features of an intention. Therefore, we made an assumption about which features of a sentence affect IAN. Then, we designed the test set by controlling the frequencies of those features in it.

We assumed that two features of a sentence affect IAN. One feature is the sentence structure (presence of main clause, predicates, and POS other than interjection in the sentence). "Complete" sentences have a main clause (e.g. "それは良かったですね。/It was great for you. [self-disclosure_plus]"). "Suspended" sentences have no main clause, but do have predicates (e.g. "今日はよく働いたから。/Because I worked a lot today. [self-disclosure_fact]"). "Predicate-omitted" sentences have no main clause and no predicates, but do have POS other than interjections (e.g. "香川が？/Kagawa? [repeat]"). "Interjection" sentences have only interjections (e.g. "うーん。/Umm. [filler]"). Apart from the classification above, "Courtesy" sentences were separately classified because they have fixed expression for some DAs (e.g. "ありがとう。/Thank you. [thanks]"). The other feature is a word at the end of a sentence, a sentence end particle of "complete" sentences ("-ne", "-ka", etc.) or a connecting particle at the end of "suspended" sentences ("-kedo", "-noni", etc.); these are considered important for expressing intentions [36]. Sentences without particles were classified as e.g. "complete_NONE". We classified sentences based on these two features, which resulted in 32 sentence classes in total.

Were we to evaluate all of the combinations of 30 DAs and 32 sentence classes, it would increase the cost of the evaluation. Moreover, not all of the combinations are equally important because there are frequent/more-important cases and rare/less-important cases; e.g., for DA of "self-disclosure", we often use "complete" sentences, but hardly use "interjection" sentences. To reduce the cost of the evaluation, we decided to

1. classify DAs on the basis of their semantic similarity into 8 classes (Tab. 1),
2. use the top 3 most frequent sentence classes for each DA class in the evaluation.

For the 2nd procedure, we investigated the frequency of each sentence class for each DA class in a text chat corpus [26]. The derived top 3 most frequent sentence classes are displayed as labels on the x-axis in Fig. 4. For each sentence class of each DA class, we used 5 sentences that were randomly selected from the text chat corpus. Since there were 3 sentence classes for each of the 8 DA classes, the test set included 120 (=$5 \times 3 \times 8$) sentences in total.

4. Results

4.1. IAN's Dependency on Sentence Class

Fig. 4 shows the results of the subjective evaluation. First, we tested whether there were interaction effects of TTS methods and sentence classes. Concretely, for each DA class, a repeated-measures two-way ANOVA was carried out on the factors TTS methods ("METHOD") and sentence classes ("SENTENCE"). The text boxes above the bar plots in Fig. 4 show the results of the analysis. The ANOVA revealed that "SENTENCE" and the interaction between TTS methods and sentence classes ("INTERACTION") had a significant effect on IAN in 7 out of 8 DA classes. These results confirm that the sentence classes in this experiment actually affected the results of the evaluation of IAN.

The ANOVA results indicate that the design of the test set affects the evaluation of IAN. In other words, reproducibility will be degraded when IAN is evaluated in accordance with the conventional procedure that randomly selects utterances for the test set. For example, suppose we compare IAN of DNN-BASELINE and DNN-DACODE for "ACKNOWLEDGEMENT" in Fig. 4; sentence classes 1 and 3 have smaller differences, but sentence class 2 has a larger one. Then, the results of an evaluation using the conventional procedure may vary depending on the proportion of class 2 sentences in the test set. Therefore, the proportion should be controlled when the test set is designed.

4.2. Comparison of TTS methods

Since the interaction effects between TTS methods and sentence classes were significant, we tested simple main effects.

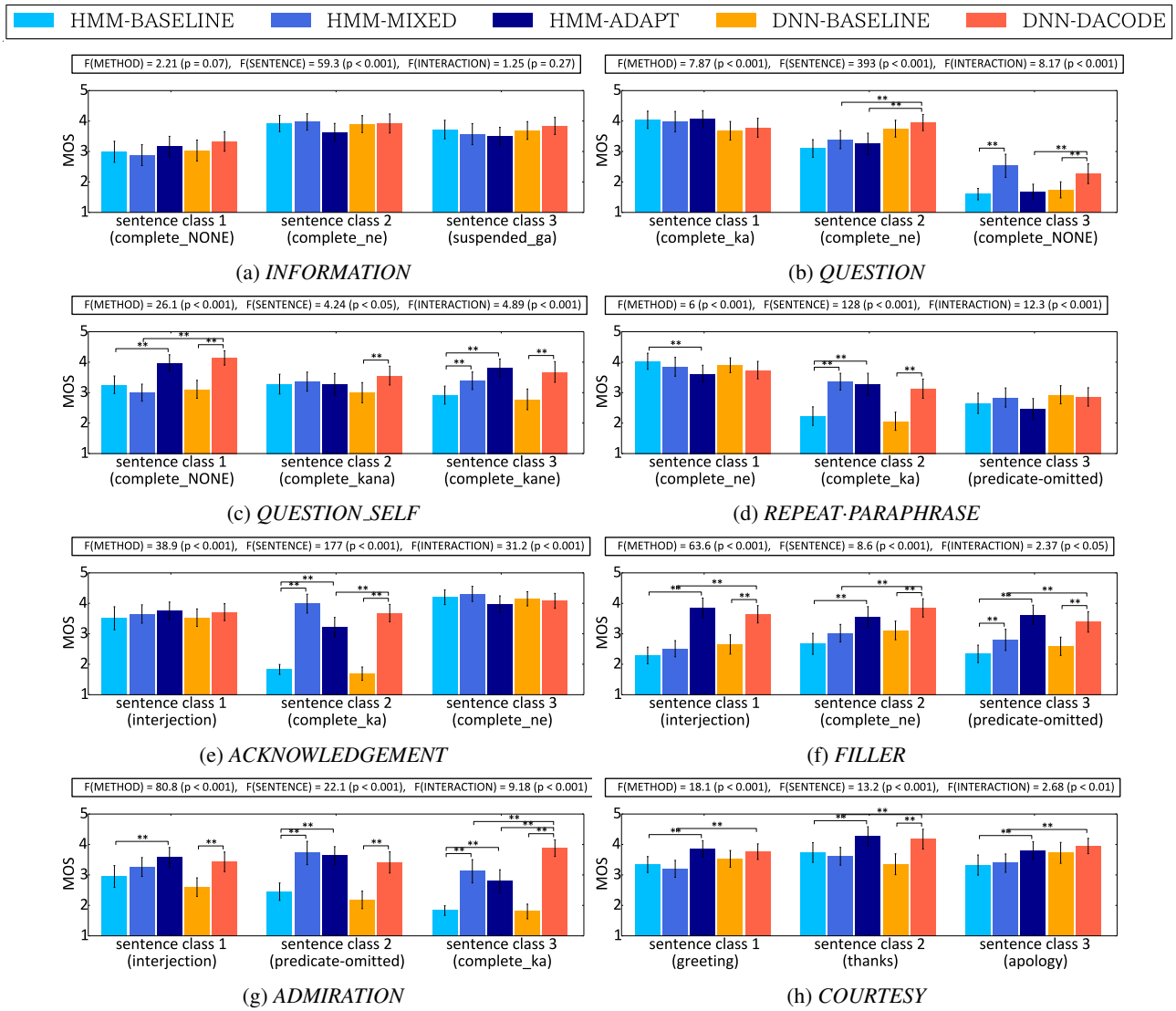


Figure 4: *Results of subjective evaluation with respect to illocutionary act naturalness. F-values and p-values calculated by two-way ANOVA are displayed above the graph for each DA. Error bars display 99% confidence intervals. '**' denotes a significant difference confirmed by t-tests ($\alpha = 0.01$) between (1) HMM-BASELINE and HMM-MIXED, (2) HMM-BASELINE and HMM-ADAPT, (3) HMM-MIXED and DNN-DACODE, (4) HMM-ADAPT and DNN-DACODE, (5) DNN-BASELINE and DNN-DACODE.*

In particular, we conducted a t-test ($\alpha = 0.01$) between the TTS methods for each sentence class and DA class. The results are plotted in Fig. 4.

Comparing DNN-DACODE with HMM-MIXED and HMM-ADAPT, we see that it is superior or comparable to the conventional methods for all of the DA and sentence classes. These results show the superiority of the proposed method to the HMM-based conventional methods. Moreover, comparing DNN-DACODE and DNN-BASELINE, we see that the proposed method is superior or comparable to the conventional one for all of the DA and sentence classes. The results show the effectiveness of considering DA.

The MOS scores of the five methods were comparable for some DA and sentence classes, but different for others. This shows that DA is helpful for some sentences, but not so much for others. It would be interesting to determine why this tendency exists by comparing the MOS scores with objective and other subjective measures.

5. Conclusions

This study aimed at improving TTS in regard to how naturally the synthesized speech conveys the system's intention, or its "illocutionary act naturalness" (IAN). For this purpose, we utilized DAs as an auxiliary feature in a DNN-based speech synthesis system. We constructed a speech database with DA tags and built five TTS systems, one of which incorporated the proposed method. We conducted a listening test that was designed to evaluate IAN. The results showed that the proposed method improved the illocutionary act naturalness compared with the conventional methods. We also found that the MOS results depend on certain features of the sentences included in the test set. Therefore, to ensure that evaluations of IAN are reproducible, we should design test sets by considering the frequencies of those features in them. Our future work will include analyzing the results by comparing them in terms of objective and other subjective measures. We will also investigate sequence models to further improve IAN.

6. References

[1]H. P. Grice, *Studies in the Way of Words*. Harvard University Press, 1991.

[2]D. Wilson and D. Sperber, *Meaning and Relevance*. Cambridge University Press, 2012.

[3]K. Maekawa, "Phonetic and phonological characteristics of paralinguistic information in spoken Japanese," in *Proc. Fifth International Conference on Spoken Language Processing*, 1998.

[4]N. Hellbernd and D. Sammler, "Prosody conveys speaker's intentions: Acoustic cues for speech act perception," *Journal of Memory and Language*, vol. 88, pp. 70–86, 2016.

[5]ITU-T, *A Method for Subjective Performance Assessment of the Quality of Speech Voice Output Devices*, International Telecommunication Union Std., 1994.

[6]J. L. Austin, *How to Do Things with Words*. Oxford University Press, 1975.

[7]J. R. Searle, F. Kiefer, and M. Bierwisch, *Speech Act Theory and Pragmatics*. Springer, 1980, vol. 10.

[8]J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 88, no. 3, pp. 502–509, 2005.

[9]J. Yamagishi, T. Kobayashi, M. Tachibana, K. Ogata, and Y. Nakano, "Model adaptation approach to speech synthesis with diverse voices and styles," in *Proc. of ICASSP 2007*, vol. 4, 2007, pp. 1233–1236.

[10]T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for HMM-based expressive speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 9, pp. 1406–1413, 2007.

[11]R. Barra-Chicote, J. Yamagishi, S. King, J. M. Montero, and J. Macias-Guarasa, "Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech," *Speech Communication*, vol. 52, no. 5, pp. 394–404, 2010.

[12]S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using LSTM-RNNs," in *Proc. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1613–1616.

[13]J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, Y. Morino, and Y. Ochiai, "Investigating different representations for modeling and controlling multiple emotions in DNN-based speech synthesis," *Speech Communication*, vol. 99, pp. 135–143, 2018.

[14]L. Xue, X. Zhu, X. An, and L. Xie, "A comparison of expressive speech synthesis approaches based on neural network," in *Proc. the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-Modal Affective Computing of Large-Scale Multimedia Data*, 2018, pp. 15–20.

[15]H. Fujisaki, "Prosody, models, and spontaneous speech," in *Computing prosody*. Springer, 1997, pp. 27–42.

[16]K. R. Scherer and J. S. Oshinsky, "Cue utilization in emotion attribution from auditory stimuli," *Motivation and Emotion*, vol. 1, no. 4, pp. 331–346, 1977.

[17]E. Ofuka, J. D. McKeown, M. G. Waterman, and P. J. Roach, "Prosodic cues for rated politeness in Japanese speech," *Speech Communication*, vol. 32, no. 3, pp. 199–217, 2000.

[18]Y. Katagiri, "Dialogue functions of Japanese sentence-final particles 'Yo' and 'Ne'," *Journal of Pragmatics*, vol. 39, no. 7, pp. 1313–1323, 2007.

[19]K. Iwata and T. Kobayashi, "Expression of speaker's intentions through sentence-final particle/Intonation combinations in Japanese conversational speech synthesis," in *Proc. Eighth ISCA Workshop on Speech Synthesis*, 2013.

[20]A. K. Syrdal and Y.-J. Kim, "Dialog speech acts and prosody: Considerations for TTS," in *Proc. Speech Prosody 2008*, 2008, pp. 661–665.

[21]P. Tsiakoulis, C. Breslin, M. Gasic, M. Henderson, D. Kim, M. Szummer, B. Thomson, and S. Young, "Dialogue context sensitive HMM-based speech synthesis," in *Proc. ICASSP 2014*, 2014, pp. 2554–2558.

[22]A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. V. Ess-Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," *Computational Linguistics*, vol. 26, no. 3, pp. 339–373, 2000.

[23]R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. J. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," *arXiv preprint arXiv:1803.09047*, 2018.

[24]N. Hojo and N. Miyazaki, "Evaluating intention communication by TTS using explicit definitions of illocutionary act performance," *Proc. INTERSPEECH 2019*, pp. 1536–1540, 2019.

[25]T. Meguro, R. Higashinaka, Y. Minami, and K. Dohsaka, "Controlling listening-oriented dialogue using partially observable markov decision processes," in *Proc. the 23rd international conference on computational linguistics*. Association for Computational Linguistics, 2010, pp. 761–769.

[26]R. Higashinaka, K. Imamura, T. Meguro, C. Miyazaki, N. Kobayashi, H. Sugiyama, T. Hirano, T. Makino, and Y. Matsuo, "Towards an open-domain conversational system fully based on natural language processing." in *Proc. COLING*, 2014, pp. 928–939.

[27]T. Nose, Y. Arao, T. Kobayashi, K. Sugiura, and Y. Shiga, "Sentence selection based on extended entropy using phonetic and prosodic contexts for statistical paramaetric speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1107–1116, 2017.

[28]M. Tachibana, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style adaptation technique for speech synthesis using HSMM and suprasegmental features," *IEICE transactions on information and systems*, vol. 89, no. 3, pp. 1092–1099, 2006.

[29]J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions. Audio, Speech, and Language Process.*, vol. 17, no. 1, pp. 66–83, 2009.

[30]N. Hojo, Y. Ijima, and H. Mizuno, "DNN-based speech synthesis using speaker codes," *IEICE TRANSACTIONS on Information and Systems*, vol. 101, no. 2, pp. 462–472, 2018.

[31]K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP 2000*, vol. 3, 2000, pp. 1315–1318.

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

[33]H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[34]H. Silén, E. Helander, J. Nurminen, and M. Gabbouj, "Ways to implement global variance in statistical speech synthesis," in *Proc. INTERSPEECH 2012*, 2012, pp. 1436–1439.

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

[36]S. Makino and M. Tsutsui, "A dictionary of basic Japanese grammer." *The Japan Times, Ltd.*, 1986.