ChatGPT-EDSS: Empathetic Dialogue Speech Synthesis Trained from ChatGPT-derived Context Word Embeddings

Yuki Saito[1], Shinnosuke Takamichi[1], Eiji Iimori[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 propose *ChatGPT-EDSS*, an empathetic dialogue speech synthesis (EDSS) method using ChatGPT for extracting dialogue context. ChatGPT is a chatbot that can deeply understand the content and purpose of an input prompt and appropriately respond to the user's request. We focus on ChatGPT's reading comprehension and introduce it to EDSS, a task of synthesizing speech that can empathize with the interlocutor's emotion. Our method first gives chat history to ChatGPT and asks it to generate three words representing the intention, emotion, and speaking style for each line in the chat. Then, it trains an EDSS model using the embeddings of ChatGPT-derived context words as the conditioning features. The experimental results demonstrate that our method performs comparably to ones using emotion labels or neural network-derived context embeddings learned from chat histories. The collected ChatGPT-derived context information is available at our project page.

Index Terms: text-to-speech, empathetic dialogue speech synthesis, dialogue context, ChatGPT, prompt engineering

1. Introduction

Dialogue speech synthesis (DSS) [1], i.e., text-to-speech (TTS) [2] for spoken dialogue systems, is a crucial technology to actualize natural speech communication between humans and robots. In contrast to TTS, which primarily aims to convey information written in the given input text correctly, DSS requires its speaking style to be more properly controlled in accordance with the dialogue situation (e.g., restaurant reservation [3] and persuasion [4]). Such control is achieved by estimating *context*, e.g., intention [5] and speakers' emotions [6], from dialogue history and conditioning a DSS model by the context [7].

Empathetic DSS (EDSS) [8] is an emerging technology for developing a friendly voice agent that can empathize with an interlocutor. As with empathetic dialogue generation [9], an EDSS model is trained to synthesize speech with an empathetic speaking style using the dialogue context. For instance, Saito et al.'s EDSS method [8] uses the speaker's emotion label as the context and improves the quality of synthetic speech. However, this method relies on the annotations of utterance-wise emotion labels for each speaker, which requires the annotators (i.e., *human dialogue advisers*) to deeply understand the empathetic dialogue lines. Although a data-driven context embedding learning [7] can provide a way to control the expressive speaking style of synthetic speech from the chat history, the learned embedding vectors are often hard to interpretable for humans.

In text-based dialogue paradigm, ChatGPT (generative pretrained Transformer)[1], a state-of-the-art artificial intelligence (AI) chatbot, has achieved meaningful breakthroughs in various

---
[1]https://chat.openai.com/chat

AI listener
(EDSS model)

AI dialogue adviser
(ChatGPT)

Speaker: *Hello!*
Listener: *Oh, did you get a good score?*
Speaker: *Bingo! I improved my scores!*
Listener: *Congratulations!*

How should I speak this line?

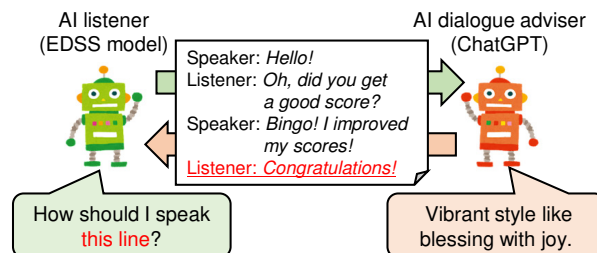Vibrant style like blessing with joy.

Figure 1: *Conceptual dialogue of our ChatGPT-EDSS.*

creative applications, such as writing novels and song lyrics. It is based on GPT-3 [10], which has been fine-tuned using supervised learning and reinforcement learning for generating response texts preferred by humans [11]. This learning mechanism enables ChatGPT to deeply understand the content and purpose of input text prompts and appropriately respond to the user's requests. Although this superior reading comprehension has the potential to extract the dialogue context from the given chat history, the applicability of ChatGPT to spoken dialogue technologies has not yet been investigated.

To this end, we propose *ChatGPT-EDSS*, a ChatGPT-powered EDSS method using ChatGPT as *an AI dialogue adviser*, as shown in Fig. 1. Our method first gives dialogue history to ChatGPT as the prompt and asks it to generate three words related to the context: intention, emotion, and speaking style for each dialogue line. Then, it trains an EDSS model using embedding vectors of the three words as conditional features. We present our methodology to collect ChatGPT-derived context words, training method for ChatGPT-EDSS, and experimental evaluation using a Japanese speech corpus of empathetic dialogue. The contributions of this study are as follows:

- We investigate a way to introduce ChatGPT to spoken dialogue research, especially in EDSS that requires deep understanding of dialogue context to properly control the speaking style of synthetic speech.
- We present the prompt design to obtain meaningful context by using ChatGPT and analyze the obtained context words.
- We demonstrate that the use of ChatGPT-derived context word embeddings can achieve naturalness and style similarity of synthetic speech comparable to that of ground-truth emotion labels or deeply-learned context embeddings [7].

2. Related Work

2.1. EDSS using explicit and implicit dialogue context

Saito et al. proposed a baseline EDSS method [8] using ground-truth emotion labels and embedding vectors of chat history as explicit and implicit dialogue context features, respectively. They introduced a conversational context encoder (CCE) [7] that extracts an embedding vector from lines of chat history
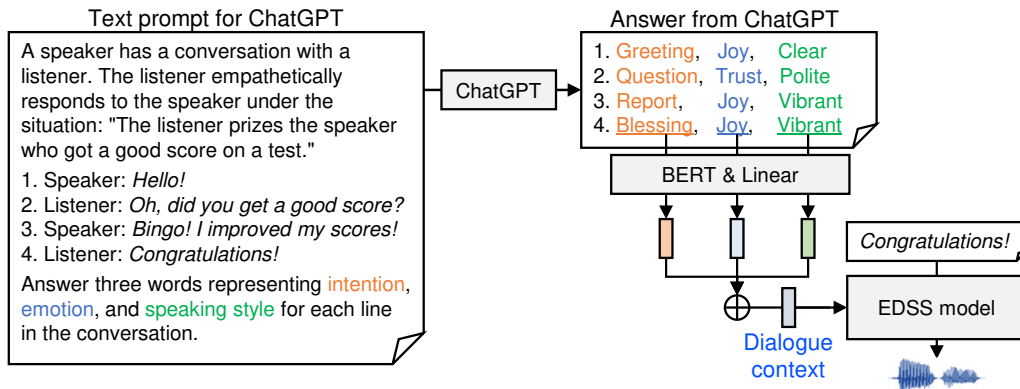
Figure 2: *Overview of our ChatGPT-EDSS.*

as the implicit context feature. Their method outperformed a FastSpeech 2 (FS2)-based TTS model [12] regarding the reproducibility of speaking style in EDSS.

2.2. Ability of ChatGPT

As of March 2023, many researchers have been exploring ChatGPT's ability in real-world situations (e.g., education, evaluation [13]) and theory of mind [14]. In addition, some have introduced ChatGPT into the assessment of human's states via texts, e.g., personalities [15] and sentiment [16]. These work motivate us to use text dialogue contexts obtained by ChatGPT for enhancing DSS technologies.

2.3. Media creation from prompt

With the advancement of deep generative modeling techniques such as denoising diffusion probabilistic models [17], media generation from a prompt has been widely studied. DaLL-E [18] is one of the first models to generate a realistic image from an input prompt. GPT-3 [10] is an autoregressive large language model that can continue to generate successive sentences from a given initial text as the prompt. AudioGen [19] and MusicLM [20] are generative models for environmental sound and music from text descriptions, respectively. These technologies offer an intuitive way to control the outcomes of media creation by changing the natural language description in the prompt.

Compared with image or text generation, research on TTS control using text prompts is still developing. One primary reason is the difficulty in constructing a sufficiently large dataset that includes many triplets of text to be spoken, speech, and natural language description to explain the speech. Guo et al. [21] dealt with this difficulty by asking proprietary experts to write prompts that describe the given speech and diversifying the prompts by using SimBERT [22]. Although their dataset contains more than 150,000 data that can be used for training a text-prompt-aware TTS model, such a method for constructing the dataset is very costly and hard to generalize.

3. ChatGPT-EDSS

As shown in Fig. 2, our ChatGPT-EDSS consists of two steps: 1) collection of dialogue context words using ChatGPT and 2) training of an EDSS model using the context words.

3.1. Collection of dialogue context words

We ask ChatGPT to generate the dialogue context words from empathetic dialogue lines. As shown in the left part of Figure 2, the text prompt consists of 1) description of dialogue setting, 2) dialogue lines, and 3) request for answering context words.

Dialogue setting description explains the roles of the

speaker and listener as well as the dialogue situation to ChatGPT. We empirically found that presenting the dialogue situation improved the relevance of outcomes.

Dialogue lines describe the content of the conversation per turn. The format is a sequence of "[turn ID] [speaker's name] [content]" for each dialogue line. We limit the maximum turns in one prompt to 5 because when ChatGPT is asked to answer about long dialogues, it tends to hang in the middle of the answer. If one dialogue consists of more than five turns, we divide it into multiple queries overlapping two turns from the previous query. For example, a dialogue taking 10 turns is divided into prompts for 1–5, 3–7, 5–9, and 7–10 dialogue turns.

Request for answering context words asks ChatGPT to generate words describing the dialogue context for each line. We consider three kinds of context words: 1) dialogue intention [5], 2) emotion [6], and 3) speaking style. The categories of answers for the emotion and speaking style are { neutral, joy, anticipation, anger, disgust, sadness, surprise, fear, trust } (i.e., neutral and eight emotions defined by Plutchik [23]) and { cute, cool, quiet, polite, intellectual, honest, clear, gentle, gravelly, vibrant }, respectively.

3.2. EDSS model training using context word embeddings

We extract embeddings of the collected words by using BERT [24] and condition an EDSS model by the embeddings. Specifically, we define a ChatGPT-derived dialogue context vector as the sum of the word embeddings and train the EDSS model to predict empathetic dialogue speech from an input text and the context vector. This method regards ChatGPT as an interactive context estimator and replaces the CCE used in Saito et al.'s baseline EDSS method [8] with ChatGPT.

3.3. Discussion

Our ChatGPT-EDSS relates to text-predicted global style tokens (TP-GSTs) [25], a TTS method that predicts an expressive speaking style from a text-derived prosody embedding. From this perspective, our method uses ChatGPT to extract style-related words from dialogue lines and predicts the prosody embedding from the extracted words. Although TP-GSTs can improve the quality of synthetic speech better than a Tacotron-based TTS model [26], it cannot consider the dialogue history to train the TTS model, which is essential for reproducing an empathetic speaking style in EDSS [8]. However, one can introduce a similar idea that uses text-derived prosody embedding to predict the weight for each GST (i.e., predicting combination weights proposed in [25]) in the ChatGPT-EDSS training.

From another perspective, one can regard our ChatGPT-EDSS as weakly supervised learning [27] of expressive TTS

Table 1: *Averaged reliability scores and most frequent words appeared in each emotion category of collected context words*

	Reliability score	Intention	Emotion	Style
Neutral	3.95	問いかけ (question)	期待 (anticipation)	落ち着いた (quiet)
Happy	4.04	祝福 (blessing)	喜び (joy)	穏やか (gentle)
Angry	3.66	共感 (**empathy**)	信頼 (trust)	丁寧 (polite)
Sad	4.03	共感 (**empathy**)	悲しみ (sadness)	丁寧 (polite)

that uses the context words to condition the EDSS model instead of the ground-truth emotion label. We discuss the reliability of the collected context words later in Section 4.2 because Chat-GPT may generate an improper answer from the given prompt.

4. Experimental evaluation

We evaluated our ChatGPT using the STUDIES [8] corpus including Japanese empathetic spoken dialogues. The dialogue domain was chit-chat between a female teacher (empathetic listener) and students in a school.

4.1. Experimental conditions

This section describes the conditions for our evaluation.

Conditions for context word collection: We collected the context words for the long (10–20 turns) and short (4 turns) dialogues included in the STUDIES corpus using ChatGPT. The numbers of dialogues were 150 and 720, respectively. We employed 31 workers who asked ChatGPT to generate the context words and annotated the reliability score for each answer with an integer between 1 ("very unreliable") and 5 ("very reliable"). The workers completed these procedures by 1) accessing Google Sheets prepared by us, 2) copying the text prompts contained in the first field, 3) pasting the prompts to the Chat-GPT query field, 4) copying and pasting the ChatGPT answer to the second field, and 5) filling the reliable score in the third field[2]. We asked workers to resend the query to ChatGPT when 1) it failed to generate the context words, 2) the answer included a sentence other than the context word (e.g., the speaker's name or the original dialogue line), and/or 3) the language of the answer was not Japanese (e.g., English or Chinese).

Conditions for EDSS: We trained an EDSS model of the STUDIES teacher with the collected context words. Following Saito et al.'s study [8], we used 726, 72, and 72 dialogues for training, validation, and evaluation, respectively. We downsampled the speech data to 22,050 Hz. We used the validation data to choose hyperparameters for the following models, whose parameters were randomly initialized.

Acoustic model for EDSS: We used FS2 [12] as an acoustic model that predicted a mel-spectrogram from text with PyTorch implementation for Japanese TTS[3]. We followed the settings of a neural network architecture and speech parameter extraction of this implementation. We used the WORLD vocoder [28, 29] to estimate F_0. The optimizer was Adam [30] with an initial learning rate η of 0.0625, β_1 of 0.9, and β_2 of 0.98. We first pretrained FS2 using the JSUT corpus [31], a Japanese speech corpus including about 10 hours of a female speaker's speech, with 200K iterations. We then fine-tuned it by using the STUDIES training data with 100K iterations.

Neural vocoder: We used a HiFi-GAN vocoder [32] for speech waveform generation from a mel-spectrogram with PyTorch implementation[4] provided by the first author of the HiFi-

---

[2]We can automate this procedure excluding the reliability scoring because the ChatGPT API has become accessible since March 2, 2023.
[3]https://github.com/Wataru-Nakata/FastSpeech2-JSUT
[4]https://github.com/jik876/hifi-gan

Table 2: *Numbers of unique context words for each ground-truth emotion label*

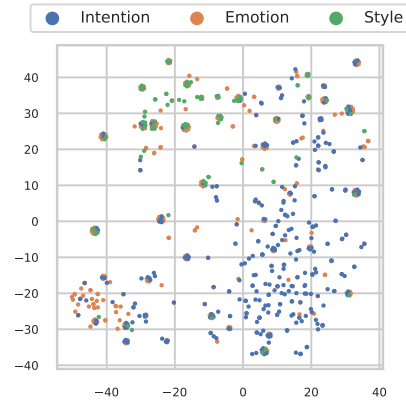|	Intention	Emotion	Style
Neutral	206	130	42
Happy	76	35	19
Angry	17	17	8
Sad	40	53	19

Figure 3: *t-SNE plot of BERT embeddings extracted from unique context words*

GAN paper. We trained HiFi-GAN by using the same training data as that for FS2 with 350K iterations. The optimizer was Adam with η of 0.0003, β_1 of 0.8, and β_2 of 0.99.

4.2. Analysis of ChatGPT-derived context words

Table 1 lists the results of the context word collection summarized in accordance with the ground-truth emotion labels of the STUDIES teacher. First, we found that the averaged reliability scores were more than 3.6 for all emotion labels. Second, the most frequent intention words for "Angry" and "Sad" utterances were "**empathy**." Third, the most frequent emotion word corresponded to the ground-truth emotion label except for "Neutral" and "Angry." Finally, the most frequent style words consisted of "quiet," "gentle," and "polite." These results suggest that ChatGPT 1) actually understands the intention of "empathetic" dialogue, 2) provides reliable weak labels for expressive TTS to some extent, and 3) roughly estimates the STUDIES teacher's speaking style as moderate.

Table 2 lists the number of unique context words for each ground-truth emotion label of the STUDIES teacher's utterances. First, the intention words were very diverse and consisted of more than 100 unique words for "Neutral" utterances. However, we found that 79% of these words only appeared five times or less. We observed a similar tendency from the results shown in the third and fourth columns in Table 2, despite the fact that we defined the categories of emotion and style in advance. We can confirm these diversities from the t-SNE visualization [33] of context word embeddings by BERT shown in Fig. 3, although the different categories tend to form roughly different clusters. These results indicate that ChatGPT 1) can

generate various candidate words for describing the context and 2) does not necessarily satisfy the pre-specified requirements for the word generation.

4.3. Subjective evaluations

We conducted subjective evaluations to investigate whether our ChatGPT-EDSS can reproduce the speaking style of empathetic dialogue speech without degrading naturalness.

Evaluation setup: We conditioned the FS2-based EDSS model on the following factors:

- **Emo**: Emotion label annotated by the corpus developers
- **CCE**: Data-driven context embedding extracted from dialogue history [7]
- **IES (ours)**: Embeddings of intention, emotion, and style words generated from ChatGPT

The **CCE** extracted the context embedding from joint vectors of one-hot encoding of speaker identity (3-dim.) and up to four sentence embedding sequences (one current sentence and up to three previous ones) obtained by using BERT (768-dim.) pre-trained by using Japanese text data[5]. The dimensionality of the context embedding was 256, and we prepared a linear layer to project the BERT-derived word embedding onto the 256-dimensional feature space.

As explained in Section 3.1, one sentence may have multiple context words due to the overlapping procedure. In that case, we aggregated the embeddings of multiple words for each intention, emotion, and style by simply taking the average of the embeddings. This aggregation may improve the robustness of our ChatGPT-EDSS towards the large variation of context words described in Section 4.2.

Evaluation criteria: We conducted a five-scale mean opinion score (MOS) test on the naturalness of synthetic speech. We presented 30 speech samples to listeners in random order. Listeners rated the naturalness of each speech sample from degrees of 1 ("very unnatural") to 5 ("very natural"). We recruited 50 listeners using our crowdsourcing subjective evaluation system. We also conducted a five-scaled MOS test on speaking-style similarity of synthetic speech. Listeners first listened to the reference speech reconstructed from a natural mel-spectrogram with HiFi-GAN and then scored the presented synthetic speech regarding the similarity of speaking style from degrees of 1 ("very dissimilar") to 5 ("very similar"). We recruited 50 listeners using our crowdsourcing subjective evaluation system.

Evaluation results: Table 3 shows the evaluation results. We found that using **IES** as the conditional features for the EDSS model performed comparably to using **Emo** or **CCE**. This result demonstrates that we can use ChatGPT as the context embedding extractor from empathetic dialogue lines instead of using the emotion label or conventional data-driven context embedding vector. We also observed that the two EDSS models using **Emo** only and the combination of **Emo** and **CCE** slightly degraded the naturalness MOS, while the one using both **Emo** and **IES** scored higher MOS values regarding the naturalness and similarity. One reason is the fine-grained (and possibly unlimited) emotion categories represented by the emotion words showing in Table 2, which enhances the expression ability of the EDSS model compared with the limited number of emotion categories in the ground-truth label (only four).

To further investigate the effects caused by introducing ChatGPT-derived context words in EDSS, we calculated the differences between the naturalness and similarity MOS of 1)

---

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

Table 3: *Results of MOS tests on speech naturalness and speaking style similarity with their 95% confidence intervals*

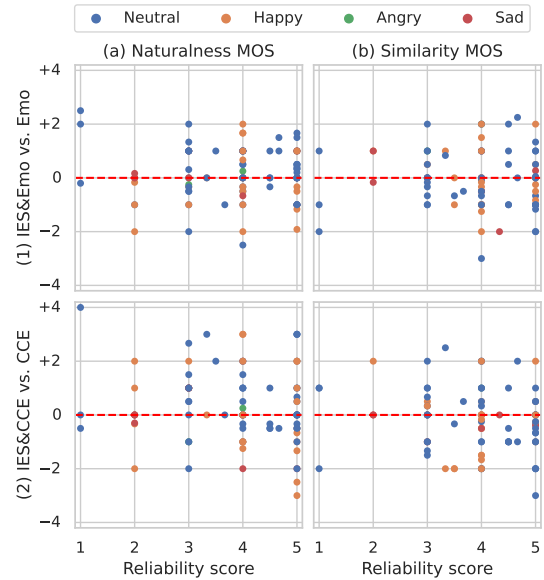| Method			MOS	
Emo	CCE	IES (ours)	Naturalness	Similarity
✓			3.43±0.14	3.20±0.15
	✓		3.54±0.14	3.24±0.14
		✓	3.52±0.14	3.19±0.15
✓		✓	3.52±0.14	3.21±0.14
✓	✓		3.43±0.14	3.24±0.14
	✓	✓	3.49±0.14	3.20±0.14

Figure 4: *MOS improvement with respect to reliability score of ChatGPT answer*

IES&Emo and **Emo** and 2) **IES&CCE** and **CCE** with respect to the reliability score for each ground-truth emotion label. Figure 4 shows the results. From this figure, we observe that there is no correlation between the reliability score and the MOS improvement and that the improvement varies widely even when the reliability score is 5. This result suggests that, although the introduction of **IES** does not negatively affect the quality of synthetic speech, we still require countermeasures against the large diversity of ChatGPT responses discussed in Section 4.2.

5. Conclusion

We proposed ChatGPT-EDSS, a ChatGPT-powered empathetic dialogue speech synthesis (EDSS) method using word embeddings of ChatGPT answers as the dialogue context. Our method first gives text-based chat history to ChatGPT as a prompt and asks it to generate three words related to the dialogue context: intention, emotion, and speaking style for each dialogue line. Then, it trains an EDSS model using embedding vectors of the three words as conditional features. The evaluation results demonstrated that our ChatGPT-EDSS performed comparably to ones using emotion labels or deeply-learned context embeddings extracted from chat histories. Our future work is to investigate the effect of the dialogue domain in ChatGPT-EDSS and to examine whether ChatGPT's hallucination occurs in our method or not.

6. References

[1] G. Bruce, B. Granström, K. Gustafson, M. Horne, D. House, B. Lastow, and P. Touati, "Speech synthesis in spoken dialogue research," in *Proc. EUROSPEECH*, Madrid, Spain, Sep. 1995, pp. 1169–1172.

[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] S. Kim and R. E. Banchs, "R-cube: a dialogue agent for restaurant recommendation and reservation," in *Proc. APSIPA ASC*, Siem Reap, Cambodia, Dec. 2014.

[4] 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.

[5] N. Hojo and N. Miyazaki, "Evaluating intention communication by tts using explicit definitions of illocutionary act performance," in *Proc. INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 1536–1540.

[6] T. S. Polzin and A. Waibela, "Emotion-sensitive human-computer interfaces," in *Proc. ITRW on Speech and Emotion*, Newcastle, Northern Ireland, U.K., Sep. 2000.

[7] 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.

[8] 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.

[9] 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, Aug. 2019, pp. 5370–5381.

[10] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2020.

[11] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Gray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Proc. NeurIPS*, New Orleans, U.S.A., Dec. 2022.

[12] 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.

[13] C. Leiter, R. Zhang, Y. Chen, J. Belouadi, D. Larionov, V. Fresen, and S. Eger, "ChatGPT: A meta-analysis after 2.5 months," *arXiv*, vol. arXiv:2302.1379, 2023.

[14] M. Kosinski, "Theory of mind may have spontaneously emerged in large language models," *arXiv*, vol. arXiv:2302.02083, 2023.

[15] H. Rao, C. Leung, and C. Miao, "Can ChatGPT assess human personalities? a general evaluation framework," *arXiv*, vol. arXiv:2303.01248, 2023.

[16] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is ChatGPT a general-purpose natural language processing task solver?" *arXiv*, vol. arXiv:2302.06476, 2023.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, Vancouver, Canada, Dec. 2020.

[18] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. ICML*, Virtual Conference, Jun. 2021, pp. 8821–8831.

[19] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually guided audio generation," in *Proc. ICLR*, Kigali, Rwanda, May 2023.

[20] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating music from text," *arXiv*, vol. abs/2301.11325, 2023.

[21] Z. Guo, Y. Leng, Y. Wu, S. Zhao, and X. Tan, "PromptTTS: Controllable text-to-speech with text descriptions," *arXiv*, vol. abs/2211.12171, 2022.

[22] J. Su, "Simbert: Integrating retrieval and generation into bert," Tech. Rep., 2020. [Online]. Available: https://github.com/ZhuiyiTechnology/simbert

[23] R. Plutchik, *Emotion: Theory, research, and experience: Vol. 1. Theories of emotion*. New York: Academic, 1980.

[24] 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.

[25] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *Proc. SLT*, Athens, Greece, Dec. 2018, pp. 595–602.

[26] 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," *arXiv*, vol. abs/1703.10135, 2017.

[27] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National Science Review*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[28] 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.

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

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

[31] 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.

[32] 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.

[33] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *The Journal of Machine Learning Research*, vol. 86, no. 9, pp. 1929–1958, 2008.