Acoustic Modeling for End-to-End Empathetic Dialogue Speech Synthesis Using Linguistic and Prosodic Contexts of Dialogue History

Yuto Nishimura[1], Yuki Saito[1], Shinnosuke Takamichi[1], Kentaro Tachibana[2], and Hiroshi Saruwatari[1]

[1]The University of Tokyo, Japan, [2]LINE Corp., Japan.

yutonishimurav20512@gmail.com, yuuki_saito@ipc.i.u-tokyo.ac.jp

Abstract

We propose an end-to-end empathetic dialogue speech synthesis (DSS) model that considers both the linguistic and prosodic contexts of dialogue history. Empathy is the active attempt by humans to get inside the interlocutor in dialogue, and empathetic DSS is a technology to implement this act in spoken dialogue systems. Our model is conditioned by the history of linguistic and prosody features for predicting appropriate dialogue context. As such, it can be regarded as an extension of the conventional linguistic-feature-based dialogue history modeling. To train the empathetic DSS model effectively, we investigate 1) a self-supervised learning model pretrained with large speech corpora, 2) a style-guided training using a prosody embedding of the current utterance to be predicted by the dialogue context embedding, 3) a cross-modal attention to combine text and speech modalities, and 4) a sentence-wise embedding to achieve fine-grained prosody modeling rather than utterance-wise modeling. The evaluation results demonstrate that 1) simply considering prosodic contexts of the dialogue history does not improve the quality of speech in empathetic DSS and 2) introducing style-guided training and sentence-wise embedding modeling achieves higher speech quality than that by the conventional method.

Index Terms: speech synthesis, spoken dialogue, dialogue speech synthesis, empathy, contexts

1. Introduction

The role of communication is to understand each other and share thoughts [1]. *Empathy*, which is the focus of this paper, is a key component of attentive-listening communication. Empathy is the act of feeling another person's thoughts as if they were the listener's own [1], and can be regarded as an active attempt to get inside the other person's head [2]. Unlike *sympathy*, where one synchronizes self with the other in emotion, empathy requires making a response by deeply understanding the other's emotion. Recent studies in dialogue have proposed natural language understanding [3] and language response generation [4] methods to implement the empathetic behavior in chat-oriented dialogue [5]. At the same time, emotion and prosody of the response speech are also vital in empathy [6]. The promising method to involve these factors in synthetic speech is called *empathetic dialogue speech synthesis (DSS)* [7]. To achieve better empathetic DSS technologies, we need to develop a method that considers the dialogue history with the user (i.e., interlocutor of the dialogue) and generates suitable speech features for the response.

Humans can produce empathetic speech by understanding *contexts* from the linguistic and prosodic features in the dialogue. One possible way to achieve context-aware empathetic DSS technologies is to condition a speech synthesis model by the dialogue history. Guo et al. [8] proposed a DSS method
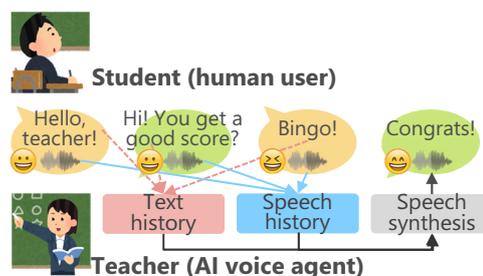

Figure 1: *Concept of our work. Speech waveform of current dialogue turn is synthesized considering text and speech history.*

incorporating a pre-trained language model to obtain *linguistic* features of the dialogue history. Their speech synthesis model is conditioned by phonemes (as with the basic speech synthesis) and a text-chat history's embedding vector extracted from the linguistic features by using a conversational context encoder (CCE). This method can train a DSS model that synthesizes natural speech considering the dialogue context in an end-to-end manner without explicitly using any annotations about the context, such as emotion labels or dialogue situation description. However, it cannot consider the *prosodic* features appearing in the speech history, i.e., the variation of speaking styles in empathetic speech, which is problematic because these can be a crucial means for empathetic communication [6]. Although several studies have attempted to produce natural response speech by using the interlocutor's previous spoken utterance (e.g., Yamazaki et al.'s F_0 control method for DSS [9]), they are unable to learn the long-term dependencies of spoken dialogue necessary for empathetic prosodic control or to benefit from the powerful recognition-synthesis capabilities of deep learning models.

We propose an end-to-end empathetic DSS model that considers both the linguistic and prosody features, as shown in Fig. 1. Our model consists of four sub-models: 1) a BERT-based sentence encoder [10] that extracts text embeddings from a text sequence of dialogue lines, 2) a prosody encoder that extracts prosody embeddings from a mel-spectrogram sequence of the spoken dialogue, 3) a cross-modal (CM) CCE that aggregates the two embeddings and predicts the conversational context from the dialogue history, and 4) a speech synthesis model based on FastSpeech 2 [11] conditioned by phoneme embeddings and the context. Therefore, our model can be regarded as an extension of conventional ones [8, 9] that use only linguistic context information or limited prosody representation in DSS. We also investigate four methods to effectively train our model, which introduce 1) a self-supervised learning (SSL) model pretrained with large speech corpora, 2) a style-guided training using a prosody embedding of a current utterance to be predicted by the dialogue context embedding, 3) a cross-modal attention to capture the text-speech modality, and 4) a sentence-wise embedding to achieve fine-grained prosody modeling rather than

utterance-wise modeling. We conducted experiments to determine our model's effectiveness using a corpus of empathetic dialogue speech synthesis [7]. The results demonstrate that 1) simply considering prosodic contexts of the dialogue history does not improve the quality of speech by empathetic DSS and 2) introducing style-guided training and sentence-wise embedding modeling achieves higher speech quality than that by the conventional method.

2. Related Work

2.1. End-to-end speech synthesis

End-to-end speech synthesis [12] is a framework to solve three sub-tasks in speech synthesis (i.e., text analysis, acoustic modeling, and waveform synthesis) using a stack of deep neural networks (DNNs). The definition of this framework depends on where we consider the end to be, e.g., linguistic features to waveform [13], text (phonemes) to acoustic features [12, 14], or text to waveform (i.e., truly end-to-end methods) [15, 16, 17]. In this study, we follow the second definition and adopt Fast-Speech 2 [11] as a baseline model from the viewpoints of fast learning speed and stable inference.

2.2. Dialogue-context-aware acoustic modeling

Context-aware acoustic modeling for end-to-end DSS has to deal with the difficulty of modeling the conversational context with long-term dependency of the dialogue (i.e., dialogue history). Guo et al. [8] proposed a DSS method introducing a CCE that predicts the contexts from a text sequence of dialogue history, without using any annotations related to the dialogue, such as emotion labels [18]. However, their method cannot consider prosodic features of the dialogue history for the context prediction, which is unfortunate because prosody is often vital for empathetic communication. Yamazaki et al. [9] proposed a spoken-response generation method that can control the F_0 values of the response speech while considering both the linguistic and prosodic contexts of the user's previous utterance. However, this method focuses only on the pitch control and ignores other prosodic contexts such as energy, which is also an essential factor to simulate the empathetic behavior in human speech communication [7]. Cong et al. [19] proposed a style-guided context learning for a DSS model, where embedding vectors of the current utterance by a dialogue agent and previous utterance by a user extracted from their mel-spectrograms become closer in the embedding space. However, this method cannot consider the complicated interaction between the linguistic and prosodic contexts of the dialogue history in empathetic spoken dialogue.

2.3. Empathetic dialogue generation

In daily communication, humans naturally control their speech prosody to empathize with each other. Empathetic dialogue generation aims to incorporate this "empathetic dialogue" behavior into a dialogue system by implementing empathy in human communication, which can improve the user experience in human-computer interactions [20]. A major challenge in building such systems is how to estimate the user's mental state (typically expressed as emotions), from the dialog history. Recently developed DNN-based natural language understanding and generation models have the ability to deeply understand the user's mental state, intentions, and emotions [21, 22]. However, context-aware empathetic DSS technology has not been achieved thus far.
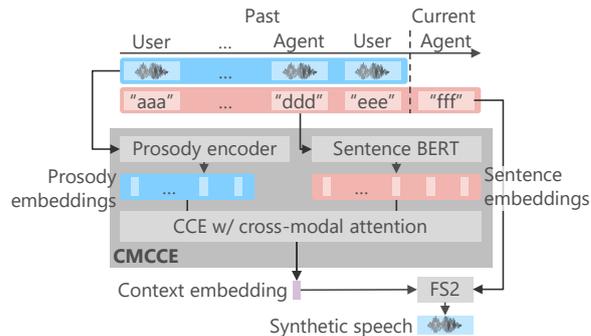

Figure 2: *Overview of proposed method. "FS2" stands for "FastSpeech2."*

3. Proposed Method

In this study, we propose a speech synthesis method conditioned on the linguistic and prosodic contexts of dialogue history between the agent and user. Fig. 2 shows an overview of the proposed method.

3.1. Baseline empathetic DSS model

Our baseline empathetic DSS model consists of a BERT-based sentence encoder [10], CCE [8], and speech synthesis model based on FastSpeech 2 [11]. Let C be a memory capacity parameter that determines the number of past utterances by the agent and user to be considered for the context prediction. In generating a dialogue response at the tth turn, the sentence BERT first takes a text sequence of the dialogue up to the past C turns, i.e., $\boldsymbol{x}_{t-C}, \boldsymbol{x}_{t-C+1}, \ldots, \boldsymbol{x}_t$, and produces their text embeddings $\boldsymbol{l}_{t-C}, \boldsymbol{l}_{t-C+1}, \ldots, \boldsymbol{l}_t$. Note that vectors with subscript $t - C < 1$ are treated as zero vectors and never contribute to the context prediction. Henceforth, we denote these sequences as $\boldsymbol{x}_{t-C:t}$ and $\boldsymbol{l}_{t-C:t}$. The CCE then extracts a conversational context embedding vector $\boldsymbol{e}_t$ from $\boldsymbol{l}_{t-C:t}$. Finally, the FastSpeech 2 model takes the text to be synthesized at turn t, $\boldsymbol{x}_t$, and the embedding vector, $\boldsymbol{e}_t$, as input and generates a mel-spectrogram of response speech $\hat{\boldsymbol{y}}_t$. The loss function for training this DSS model is defined as the speech parameter prediction error (e.g., L1 loss) between $\hat{\boldsymbol{y}}_t$ and $\boldsymbol{y}_t$, i.e., the mel-spectrogram extracted from the agent's speech at the tth turn.

3.2. Cross-modal CCE (CMCCE)

We introduce a CMCCE, an extension of the conventional CCE, to our empathetic DSS model so that it can consider both the linguistic and prosodic contexts of long-term dialogue history. Figure 2 shows the architecture of our CMCCE. In the conversational context prediction using CMCCE, a DNN-based prosody encoder first takes a mel-spectrogram sequence of the dialogue up to the past C turns excluding the current turn t, i.e., $\boldsymbol{y}_{t-C:t-1}$, and produces their prosody embeddings $\boldsymbol{p}_{t-C:t-1}$. Then, a sentence BERT extracts the text embeddings $\boldsymbol{l}_{t-C:t}$ from a dialogue text sequence $\boldsymbol{x}_{t-C:t}$ in the same manner as described in Section 3.1. Finally, our CMCCE aggregates the two embeddings $\boldsymbol{p}_{t-C:t-1}$ and $\boldsymbol{l}_{t-C:t}$ to predict the context embedding $\boldsymbol{e}_t$.

In the following subsections, we present four strategies to make our CMCCE capture the linguistic and prosodic contexts of dialogue history effectively.

3.2.1. SSL model as prosody encoder

One can train a random-initialized DNN (e.g., Du et al.'s work [23]) from scratch. Another approach is to use an SSL
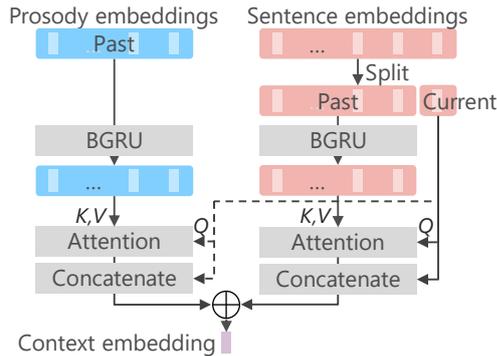
Figure 3: *Attention-based cross-modal aggregation architecture. Blue and red boxes represent prosody and text embeddings, respectively. $Q, K,$ and V are query, key, and value of attention, respectively.*

model pretrained with large speech corpora as the prosody encoder and fix it during the training, similar to the sentence BERT used for the text-related embedding extraction.

3.2.2. Style-guided context embedding learning

The context embedding e_t should correspond to the appropriate speaking style for the text at the tth turn to be synthesized. To explicitly consider this requirement, we introduce the style-guided context embedding learning proposed by Cong et al. [19] into empathetic DSS. Specifically, we train our empathetic DSS model to minimize not only the speech parameter prediction error (described in Section 3.1) but also the embedding matching loss defined as the mean squared error between the context embedding e_t and prosody embedding p_t. Specifically, we train the DSS model so that the embedding obtained from the past dialogue history matches the embedding obtained from the ground-truth speech at the tth turn. The difference between our method and Cong et al.'s is the dialogue history information to be used for the context prediction. In other words, the former uses the linguistic and prosodic context with long-term history while the latter only considers the prosody embedding of a user's utterance as the context.

3.2.3. Attention-based cross-modal aggregation

One simple approach to aggregate the embeddings from two different modalities, i.e., text and speech, is to compress each embedding sequence into a fixed-dimensional embedding independently, and take the element-wise sum of the two outcomes. However, this approach does not necessarily model the cross-modality of text and speech in empathetic dialogue history. Therefore, we investigate another approach to aggregate the cross-modal embeddings by using attention networks, as shown in Fig. 3. In this approach, the text embedding sequence $l_{t-C:t}$ is first split into the current and past embeddings, l_t and $l_{t-C:t-1}$. Then, the past text/prosody embedding sequences $l_{t-C:t-1}$ and $p_{t-C:t-1}$ are fed into a bi-directional gated recurrent unit (BGRU) independently to capture their time dependency. The outcomes of this operation are regarded as the key and value of the following attention networks, and the current text embedding l_t is treated as a query of the attention. These attention networks can be expected to learn which point in the past information should be used for predicting the conversational context while considering both text and speech modalities.

Table 1: *Number of dialogues in training/validation/evaluation sets for experimental evaluation. Numbers in parentheses mean the number of utterances by the teacher.*

Set	Long		Short	
Training	126	(1,743)	600	(1,200)
Validation	12	(91)	60	(120)
Evaluation	12	(91)	60	(120)

3.2.4. Fine-grained context embedding modeling

Guo et al.'s conventional CCE predicts an *utterance-wise* context embedding, i.e., it assumes that speech utterance in each turn has only one speaking style. However, empathetic speech is a bit more complicated, as an agent typically considers a user's mental state and then responds to the user with one or more speaking styles in the same turn. Therefore, we instead use *sentence-wise* context embeddings for modeling fine-grained change of the speaking style in empathetic dialogue within a single utterance. Note that this fine-grained modeling is only applied to generation for the current sentence, and utterance-wise embeddings of past information about the dialogue are stored in the dialogue history. The reason is to prevent the dialogue history from being filled only with the sentence-wise embeddings of one speaker (i.e., agent or user).

4. Experimental Evaluation

4.1. Experimental conditions

We used the STUDIES corpus [7], which contains simulated empathetic chat-dialogue uttered by three actors playing the roles of a female teacher and her male/female students at a tutoring school, for the experimental evaluation. We built the empathetic DSS model of the teacher in the STUDIES corpus. We split the teacher's speech data included in the long dialogue (10–20 turns) and short dialogue (four turns) subsets for training, validation, and evaluation, as shown in Table 1. We downsampled the speech data to 22,050 Hz.

We used FastSpeech 2 [11] with the PyTorch implementation for Japanese speech synthesis [24]. We followed the default settings of the DNN architecture and speech parameter extraction of this implementation. We used the WORLD vocoder [25, 26] to estimate F_0. We modified the variance adaptors to predict the speech features averaged over phonemes, referring to FastPitch [27] for stable training and higher synthetic speech quality. The optimization algorithm was Adam [28] with an initial learning rate η of 0.0625, β_1 of 0.9, and β_2 of 0.98. We first pretrained FastSpeech 2 using the JSUT corpus [29] for 500 epochs and then fine-tuned it using the STUDIES corpus for 500 epochs.

We used the HiFi-GAN vocoder [30] for speech waveform generation from 80-dimensional mel-spectrograms. We first pretrained HiFi-GAN using the JSUT corpus [29] for 100 epochs. Then, we fine-tuned it by using the same training data as that for FastSpeech 2 (shown in Table 1) for 200 epochs. The optimization algorithm was Adam with an initial learning rate η of 0.0003, β_1 of 0.8, and β_2 of 0.99.

We used the CCE [8] with the open-source PyTorch implementation [31]. However, since the open-sourced implementation of context aggregation was different from that in the original paper, we changed the structure to be the same as in the original paper. In addition, although Guo et al.'s method also used an auxiliary encoder to consider hand-crafted linguistic features, we excluded it in this study so as to focus only on the context embedding prediction from the dialogue history in-

formation. We used a sentence BERT pretrained using Japanese text data[1] for the text embedding extraction in both the baseline and proposed models. For the pretrained prosody encoder, we used the wav2vec 2.0 SSL model[2] [32] pretrained with Japanese speech data in Common Voice [33], CSS10 [34], and JSUT [29] corpora. In our DSS method, the prosody embeddings were extracted from the agent's ground-truth speech and synthetic speech for the training and inference phases, respectively.

4.2. Subjective evaluations

We conducted preference AB tests on the naturalness of synthetic speech by the "**Baseline**" model (FastSpeech 2 conditioned by text-embedding-derived context embedding [8]) and our empathetic DSS model with the four additional strategies (described in Section 3.2). Similarly, we conducted preference XAB tests on the speaking-style similarity of synthetic speech to the teacher's corresponding speech sample analysis-synthesized by the HiFi-GAN vocoder. In the XAB test, we first asked participants to listen to sample "X" (analysis-synthesized speech) and then to choose which speaking style (way of expressing emotion) was closer to "X" among the "A" and "B" samples that were played afterward. Twenty-five listeners participated in each of the evaluations using our crowdsourcing evaluation system. Each listener evaluated ten samples. In this evaluation, we first verified the effectiveness of the four strategies (Section 3.2) to be considered individually. Then, for those that showed positive effect on the quality of synthetic speech, we further evaluated the interaction of multiple strategies.

Table 2 shows the evaluation results, where "**SG**," "**Attn**," and "**FG**" stand for style-guided embedding learning, attention for cross-modal aggregation, and fine-grained context modeling for our empathetic DSS model, respectively. For the first investigation results, we found that just considering the speech modality in empathetic DSS did not necessarily improve the quality of synthetic speech. At the same time, introducing one of our strategies, i.e., "SG" for the proposed model without the SSL model and "Attn" for that with the SSL model, could outperform the baseline model regarding the speech quality. Furthermore, the proposed model with multiple strategies, i.e., without the SSL model and with the "SG" and "FG" strategies, achieved significantly higher scores than the baseline in both naturalness and speaking-style similarity. These results suggest that it may be possible to estimate the dialogue context that can achieve higher quality empathetic DSS by considering not only textual features but also prosodic ones of the dialogue history. It should be noted, however, that inappropriate strategies (e.g., "FG" for the proposed model without the SSL model) can degrade the quality of the synthesized speech over the baseline. One reason may be the difficulty of learning complicated contexts with insufficient modeling performance.

Finally, we conducted a five-scale mean opinion score (MOS) test on the utterance-level naturalness of synthetic speech to evaluate the absolute performance of the baseline and proposed models. In this test, we compared the baseline model with the best proposed model (i.e., without SSL model and with "SG" and "FG" strategies) and its variants for an ablation study. We presented 35 synthetic speech samples to listeners in random order. Listeners rated the naturalness of each sample from degrees of 1 ("very unnatural") to 5 ("very natural").

---

[1] https://huggingface.co/colorfulscoop/sbert-base-ja

[2] https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-japanese

Table 2: *Results of preference AB and XAB tests.* **BOLD** *values have significant difference with* $p < 0.05$ *(Baseline vs Proposed method with some of techniques described in Section 3.2)*

(a) Results of proposed model without SSL model

Baseline	Naturalness	Similarity	Proposed (w/o SSL)		
			SG	Attn	FG
	0.45 vs. 0.55	**0.54 vs. 0.46**			
	0.44 vs. 0.56	0.53 vs. 0.47	✓		
	0.50 vs. 0.50	0.54 vs. 0.46		✓	
	0.48 vs. 0.52	**0.54 vs. 0.46**			✓
	0.43 vs. 0.57	0.47 vs. 0.53	✓	✓	
	0.45 vs. 0.55	**0.45 vs. 0.55**	✓		✓

(b) Results of proposed model with SSL model

Baseline	Naturalness	Similarity	Proposed (w/ SSL)		
			SG	Attn	FG
	0.50 vs. 0.50	**0.61 vs. 0.39**			
	0.53 vs. 0.47	0.46 vs. 0.54	✓		
	0.51 vs. 0.49	**0.44 vs. 0.56**		✓	
	0.52 vs. 0.48	0.50 vs. 0.50			✓
	0.43 vs. 0.57	0.50 vs. 0.50	✓	✓	
	0.46 vs. 0.54	**0.54 vs. 0.46**		✓	✓

Table 3: *Results of MOS tests on speech naturalness with 95% confidence intervals*

Method			Naturalness MOS
Proposed (w/o SSL)			
SG	Attn	FG	
			3.59±0.10
✓			3.62±0.10
		✓	3.59±0.10
✓		✓	3.66±0.10
✓	✓	✓	3.55±0.10
Baseline			3.55±0.10

Table 3 shows the evaluation results. Among the seven compared methods, our model introducing "SG" and "FG" without the SSL model achieved the highest MOS, thus demonstrating the effectiveness of our model in improving the quality of synthetic speech. Interestingly, the proposed method that introduced all strategies excluding "SSL model" performed similarly to the baseline. This result suggests that in empathetic DSS, using a richer model does not necessarily lead to an improvement in speech quality. The speech samples used in the subjective evaluations are available at http://sython.org/Corpus/STUDIES/demo_empDSS.html.

5. Conclusion

In this paper, we proposed an end-to-end empathetic DSS model that considers both linguistic and prosody features, and investigated the effectiveness of four methods to train our model. The results demonstrated that 1) simply considering the prosodic contexts of the dialogue history did not improve the quality of speech by empathetic DSS and 2) introducing style-guided training and sentence-wise embedding modeling achieved higher speech quality than that by the conventional method. In future work, we will investigate a method for building our empathetic DSS model in a supervised or semi-supervised setting, i.e., emotion labels of agent or user utterances are fully or partially available.

6. References

[1] N. Komiya, *Professional Counselor's Lesson on Listening Skills for Each Situation*. Natsumesha CO., LTD, 2015 (in Japanese).

[2] M. H. Davis, *Empathy: A Social Psychological Approach*. Routledge, 2018.

[3] Q. Lui, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, "EmpDG: Multi-resolution interactive empathetic dialogue generation," in *Proc. COLING*, Barcelona, Spain, Dec. 2020, pp. 4454–4466.

[4] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. ACL*, Florence, Italy, Jul. 2019, pp. 5370–5381.

[5] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *ACM SIGKDD Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.

[6] C. Regenbogen, D. A. Schneider, A. Finkelmeyer, N. Kohn, B. Derntl, T. Kellermann, R. E. Gur, F. Schneider, and U. Habel, "The differential contribution of facial expressions, prosody, and speech content to empathy," *Cognition & Emotion*, vol. 26, no. 6, pp. 995–1014, 2012.

[7] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "STUDIES: Corpus of japanese empathetic dialogue speech towards user-friendly voice agent," in *Proc. INTERSPEECH (Accepted)*, Incheon, South Korea, Sep. 2022.

[8] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end TTS for voice agents," in *Proc. SLT*, Shenzhen, China, Jan. 2021, pp. 403–409.

[9] Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, "Neural spoken-response generation using prosodic and linguistic context for conversational systems," in *Proc. INTERSPEECH*, Brno, Czech Republic, Sep. 2021, pp. 246–250.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, U.S.A., Jun. 2019, pp. 4171–4186.

[11] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *Proc. ICLR*, Vienna, Austria, May 2021.

[12] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R.-J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R.-A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 4006–4010.

[13] A. v. d. 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*, vol. abs/1609.03499, 2016.

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

[15] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proc. ICLR Workshop*, Toulon, France, May 2017.

[16] W. Ping, K. Peng, and J. Chen, "ClariNet: Parallel wave generation in end-to-end text-to-speech," in *Proc. ICLR*, New Orleans, U.S.A., May 2019.

[17] R. J. Weiss, R. Skerry-Ryan, E. Battenberg, S. Mariooryad, and D. P. Kingma, "Wave-Tacotron: Spectrogram-free end-to-end text-to-speech synthesis," in *Proc. ICASSP*, Montreal, Canada, Jun. 2021, pp. 5679–5683.

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

[19] J. Cong, S. Yang, N. Hu, G. Li, L. Xie, and D. Su, "Controllable context-aware conversational speech synthesis," in *Proc. INTERSPEECH*, Brno, Czech Republic, Sep. 2021, pp. 4658–4662.

[20] K. K. Liu and R. W. Picard, "Embedded empathy in continuous interactive health assessment," in *Proc. CHI Workshop on HCI Challenges in Health Assessment*, Portland, U.S.A., Apr. 2005.

[21] Y. Liu, W. Maier, W. Minker, and S. Ultes, "Empathetic dialogue generation with pre-trained RoBERTa-GPT2 and external knowledge," in *Proc. IWSDS*, Singapore, Nov. 2021.

[22] Y. Xie and P. Pu, "Empathetic dialog generation with fine-grained intents," in *Proc. CoNLL*, Punta Cana, Dominican Republic, Nov. 2021, pp. 133–147.

[23] C. Du and K. Yu, "Phone-level prosody modelling with GMM-based MDN for diverse and controllable speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 190–201, Jan. 2022.

[24] Wataru-Nakata, "FastSpeech2-JSUT," https://github.com/Wataru-Nakata/FastSpeech2-JSUT, 2021.

[25] 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.

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

[27] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*, Montreal, Canada, Jun. 2021, pp. 6588–6592.

[28] D. Kingma and B. Jimmy, "Adam: A method for stochastic optimization," in *arXiv preprint arXiv:1412.6980*, 2014.

[29] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, Sep. 2020.

[30] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2020.

[31] keonlee9420, "Expressive-FastSpeech2," https://github.com/keonlee9420/Expressive-FastSpeech2/tree/conversational, 2021.

[32] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2020.

[33] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common Voice: A massively-multilingual speech corpus," in *Proc. LREC*, Marseille, France, May 2020, pp. 4218–4222.

[34] K. Park and T. Mulc, "CSS10: A collection of single speaker speech datasets for 10 languages," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1566–1570.