*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# Neural Spoken-Response Generation Using Prosodic and Linguistic Context for Conversational Systems

*Yoshihiro Yamazaki*[1], *Yuya Chiba*[2], *Takashi Nose*[1], *Akinori Ito*[1]

[1] Graduate School of Engineering, Tohoku University, Japan
[2] NTT Communication Science Laboratories, Japan

yoshihiro.yamazaki.t2@dc.tohoku.ac.jp, yuuya.chiba.ax@hco.ntt.co.jp,
takashi.nose.b7@tohoku.ac.jp, aito@spcom.ecei.tohoku.ac.jp

## Abstract

Spoken dialogue systems have become widely used in daily life. Such a system must interact with the user socially to truly operate as a partner with humans. In studies of recent dialogue systems, neural response generation led to natural response generation. However, these studies have not considered the acoustic aspects of conversational phenomena, such as the adaptation of prosody. We propose a spoken-response generation model that extends a neural conversational model to deal with pitch control signals. Our proposed model is trained using multimodal dialogue between humans. The generated pitch control signals are input to a speech synthesis system to control the pitch of synthesized speech. Our experiment shows that the proposed system can generate synthesized speech with an appropriate F0 contour as an utterance in context compared to the output of a system without pitch control, although language generation remains an issue.

## 1. Introduction

Sociable conversational dialogue systems are expected to improve the user experience in many domains [1, 2]. In recent years, neural-based response generation has been studied for dialogue systems [3–6], and the fruits of these studies have significantly improved the naturalness of language generation in dialogue systems.

In addition to linguistic expressions, it is necessary to consider various aspects of dialogue phenomena to produce human-like spoken dialogue. One such dialogue phenomenon is the entrainment of prosody [7], where speakers adapt their speaking style to that of their dialogue partners. Entrainment occurs in several factors of speech, such as energy [7, 8], fundamental frequency [9, 10], and speech rate [11]. Another phenomenon in spoken dialogue is placing emphasis on a specific word. In human-human dialogue, speakers typically deaccent items that represent given information in discourse [12, 13]. Furthermore, phrases beginning new topics tend to have specific characteristics of prosody, such as a wider pitch range [13]. Therefore, a response generation model for a spoken dialogue system should consider both linguistic expressions and prosodic information of the dialogue context to replicate such phenomena.

Neural-based approaches have also been developed in the area of speech synthesis [14, 15]. For speech synthesis based on deep learning, Yamada et al. [16, 17] improved parametric speech synthesis based on a deep neural network (DNN) [18] to control the pitch of synthesized speech. This method can customize the pitch of arbitrary segments by using pitch control signals that compose what is called the differential F0 context.

We propose a spoken-response generation method to generate system utterances that have pitch contours close to those of human-human dialogue. Our proposed method is based on the encoder-decoder model, and it converts the linguistic and prosodic information of a user utterance into a system utterance and pitch control signals. The speech synthesizer based on the differential F0 context can control the pitch word by word. Therefore, the network can train a local relationship between a word sequence and prosody by combining the language and prosody generation models.

## 2. Related Studies

### 2.1. Spoken dialogue systems with prosodic control

Several groups have developed systems that consider the prosodic aspect of a system utterance [19–24]. They reported that adapting the pitch, intensity, and speech rates of system utterances to users improved how users evaluated them in terms of naturalness, friendliness, and reliability. Recently, Fuscone et al. [25] proposed a neural-network-based approach to estimating the statistics of the acoustic parameters of an upcoming utterance from the preceding utterances. However, that study only controlled prosody-related parameters at the sentence level, and the generated speech did not always display prosody variation close to the speech of a human-human conversation.

In our study, we control the pitch of the generated utterance word by word. Thus, the proposed system can generate utterances with more human-like prosody that fits the context of the conversation.

### 2.2. Conversational speech synthesis

Conversational corpora have been used to improve the expressiveness of synthesized speech [26]. Székely et al. [27] introduced transfer learning by using large-scale podcast data for end-to-end speech synthesis [14] trained using a read speech corpus. This method resulted in natural speech synthesis for public speaking and casual conversation. However, the synthesized speech did not always fit the conversation because speech was synthesized without dialogue context. Several studies have introduced dialogue-related information to speech synthesis, such as special tokens to represent emphasis [28] and dialogue acts [29]. In addition, Guo et al. [30] proposed an end-to-end model using the dialogue history up to 10 turns. However, it is difficult for these approaches to generate synthesized speech that incorporates the prosodic information of user speech because they employ only linguistic features.

### 2.3. Pitch control using differential F0 context

Yamada et al. proposed a speech synthesis system with prosody control [16, 17], which applies DNN-based parametric speech



http://dx.doi.org/10.21437/Interspeech.2021-381

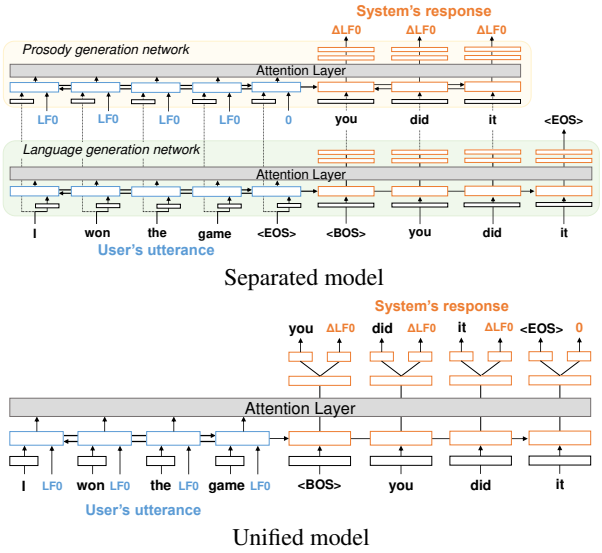


Figure 1: *Overview of spoken response generation network.*

synthesis [18]. This method is called *tailor-made speech synthesis*, which enables the user of a speech synthesizer to manually adjust the pitch of the synthesized speech. To this end, this method appends additional information called *differential F0 context* to the word sequence. The differential F0 context is the relative amount of pitch change. For example, giving the differential F0 context of $+1.0$ or $-1.0$ to a word makes the pitch of the word higher or lower, respectively.

When training the model, we first prepare spoken sentences, and then the references of the differential F0 context for those utterances are calculated by subtracting the F0 of the synthesized speech signals of the same content from that of the original utterances.

While the conventional prosody control method assumes manual control, we predict the differential F0 context (the pitch control signal) so that the prosody of the synthesized speech becomes similar to that of real conversation. Our proposed method predicts the pitch control signals using a user's prosodic information in addition to the linguistic information. Therefore, we expect the neural network to generate a pitch that can emphasize the focus or entrainment based on bi-modal (linguistic and prosodic) information without additional annotation.

## 3. Neural Spoken Response Generation

Figure 1 illustrates our proposed spoken response generation models. We examined the two architectures shown in the figure.

The first architecture is the *separated model*, shown in the upper panel of the figure. This model consists of two encoder-decoder networks: a language generation network and a prosody generation network. The language generation network takes the word sequence of the user utterance as input and generates the word sequence of the system utterance. In the prosody generation network, the word sequence and average log F0 sequence of the user utterance are input to the encoder to output a representation vector. The decoder takes this representation vector and the word sequence of the system utterance as the input used to output the differential F0 context sequence. With this model, the language generation network can be substituted for another recent state-of-the-art language model (e.g., [6]).

The second architecture is the *unified model* shown in the lower panel of the figure. This model trains the relationship between the linguistic and prosodic information using a single encoder-decoder network. The encoder converts the word and average log F0 sequences into a single representation vector. The decoder takes this representation vector as input and generates the word and differential F0 sequences simultaneously. This model is expected to effectively train the relationship between the modalities by converting the linguistic and prosodic information using a single encoder-decoder.

Finally, the generated word and differential F0 context sequence are fed into the speech synthesizer to generate the response speech.

## 4. Experimental Conditions

In our experiments, we compared the synthesized speech produced by the proposed system with that produced by a system without pitch control (*baseline model*). In addition, we evaluated the effectiveness of the linguistic and prosodic features of a preceding utterance by changing the input to the network.

### 4.1. Corpora

We used Spontaneous Multimodal One-on-one Chat-talk (SMOC) [31], a corpus of Japanese multimodal chat-talk with 107 speakers (female: 33, male: 74), to train the networks. This corpus contains video and speech data with transcriptions of dyadic conversation. The data is made up of dialogue by 102 pairs of speakers for a total of 56 hours of speech.

The dialogue was separated into utterance pairs of the two speakers. Then, the experimental data were separated into training, development, and test sets so that these sets would not share the same speaker. The numbers of pairs in the training, development, and test sets were 22,746, 3,425, and 3,197, respectively.

In addition, we used five million Japanese tweet-reply pairs collected from Twitter from June 2 to December 18, 2019, for pre-training of the language generation model.

### 4.2. Extraction of pitch control signals

First, F0 was extracted using Harvest [32] included in the WORLD vocoder [33] with a 5-ms frame shift. The word-by-word average of log F0 was extracted from the preceding utterances, based on the time information of the transcriptions. By contrast, differential F0 contexts were calculated from succeeding utterances. To calculate the differential F0 contexts, speech of the same sentence as a target utterance was synthesized in advance. The word-by-word average log F0 values were calculated for both target and synthesized speech. Then, the average log F0 sequence of the synthesized speech was subtracted from that of the target speech to obtain the differential F0 contexts. Here, the average log F0 sequences were normalized by using the average of the speaker to remove the speaker difference.

### 4.3. Training of spoken response generation model

We used the Long Short-Term Memory (LSTM) and the Bidirectional LSTM (BLSTM) for the encoder and decoder of the unified model. In the separated model, LSTM was used only for the decoder of the language generation network, while BLSTM was used for the other parts. We stacked both the LSTM and BLSTM in 2 layers. These encoders and decoders also had an attention mechanism. The number of units for the hidden layers and the embedding layer was 1,024. In the training step,

the batch size was set to 64, and the gradient clipping was set to 5.0. The dropout rate was 0.2. The optimization algorithm was Adam with learning parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rates of the encoders and decoders were 0.0001 and 0.0005, respectively.

Pre-training was conducted for the language generation network using the Twitter data. The vocabulary used in the network consisted of the 50,000 most frequent words that appeared in the pre-training data, and the out-of-vocabulary words were replaced into a single ⟨UNK⟩ token. The maximum number of epochs of pre-training was 20, and the model with the minimum validation loss was used for fine-tuning. In the pre-training step, we applied scheduled sampling [34] with a teacher forcing ratio of 0.5 every step.

After the pre-training, we fine-tuned the network using the SMOC corpus. In the fine-tuning step of the unified model, we expanded the networks to deal with pitch control signals and additional vocabulary. For the pitch control signals, the input layer of the encoder was expanded by one unit, and the output layer of the decoder by one full-connection layer and an output layer of one unit. In addition, 371 words that occurred more than four times only in the SMOC corpus were appended to the vocabulary of the embedding and the word projection layers.

In the separated model, the language generation network was fine-tuned using the transcriptions of the SMOC corpus. Then, the prosody generation network was trained using the ground-truth of the transcriptions and pitch control signals of the SMOC corpus. Here, the parameters of the embedding layers of the prosody generation network were fixed to be the same as those of the fine-tuned language generation network. The number of training epochs was fixed at 20, and the teacher forcing ratio was set to 1.0.

We used the cross-entropy loss for language generation and the mean squared error (MSE) for prosody generation. The weighted average of these losses was used for the unified model. The weights of the loss were determined by preliminary experiments. The weight of the MSE was 1.0 and that of the cross entropy loss was $10^2$.

### 4.4. Speech synthesizer with pitch control

The speech synthesis model based on differential F0 contexts was trained according to the method used by Yamada et al. [16].

The DNN-based acoustic model (*the baseline model* that has no additional pitch control) takes the 1,994-dimensional input vectors that include the relative position within the phone and the 1,993-dimensional binary answers to questions about linguistic contexts. The output is 139 dimensional vectors including 40 dimensional mel-frequency cepstrum, log F0, and five-band aperiodicity values (0–1, 1–2, 2–4, 4–6, and 6–8 kHz) with $\Delta$ and $\Delta\Delta$ coefficients, and the voiced/unvoiced flag. The network was the three-layered Multilayer Perceptron (MLP) with 1,024 hidden units. In the training step, we set the dropout rate to 0.5 and the batch size to 100.

The input of the acoustic model with pitch control was the differential log F0 of the preceding, current, and succeeding words in addition to the same 1,994-dimensional features as the base acoustic model. The output was the 138-dimensional differential acoustic features excluding the voiced/unvoiced flag. The hyperparameters were the same as those of the baseline model.

We used the 4,900 Japanese utterances by the single female speaker contained in the basic5000 subset of JSUT [35] as the training data.

Table 1: *Objective evaluation results: L and P denote that the word sequence and the average log F0 sequence were used as the encoder input. The metrics marked with * were calculated from the results generated by teacher forcing.*

(a) Objective evaluation of language generation.

| Model | Modality | PPL* | BLEU-4 | Dist-2 |
|---|---|---|---|---|
| Unified model | L | 258.16 | 1.61 | 21.31 |
| | P | **97.60** | 0.93 | 0.22 |
| | L+P | 257.09 | 1.70 | **21.33** |
| Separated model (Language generation) | L | 255.38 | **1.78** | 20.08 |

(b) Objective evaluation of prosody generation.

| Model | Modality | ΔLF0 MSE* | LF0 RMSE* [cent] |
|---|---|---|---|
| Unified model | L | 0.0615 | 440.7 |
| | P | 0.0606 | 439.9 |
| | L+P | 0.0618 | 429.7 |
| Separated model (Prosody generation) | L | 0.0501 | 404.7 |
| | P | 0.0503 | 405.8 |
| | L+P | **0.0492** | 401.7 |
| | w/o Encoder | 0.0505 | **391.9** |
| Baseline | – | – | 433.3 |

### 4.5. Conditions of objective and subjective evaluations

We compared two models (the unified and separated models) to evaluate language generation and three models (the baseline, unified, and separated models) to evaluate the generated speech. The baseline system simply synthesizes the speech of reference texts with the DNN-based acoustic model without any pitch control. We compared the baseline with the proposed systems to verify the effectiveness of the pitch control in spoken dialogues. The unified model had three sub-conditions: L, P, and L+P, where L used only the linguistic feature of the previous utterance, P used only the prosodic feature (log F0), and L+P used both features. The separated model also had the three sub-conditions, but it had only the L condition for language generation because the language-generation network in the separated model only uses the linguistic feature. We prepared another condition, "w/o Encoder," for the separated model. In this condition, we did not use any feature from the previous utterance but generated the differential F0 context only from the generated sentence.

## 5. Experimental Results

### 5.1. Objective evaluation

We used perplexity (PPL), BLEU-4, and Dist-2 for the evaluation of language generation. We used the language generation results by teacher forcing to calculate PPL, MSE, and RMSE. In particular, the generated sentence has to match the reference sentence for calculating MSE and RMSE. For other metrics, we used the generation results by beam-search with a beam width of 20. We repeated the processes of fine-tuning and evaluation three times and then compared their average.

For the evaluation of prosody generation, we used the mean-squared error (MSE) of the differential log F0 and the root mean squared error (RMSE) of the log F0 between the synthesized speech and the reference speech. In this evaluation, we gave the reference sentence to the network and controlled only the pitch. Table 1 shows the objective evaluation results when changing the combination of features used as the input of the encoder. As shown in the table, the PPLs of the networks were

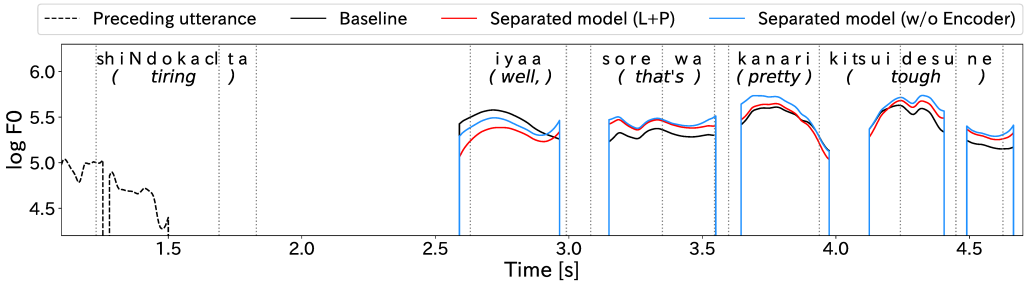


Figure 2: *F0 contour of synthesized speech. The figure shows the system utterance "Well, that's pretty tough." added to the end of the preceding utterance "It was so tiring." English words in parentheses are translations from Japanese.*

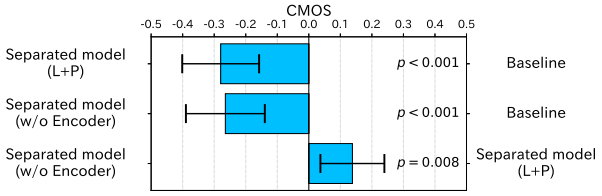


Figure 3: *Subjective evaluation results. A positive score indicates that the right-side system is more appropriate, while a negative score indicates the opposite result.*

relatively high. This result reflects the insufficient training of the language generation model. Therefore, we plan to substitute the language generation model with the Transformer-based model [6, 36] in future studies.

From the comparison of the MSE, the separated model surpassed the baseline. The separated model seemed to have been sufficiently trained as a result of focusing on the individual task, in contrast to the unified model. In addition, the BLSTM-based decoder in the prosodic generation network seemed to contribute to improving the performance of the separated model. In terms of the RMSE, the separated model without the encoder showed the best performance. This result suggests that the separated model can generate an F0 sequence close to that of spontaneous speech.

### 5.2. Subjective evaluation

Next, we conducted a subjective evaluation of the perceived appropriateness in the dialogue context of the speech by paired comparison. We compared the systems that had higher objective evaluation results: the separated model taking the bi-modal features as input ("Separated model (L+P)") and the separated model without the encoder ("Separated model (w/o Encoder)"). In addition, we used the baseline system for comparison. We used 20 utterance pairs randomly selected from the test set. Each sample contained preceding and succeeding utterances. We substituted the succeeding utterance for synthesized speech. The participants listened to the speech synthesized by two systems in random order and then evaluated which one was most appropriate in the dialogue context on a scale from 2 (former is appropriate) to $-2$ (latter is appropriate). Seventeen evaluators participated in the experiment.

Figure 3 shows the subjective evaluation results. The figure shows the average Comparison Mean Opinion Score (CMOS), and the error bars are at the 95% confidence interval. A positive score indicates that the right-side system is more appropriate than the left-side one, and a negative score indicates the opposite. Significant differences were found between all systems by a one-sample $t$-test. Comparison with the baseline showed that the differential F0 context is effective for generating synthesized speech that is appropriate in the dialogue context. In addition, the "Separated model (L+P)" was more preferred over the "Separated model (w/o Encoder)." This result indicates that the preceding utterance helps to improve the perceived appropriateness.

Finally, the log F0 contour of the synthesized speech is shown in Figure 2. English words in parentheses are translations from Japanese. We selected a sample in which the "Separated model (L+P)" was preferred over the "Separate model (w/o Encoder)." The CMOS of the example was +0.71 in the comparison. The example shows the generated response "Well, that's pretty tough." added to the preceding utterance "It was so tiring." The figure shows that the F0 contours of the separated models were different from those of the original speech synthesis. Therefore, the pitch control itself in a word-by-word manner contributed to improving the naturalness of the speech. One possible reason that the "Separated model (L+P)" was most preferred is the effect achieved by the entrainment of the prosody. As shown in the figure, some of the samples with higher scores had F0 contours that were closer to the end of the preceding utterance at their beginning point.

## 6. Conclusion

We proposed a spoken-response generation model to control the pitch of system utterances according to the context of dialogue. Our experimental results show that the proposed method based on a basic encoder-decoder model with LSTM could generate speech with appropriate F0 contour in the dialogue context, although the performance of language generation remains an issue. In particular, our proposed model using preceding utterances improved the perceived appropriateness in the dialogue of the synthesized speech.

In future studies, we plan to introduce a recent speech synthesis model for multilevel control of prosody attributes [37] and a state-of-the-art language generation model (e.g., [6]) to improve the performance of the spoken-response generation. In addition, we plan to expand the method to deal with other parameters, such as speech rate, intensity, and facial expressions.

## 7. Acknowledgments

A part of this work was supported by JST COI Grant Number JPMJCE1303 and the JSPS Grant-in-Aid for Scientific Research JP17H00823 and JP20K19903.

# 8. References


[1] A. Ram, R. Prasad, C. Khatri *et al.*, "Conversational AI: The science behind the Alexa Prize," *arXiv preprint arXiv:1801.03604*, pp. 1–18, 2018.

[2] S. Lee and J. Choi, "Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity," *International Journal of Human-Computer Studies*, vol. 103, pp. 95–105, 2017.

[3] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Proc. AAAI*, vol. 32, no. 1, 2018, pp. 5610–5617.

[4] S. Yavuz, A. Rastogi, G.-L. Chao, and D. Hakkani-Tur, "Deepcopy: Grounded response generation with hierarchical pointer networks," in *Proc. SIGDIAL*, 2019, pp. 122–132.

[5] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu, and Q. V. Le, "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.

[6] E. M. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau, "Can you put it all together: Evaluating conversational agents' ability to blend skills," in *Proc. ACL*, 2020, pp. 2021–2030.

[7] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proc. INTERSPEECH*, 2011, pp. 3081–3084.

[8] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability." *Journal of Personality and Social Psychology*, vol. 32, no. 5, pp. 790–804, 1975.

[9] A. Gravano, Š. Beňuš, R. Levitan, and J. Hirschberg, "Three ToBI-based measures of prosodic entrainment and their correlations with speaker engagement," in *Proc. SLT*, 2014, pp. 578–583.

[10] H. Giles and P. Powesland, "Accommodation theory," in *Sociolinguistics*, 1997, pp. 232–239.

[11] R. Street, "Speech convergence and speech evaluation in fact-finding interviews," *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[12] E. F. Prince, "Towards a taxonomy of given-new information," *Radical Pragmatics*, 1981.

[13] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1-2, pp. 31–43, 2002.

[14] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[15] A. 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.

[16] S. Hanabusa, T. Nose, and A. Ito, "Segmental pitch control using speech input based on differential contexts and features for customizable neural speech synthesis," in *Proc. IIH-MSP*, 2018, pp. 124–131.

[17] S. Yamada, T. Nose, and A. Ito, "A study on tailor-made speech synthesis based on deep neural networks," in *Proc. IIH-MSP*, 2017, pp. 159–166.

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

[19] N. Lubold, H. Pon-Barry, and E. Walker, "Naturalness and rapport in a pitch adaptive learning companion," in *Proc. ASRU*, 2015, pp. 103–110.

[20] T. Kase, T. Nose, and A. Ito, "On appropriateness and estimation of the emotion of synthesized response speech in a spoken dialogue system," in *Proc. HCI International*, 2015, pp. 747–752.

[21] R. Levitan, S. Benus, R. H. Gálvez, A. Gravano, F. Savoretti, M. Trnka, A. Weise, and J. Hirschberg, "Implementing acoustic-prosodic entrainment in a conversational avatar." in *Proc. INTERSPEECH*, vol. 16, 2016, pp. 1166–1170.

[22] N. Sadouohi, A. Pereira, R. Jain, L. Leite, and J. F. Lehman, "Creating prosodic synchrony for a robot co-player in a speech-controlled game for children," in *Proc. HRI*, 2017, pp. 91–99.

[23] Y. Chiba, T. Nose, M. Yamanaka, T. Kase, and A. Ito, "An analysis of the effect of emotional speech synthesis on non-task-oriented dialogue system," in *Proc. SIGDIAL*, 2018, pp. 371–375.

[24] R. Hoegen, D. Aneja, D. McDuff, and M. Czerwinski, "An end-to-end conversational style matching agent," in *Proc. IVA*, 2019, pp. 111–118.

[25] S. Fuscone, B. Favre, and L. Prévot, "Neural representations of dialogical history for improving upcoming turn acoustic parameters prediction," in *Proc. INTERSPEECH*, 2020, pp. 4203–4207.

[26] T. Koriyama, T. Nose, and T. Kobayashi, "Conversational spontaneous speech synthesis using average voice model," in *Proc. INTERSPEECH*, 2010, pp. 853–856.

[27] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data," in *Proc. INTERSPEECH*, 2019, pp. 4435–4439.

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

[29] N. Hojo, Y. Ijima, H. Sugiyama, N. Miyazaki, T. Kawanishi, and K. Kashino, "DNN-based speech synthesis considering dialogue-act information and its evaluation with respect to illocutionary act naturalness," in *Proc. Speech Prosody*, 2020, pp. 975–979.

[30] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end tts for voice agent," *arXiv preprint arXiv:2005.10438*, 2020.

[31] Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, "Construction and analysis of a multimodal chat-talk corpus for dialog systems considering interpersonal closeness," in *Proc. LREC*, 2020, pp. 443–448.

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

[33] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transaction on Information and Systems*, vol. E99-D, no. 7, pp. 1877–1884, 2016.

[34] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer, "Scheduled sampling for sequence prediction with recurrent neural networks," in *Proc. NIPS*, 2015, pp. 1171–1179.

[35] 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.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[37] G. Sun, Y. Zhang, R. J. Weiss, Y. Cao, H. Zen, and Y. Wu, "Fully-hierarchical fine-grained prosody modeling for interpretable speech synthesis," in *Proc. ICASSP*, 2020, pp. 6264–6268.