CALLS: Japanese Empathetic Dialogue Speech Corpus of Complaint Handling and Attentive Listening in Customer Center

Yuki Saito[1], Eiji Iimori[1], Shinnosuke Takamichi[1], Kentaro Tachibana[2], and Hiroshi Saruwatari[1]

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

yuuki_saito@ipc.i.u-tokyo.ac.jp

Abstract

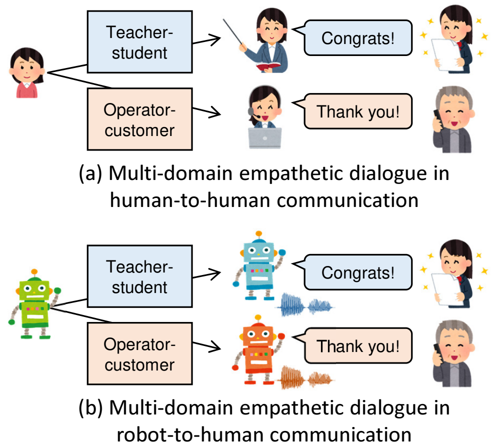
We present *CALLS*, a Japanese speech corpus that considers phone calls in a customer center as a new domain of empathetic spoken dialogue. The existing STUDIES corpus covers only empathetic dialogue between a teacher and student in a school. To extend the application range of empathetic dialogue speech synthesis (EDSS), we designed our corpus to include the same female speaker as the STUDIES teacher, acting as an operator in simulated phone calls. We describe a corpus construction methodology and analyze the recorded speech. We also conduct EDSS experiments using the CALLS and STUDIES corpora to investigate the effect of domain differences. The results show that mixing the two corpora during training causes biased improvements in the quality of synthetic speech due to the different degrees of expressiveness. Our project page of the corpus is http://sython.org/Corpus/STUDIES-2.

Index Terms: speech corpus, empathetic spoken dialogue, empathetic dialogue speech synthesis, empathy, domain difference

1. Introduction

Empathetic dialogue speech synthesis (EDSS) [1] is an emerging technology in the text-to-speech (TTS) [2] research field. Its goal is to develop a friendly voice agent that can empathetically talk to humans with a speaking style suitable for the dialogue situation. Saito et al. [1] first constructed a Japanese speech corpus of empathetic dialogues named *STUDIES*. This corpus considers chit-chat between a teacher and student in a school [3] as the empathetic dialogue situation and includes voices of a female actor who performs the teacher. Their baseline end-to-end EDSS model well reproduced the teacher's empathetic speaking styles by using the speaker's emotion label and chat history as the conditional features. Since EDSS is essential for the next-generation society where humans and robots will collaborate, developing speech corpora for various empathetic dialogue situations is vital for TTS research and its related fields, such as natural language processing and spoken dialogue systems.

Although humans can speak empathetically with their interlocutors in various dialogue domains, as shown in Fig. 1(a), the STUDIES corpus assumes only a single dialogue domain. Concretely, the teacher attempts to motivate her students to increase their enjoyment of school life through her *informal* (i.e., basically without honorific words) and intensely expressive speaking styles. However, such styles are only sometimes suitable for other dialogue domains where an empathetic listener talks more *formally*, such as a doctor counseling a patient [4]. Therefore, developing speech corpora containing a single speaker's empathetic dialogues in various domains will contribute to multi-domain EDSS that can appropriately control empathetic speaking styles in accordance with domains (Fig. 1(b)).



(a) Multi-domain empathetic dialogue in human-to-human communication

(b) Multi-domain empathetic dialogue in robot-to-human communication

Figure 1: *Multi-domain empathetic spoken dialogues in (a) human-to-human and (b) robot-to-human communications.*

To facilitate multi-domain EDSS research, we present *CALLS* (**C**omplaint handling and **A**ttentive **L**istening **L**ines **S**peech), a new empathetic dialogue speech corpus for developing EDSS technologies suitable for polite and formal dialogue domains. It includes the same female speaker as the STUDIES teacher, acting as a customer-center operator in simulated phone calls. This design enables a single speaker's different empathetic speaking styles to be analyzed in multiple dialogue domains. This paper presents 1) a methodology to construct our corpus, 2) the analysis results of the recorded speech, and 3) the results of EDSS experiments using the CALLS and STUDIES corpora to investigate the effect of domain differences. The contributions of this study are as follows:

- We construct a new corpus for developing a polite voice agent and building a multi-domain EDSS model. Our corpus is open-sourced for research purposes only from this page.
- We analyze the dialogue domain differences between our CALLS and existing STUDIES corpora regarding the prosodic and textual features.
- We present the results of EDSS experiments and demonstrate that mixing the two corpora during the training causes biased improvements in the quality of synthetic speech due to the different degrees of expressiveness.

2. Corpus Construction Methodology for Simulated Customer-Center Dialogues

2.1. Dialogue scenario

Empathetic behaviors by a customer-center operator can engender highly positive customer responses [5] and affect customer satisfaction [6]. However, recording such behaviors in actual operator-customer conversations suitable for TTS takes much work. The reason is that the contents may include much personal and confidential information, and the bandwidth is typically limited. Therefore, we instead considered simulated customer-center dialogues as the situation and constructed the CALLS corpus including two subsets: 1) situation-oriented complaint handling and 2) positive attentive listening.

Situation-oriented complaint handling: The first subset focused on situations where customers feel dissatisfied with a product or service provided by a specific company or organization. The operator responds to the complaining customers by phone calls. We specified the operator's persona as "a female in her early twenties, Tokyo dialect speaker, gentle tone of voice," which was consistent throughout all dialogue lines. We first picked out several dialogue situations from the FKC corpus [7] to create realistic dialogue scenarios. The FKC corpus includes approximately 5M text data describing anonymous users' complaints about a specific service or product, such as "The laundry detergent powder leaves white residue on my clothes." Some data optionally contain a user's proposal to solve the issue, e.g., "I wish the laundry detergent would dissolve more easily," and user profile information consisting of age, gender, job, and locale. We chose 10 complaint data containing both the proposal and user profile information for each of eight categories (e.g., "daily commodities," "food and drink," and "health and beauty") and created 2 (male or female customer) $\times$ 8 (categories) $\times$ 10 (complaints) $=$ 160 dialogue situations.

Positive attentive listening: The second subset considered situations where the operator talks with customers about positive content. For example, a customer feels satisfied with the quality of service, successful transaction, or the employee's behavior and tells the operator about their satisfaction. This design aims to make a voice agent acquire the attentive listening skills essential for social activities in Japanese speech communication [8]. We used the same persona for the operator as that for the first subset and did not specify the personas of customers.

Our CALLS corpus does not include the customers' initial emotions before the dialogues start. The reason is that customers' initial emotions can be roughly classified as negative (angry or sad) or positive (happy), depending on the subset. In addition, our corpus does not contain any angry utterances by the operator because they can make a customer dissatisfied and even sympathize with the operator's angry attitude.

2.2. Crowdsourcing phone-call dialogue lines

We crowdsourced lines of phone calls between the operator and customers. We collected the dialogue lines by microtask crowdsourcing [9], where any worker can participate in the task without approval from the client.

For collecting the complaint handling lines, we first prepared Google Forms containing task description (i.e., writing lines of the customer-operator dialogues), personas for the speakers, dialogue turns (4–10), and the dialogue situations. Then, we asked crowdworkers to access the forms and write the dialogue lines. In writing dialogue lines, we instructed crowdworkers to 1) use the user's proposal and make the dialogue as constructive as possible, 2) write each dialogue line following the format "[speaker's name] [content] [emotion]," and 3) anonymize the name of a particular company or product if the dialogue situation included it. Finally, we excluded outcomes from spamming workers (e.g., only writing a single character) and workers who did not follow our instructions.

Microtask crowdsourcing for collecting short (4 turns) attentive-listening lines was similar to that used in Saito et al.'s methodology [1]. The instruction to crowdworkers was: "*Create 4-turn dialogue lines where a customer and operator are talking happily on a phone call. You are free to think about how to end the dialogue, but please consider the assumption that the customer feels happy.*" We prohibited crowdworkers from including the name of a specific service or product in the dialogue lines. Finally, we proofread all the collected lines and removed ones that 1) never contained positive emotions and 2) included offensive expressions.

Table 1: *Number of utterances per each subset. "CH" and "AL" in this table denote complaint handling and attentive listening, respectively*

Speaker	CH	AL	Total (hours)
Operator	2,072	1,200	3,272 (6.5)
Customer	2,112	1,200	3,312 (N/A)

Table 2: *Number of utterances for each emotion*

Speaker	Neutral	Happy	Sad	Angry
Operator	657	1,669	946	0
Customer	1,149	934	284	945

2.3. Voice recording

We employed the same female speaker as the STUDIES teacher and asked her to perform the operator's role. She uttered the phone-call lines in a recording studio as if she responded to the customers considering their emotions. The recording was conducted on separate days using a unidirectional desktop condenser microphone. The sampling rate was 48 kHz.

Although we collected the customers' utterances, we did not record their voices. The reason came from Nishimura et al.'s study [10], which suggests that using only text-based chat history is sufficient to improve speech quality in EDSS.

3. Corpus Analysis

3.1. Crowdsourcing setting and results

We used Lancers[1] as the crowdsourcing platform for collecting dialogue lines. The overall period of the crowdsourcing ran from July to August 2022. We employed 150 and 1,400 crowdworkers (with overlap per task) to collect dialogue lines for the complaints handling and attentive listening subsets, respectively. We paid crowdworker in the first and second tasks \$4.04 and \$0.4 as compensation, respectively. As a result of our screening, we obtained 820 lines of complaint handling and 600 lines of attentive listening.

3.2. Corpus specification

Table 1 shows the number of utterances per subset of the CALLS corpus consisting of 3,272 utterances by the operator (6.5 hours) and 3,312 utterances by the customers (without their speech). Table 2 shows the number of utterances per emotion label. As we explained in Section 2.1, our corpus does not include any angry utterances by the operator. In addition, the operator and customer utterances are generally imbalanced in terms of emotion labels, which may make modeling the operator's empathetic speaking styles using the interlocutor's emotion more complicated than that for the STUDIES corpus.

3.3. Comparison with existing corpora

Table 3 lists existing Japanese dialogue speech corpora related to the CALLS corpus. Our corpus extends the application range of EDSS technologies suitable for more polite dialogue situations than STUDIES, although ours only includes a single

[1] https://www.lancers.jp/

Table 3: *List of existing Japanese dialogue speech corpora related to CALLS*

Corpus	Dialogue type	Open-sourced	Dur [hour]	# of speakers	Emotion label
Hiraoka et al. [11]	Persuasive	No	5.7	22	No
Kawahara et al. [12]	Attentive listening	No	2.3	8	No
STUDIES	Empathetic (teacher–student)	Yes	8.0	3	Yes
CALLS (ours)	Empathetic (operator–customer)	Yes	6.5	1	Yes

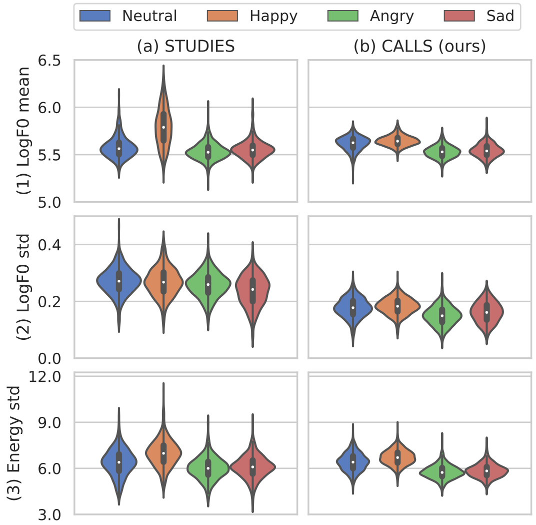


Figure 2: *Prosodic feature statistics aggregated based on the interlocutor's emotion.*

speaker's simulated spoken dialogue lines. However, mixing the CALLS operator's data with the STUDIES teacher's data (about 5 hours) yields the speaker's expressive speech corpus of over 10 hours, comparable to the JSUT corpus [13] widely used for training high-quality end-to-end TTS models in Japanese.

Our CALLS corpus closely relates to a persuasive dialogue corpus developed by Hiraoka et al. [11]. This dialogue aims to persuade an interlocutor to do something profitable, such as encouraging daily exercises and purchasing a new product. One can regard this dialogue as an example of empathetic dialogues in the task-oriented situation.

The attentive listening subset in our corpus is inspired by a counseling dialogue corpus developed by Kawahara et al. [12]. It includes simulated counseling dialogues between six subjects and professional counselors, which contain some acknowledging acts such as "うん うん (un un)" to show the listener understands and agrees with the speaker [14]. Since our CALLS corpus does not cover such behaviors that occur in spontaneous conversation, we will continue to record empathetic spoken dialogues in face-to-face situations.

3.4. Domain differences in empathetic spoken dialogues

We analyzed the differences between our CALLS corpus and existing STUDIES corpus to investigate the effects of domain difference in empathetic spoken dialogues.

Prosodic feature differences: Figure 2 shows the violin plot of the prosodic feature statistics, aggregated on the basis of the interlocutor's emotion. Rows in this figure correspond to the (1) mean and (2) standard deviation (std) of log F_0, and (3) std of energy, respectively. We used the WORLD vocoder [15, 16] for the F_0 extraction. From the results, we can see the the most significant difference is observed in the log F_0 mean of voices spoken to the happy interlocutors (Fig. 2(1)). Concretely, the log F_0 mean values for the teacher's voices widely distribute around its median (5.79), while those for the operator's voices narrowly distribute around its median (5.64). Another observation is that the log F_0 std values of the teacher's voices are generally higher than those of the operator's voices (Fig. 2(2)). A similar tendency is observed in the differences of energy std shown in Fig. 2(3). These results suggest that the expressiveness of speech in empathetic dialogues tends to become greater in casual situations and smaller in formal, polite situations.

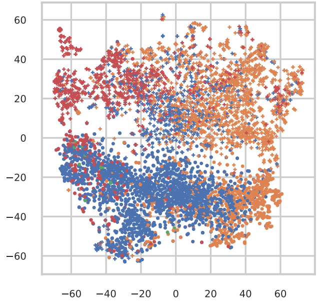


Figure 3: *t-SNE plot of sentence embeddings. Colors showing emotion labels correspond to those used in Fig. 2. Circle and cross marks denote utterances by the STUDIES teacher and CALLS operator, respectively.*

Table 4: *Number of utterances in training/validation/evaluation sets of CALLS operator and STUDIES teacher*

Set	CALLS	STUDIES
Training	2,824	2,209
Validation	224	221
Evaluation	224	211

Textual feature differences: Figure 3 shows the t-SNE plot of sentence embedding vectors extracted from the text data of STUDIES teacher and CALLS operator by using BERT [17]. We used sentence BERT pretrained with Japanese text data[2]. From this figure, we can see that the embeddings from the two different speakers are clearly isolated on the basis of the values of the y-axis, although the values of the x-axis roughly differentiate the speaker's emotion regardless of the domain difference. This result shows that texts included in our CALLS corpus cover different ranges from the existing STUDIES corpus.

4. EDSS Experiment

We report the results of two EDSS experiments: single-/multi-domain EDSS using the CALLS and STUDIES corpora.

4.1. Experimental conditions

This section describes conditions common between the single-/multi-domain EDSS experiments.

Datasets: We split speech data of the CALLS operator and STUDIES teacher for training, validation, and evaluation sets as shown in Table 4. The numbers of utterances per dialogue subset (i.e., complaint handling and attentive listening in CALLS and long-/short-dialogues in STUDIES) were roughly balanced.

[2] https://huggingface.co/koheiduck/bert-japanese-finetuned-sentiment

We downsampled the speech data to 22,050 Hz. We used the validation data to choose hyperparameters for our backbone TTS model, whose parameters were randomly initialized.

Backbone TTS model: An acoustic model for predicting a mel-spectrogram from text was FastSpeech 2 (FS2) [18], with the PyTorch implementation for Japanese TTS[3]. We followed the settings of a neural network architecture and speech parameter extraction of this implementation. The optimizer was Adam [19] with a learning rate η of 0.0625, β_1 of 0.9, and β_2 of 0.98. We first pretrained FS2 by using the JSUT corpus [13] with 200K iterations and then fine-tuned it by using the training data with 100K iterations. A neural vocoder for synthesizing speech from mel-spectrogram was HiFi-GAN [20], with the official PyTorch implementation[4]. We trained HiFi-GAN by using the training data with 350K iterations. The optimizer was Adam with η of 0.0002, β_1 of 0.8, and β_2 of 0.99.

Evaluation criteria: We conducted two kinds of five-scaled mean opinion score (MOS) tests regarding the naturalness and speaking-style similarity of synthetic speech. In the naturalness test, we presented 30 synthetic speech samples to listeners in random order. Listeners rated the naturalness of each speech sample. In the similarity test, listeners first listened to HiFi-GAN-vocoded natural speech as reference. Then, listeners scored how similar the speaking style of the presented voice was to the reference. The number of presented synthetic speech samples was 30. We conducted 8 MOS tests to evaluate the naturalness or similarity of speech samples by the CALLS operator or STUDIES teacher synthesized with single-/multi-domain EDSS model. Fifty listeners participated in each test using our crowdsourcing evaluation system, and the total number of listeners was 400.

4.2. Evaluation of single-domain EDSS

We independently trained two EDSS models using the CALLS or STUDIES corpus, which made each model focus on the given dialogue domain only. Following Saito et al.'s work [1], we investigated three conditional features: 1) the agent's emotion (**A-Emo**), 2) the user's emotion (**U-Emo**), and 3) context information learned from a conversational context encoder (**CCE**) [21]. The **CCE** extracted an embedding vector of dialogue context from joint vectors of one-hot coded speaker identity and sentence embedding sequences (one current and three previous utterances) obtained by using the same BERT as we explained in Section 3.4. The dimensionality of BERT embedding was 768, and we projected the embedding onto the 256-dimensional feature space using a fully-connected layer.

Figure 4(1) shows the evaluation results. For the two models trained only on the CALLS or STUDIES corpus, the highest naturalness MOS values were different (3.82 and 3.59, respectively). In addition, the latter was generally inferior to the former regarding the similarity MOS values. These results indicate that the CALLS operator's speaking styles are easier to reproduce than the STUDIES teacher's because the former have lower variances of prosodic features. Among the three compared factors, **CCE** improved both the naturalness and similarity, and the combination of **CCE** and **A-Emo** achieved the highest similarity MOS for the CALLS operator, which suggest that **CCE** by itself can learn expressive speaking styles and additional conditioning by emotion labels may enhance this ability.

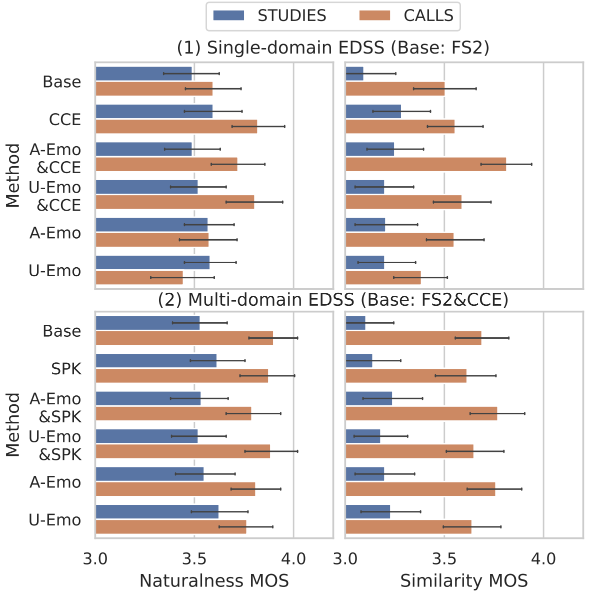


Figure 4: *MOS test results with their 95% confidence intervals*

4.3. Evaluation of multi-domain EDSS

We trained a single EDSS model using the mixture of CALLS and STUDIES corpora, which increased the number of training data but made the training more challenging due to the domain differences. Here, we set the baseline model as **FS2&CCE** in accordance with the results of the previous experiment and considered conditioning the model by a speaker embedding (**SPK**) as an additional factor to **A-Emo** and **U-Emo**. This investigation aimed to clarify whether we should explicitly separate the different empathetic speaking styles by the two speakers during the EDSS model training or not.

Figure 4(2) shows the evaluation results. From the results, learning with multiple empathetic dialogue speech corpora improved the overall naturalness and style similarity of the CALLS operator's synthetic speech. However, the improvement for the STUDIES teacher's voices was limited and the achievable style similarity was less than when trained on a single corpus. These results suggest that, even if we condition an EDSS model by speaker and emotion labels, we need to incorporate a method, such as the domain-adversarial training [22], for capturing the characteristics of different domains explicitly.

5. Conclusion

We presented CALLS, a new speech corpus to advance multi-domain empathetic dialogue speech synthesis (EDSS) research. Our corpus includes the same female speaker as the teacher in the STUDIES corpus, acting as an operator in simulated phone calls. This design enables a single speaker's different empathetic speaking styles to be analyzed in multiple dialogue domains. The results of EDSS experiments showed that mixing the two corpora during the training caused biased improvements in the quality of synthetic speech due to the different degrees of expressiveness. In future, we introduce data augmentation [23] to cope with the domain differences in multi-domain EDSS and evaluate our empathetic spoken dialogue system in face-to-face communication.

Acknowledgements: This research was conducted as joint research between LINE Corporation and Saruwatari-Takamichi Laboratory of The University of Tokyo, Japan. Part of this work was supported by JSPS KAKENHI Grant Number 21K21305.

[3] https://github.com/Wataru-Nakata/FastSpeech2-JSUT

[4] https://github.com/jik876/hifi-gan

6. References

[1] Y. Saito, Y. Nishimura, S. Takamichi, K. Tachibana, and H. Saruwatari, "STUDIES: Corpus of Japanese empathetic dialogue speech towards friendly voice agent," in *Proc. INTERSPEECH*, Incheon, South Korea, Sep. 2022, pp. 5155–5159.

[2] Y. Sagisaka, "Speech synthesis by rule using an optimal selection of non-uniform synthesis units," in *Proc. ICASSP*, New York, U.S.A., Apr. 1988, pp. 679–682.

[3] C. Warren and B. K. Hotchkins, "Teacher education and the enduring significance of "false empathy"," *The Urban Review*, vol. 47, no. 2, pp. 266–292, Jun. 2015.

[4] M. Hojat, "Ten approaches for enhancing empathy in health and human services cultures," *Journal of Health and Human Services Administration*, vol. 31, no. 4, pp. 412–450, 2009.

[5] C. M. Clark, U. M. Murfett, P. S. Rogers, and S. Ang, "Is empathy effective for customer service? Evidence from call center interactions," *Journal of Business and Technical Communication*, vol. 27, no. 2, pp. 123–153, 2013.

[6] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, Y. Aono, and T. Toda, "Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 715–728, 2020.

[7] K. Mitsuzawa, M. Tauchi, M. Domoulin, M. Nakashima, and T. Mizumoto, "FKC Corpus: a Japanese corpus from New Opinion Survey Service," in *Proc. LREC NIEUW Workshop*, Portorož, Slovenia, May 2016, pp. 11–18.

[8] H. M. Cook, "Language socialization in Japanese elementary schools: Attentive listening and reaction turns," *Journal of Pragmatics*, vol. 31, no. 11, pp. 1443–1465, Nov. 1999.

[9] E. Simperl, "How to use crowdsourcing effectively: Guidelines and examples," *LIBER Quarterly*, vol. 25, no. 1, pp. 18–39, Aug. 2015.

[10] Y. Nishimura, Y. Saito, S. Takamichi, K. Tachibana, and H. Saruwatari, "Acoustic modeling for end-to-end empathetic dialogue speech synthesis using linguistic and prosodic contexts of dialogue history," in *Proc. INTERSPEECH*, Incheon, South Korea, Sep. 2022, pp. 3373–3377.

[11] T. Hiraoka, G. Neubig, S. Sakti, T. Toda, and S. Nakamura, "Learning cooperative persuasive dialogue policies using framing," *Speech Communication*, vol. 84, pp. 83–96, Nov. 2016.

[12] T. Kawahara, M. Uesato, K. Yoshino, and K. Takanashi, "Toward adaptive generation of backchannels for attentive listening agents," in *Proc. IWSDS*, Busan, South Korea, Jan. 2015.

[13] 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.

[14] S. White, "Backchannels across cultures: A study of Americans and Japanese," *Language in Society*, vol. 18, no. 1, pp. 59–76, Mar. 1989.

[15] 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.

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

[17] 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.

[18] 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.

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

[20] 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.

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

[22] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, Apr. 2016.

[23] R. Terashima, R. Yamamoto, E. Song, Y. Shirahata, H.-W. Yoon, J.-M. Kim, and K. Tachibana, "Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation," in *Proc. INTERSPEECH*, Incheon, South Korea, Sep. 2022, pp. 3018–3022.