



Improving User Impression of Spoken Dialogue Systems by Controlling Para-linguistic Expression Based on Intimacy

Shoki Kawanishi¹, Akinori Ito¹, Yuya Chiba², Takashi Nose¹

¹Graduate School of Engineering, Tohoku University, Japan

²NTT Communication Science Laboratories, Japan

shoki.kawanishi.t6@dc.tohoku.ac.jp, aito.spc@tohoku.ac.jp,
yuya.chiba@ntt.com, takashi.nose.b7@tohoku.ac.jp

Abstract

The advent of large language models (LLMs) has improved the naturalness of responses in dialogue systems; however, conversations with these systems still differ significantly from those between humans. For instance, humans adjust their manner of speaking based on their relationship with the conversation partner, whereas dialogue systems respond uniformly, even after repeated interactions. By producing responses that consider the progression of relationships, it may be possible to develop dialogue systems that are more appealing to users. Previous studies attempting to achieve such systems have primarily focused on the linguistic expressions of responses, while para-linguistic expressions have been largely overlooked. In this paper, we propose a dialogue system that adapts both linguistic and para-linguistic expressions as the number of interactions increases. We also evaluate its effectiveness through dialogue experiments.

Index Terms: Spoken dialogue systems, intimacy-based dialogue management, conversational speech synthesis

1. Introduction

The development of large language models (LLMs) has significantly improved the naturalness of responses in dialogue systems. Recent AI systems, exemplified by ChatGPT, have made dialogue interfaces ubiquitous. However, conversations with such dialogue systems differ significantly from human-to-human interactions. One key difference is that these systems cannot engage in conversations while taking into account the relationship with the conversation partner. For instance, in typical dialogue systems, no matter how many times a user interacts with them, the assumed relationship remains unchanged, and thus the conversation is always conducted in a uniform tone. For dialogue systems to be more widely accepted as social entities, it is necessary to introduce dialogue strategies that incorporate the framework of human communication.

In human conversations, individuals are known to adjust their behavior based on their sense of closeness to their dialogue partner. According to the well-established politeness theory [1], people modify their language use and behavior depending on social distance and relative power dynamics in an interaction. Some languages, such as Japanese, have an explicit linguistic mechanism called honorifics, which alters speech style based on the relative social position or distance between speakers [2]. Adjustments in behavior corresponding to social distance are not limited to linguistic expressions. Research has shown that nonverbal behaviors, such as prosody, gaze, and pauses, also vary depending on the level of intimacy between dialogue partners [3]. Furthermore, several studies have analyzed dialogue behaviors by examining different stages of re-

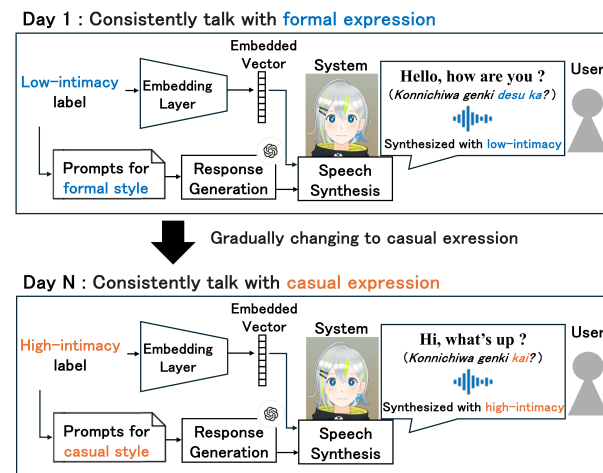


Figure 1: Gradual control of linguistic and para-linguistic expressions in system utterances.

lationships, such as friendships, acquaintances, and close confidants. Hornstein [4] found that friends tend to use more implicit openings, introduce topics more frequently, and exhibit greater responsiveness through questions. Additionally, factors such as floor time distribution, the number of interruptions [5], and various shared activities [6] also change depending on the stage of the relationship. Humans tend to expect similar interaction patterns even when communicating with non-human entities, such as computers [7]. Therefore, incorporating relationship dynamics into dialogue systems is considered promising. Various methods have been proposed to enable dialogue systems to comprehend and then leverage the relationship with users (e.g., [8]). Among these approaches, Kageyama et al. [9] focused on Japanese honorifics and investigated a method for gradually adjusting a system's utterance style based on the number of interactions. Their study demonstrated that transitioning a system's utterances from a formal to a casual style over three days of interaction improved user impressions of the system. However, their research primarily emphasized linguistic changes and overlooked variations in para-linguistic expressions, such as prosody and intonation.

In this study, we propose a method that modifies both linguistic expressions and para-linguistic expressions (Fig. 1). In the proposed method, the system adjusts its speech style from formal to casual, based on the number of interactions with the user, while simultaneously increasing the level of intimacy in speech from low to high. To achieve this, we developed a speech synthesis system capable of generating speech with varying lev-

els of intimacy. The effectiveness of the proposed method is evaluated through dialogue experiments.

2. Related work

Various studies on dialogue systems have been conducted with a focus on the relationship between the system and the user. Bickmore et al. [10] built a relational agent that introduced “immediacy” behavior [8] to support multiple interactions with users over an extended period. Kanda et al. [11] demonstrated that a guidance robot in a shopping mall could sustain user engagement by delivering friendly utterances to returning users. Kim et al. [12] confirmed that systems that greet users by name are perceived as more friendly. A study by Kageyama et al. [9], which focuses on the system’s speech style, also falls into this category. For response generation using an LLM, other studies have introduced information to the relationship with the user in order to facilitate long-term dialogue [13, 14]. While these studies consider simple agent behaviors, they primarily focus on linguistic expressions. As a result, the effects of para-linguistic expressions in speech have not been sufficiently examined.

On the other hand, a speech synthesis system plays a crucial role in conveying intimacy through para-linguistic expressions in speech. In recent years, several efforts have been made to synthesize speech with a more natural conversational tone. Iizuka et al. [15] developed a speech synthesis system trained on spontaneous conversational speech and demonstrated that it was highly evaluated by users in dialogue interactions. Moreover, research on spoken dialogue models has been rapidly progressing. For example, end-to-end models have been developed to process human-to-human spoken dialogue as both input and output, exemplified by the Generative Spoken Language Model (GSLM) [16, 17] and Moshi [18]. While these models can generate highly natural conversational-style speech, they cannot control speech based on system states such as intimacy. In contrast, in this study, we develop a speech synthesis system capable of generating conversational speech that expresses different levels of intimacy.

3. Intimacy-based Dialogue Management

3.1. Speech synthesis system to condition for the level of intimacy

First, we describe a speech synthesis system for controlling the intimacy level of para-linguistic expressions. To implement the proposed dialogue system, we develop a speech synthesis system that learns para-linguistic features corresponding to different levels of intimacy. The target speech synthesis system generates speech with acoustic characteristics based on the specified intimacy level, taking an utterance text and a label representing the degree of intimacy as input. The model is based on the multi-speaker synthesis model proposed by Cooper et al. [19]. Their model consists of three components: a speaker embedding network, a Mel-cepstrum generation component from Tacotron 2 [20], and a neural vocoder. The speaker embedding network takes a label representing the speaker as input and generates speaker-dependent speech. This network is connected to the Mel-cepstrum generation component. We replace the speaker embedding network with an intimacy embedding network that takes the level of intimacy as input. This enables the model to synthesize speech according to the level of intimacy.

The model is trained using human-to-human dialogue data annotated with subjective labels of intimacy toward the conver-

sational partner. The intimacy level, as in the previous study [9], is treated as a binary label (high or low), and the intimacy embedding network accepts binary inputs. In this model, the output of the intimacy embedding network is added to the encoder output of the Mel-cepstrum generation component. While the original model [19] employed a multi-layer encoder for speaker embeddings, we use a single-layer encoder because the intimacy label is represented as a binary vector.

3.2. Gradually adjusting intimacy-level of system speech

In the proposed dialogue management, the system gradually adjusts the level of intimacy in its responses as the dialogue progresses. The intimacy of the system’s utterance is expressed through both linguistic and para-linguistic expressions. The system’s responses are generated using an LLM. Changes in the content of the system’s utterances and the level of intimacy are guided by prompts. Two types of prompts are prepared for generating system responses: one for high intimacy and one for low intimacy. While the basic content of the two prompts is identical, the instructions regarding the level of intimacy differ. In the common part of the prompts, the system is instructed to act as a conversational agent that is knowledgeable in a wide range of topics. In addition, the system is instructed to include in its responses: 1) a reaction to the user’s utterance, 2) self-disclosure on that topic, and 3) a follow-up question related to the previous user utterance.

To control the intimacy level of linguistic expression, the system is instructed to adjust the style of utterances. In the prompt for low intimacy, the system is directed to use a polite and formal style. In contrast, for high intimacy, the system is instructed to speak in a casual style, as one would to a close friend. This approach, similar to the previous study [9], is expected to adjust the style of the system utterance according to the level of intimacy while minimizing changes to the content of the utterances to the extent possible.

4. Experimental Conditions

4.1. Dataset for training speech synthesis models

We used the Japanese Spontaneous Multimodal One-on-one Chat-talk (SMOC) corpus [21] as training data for speech synthesis. This corpus contains 510 Japanese dialogues recorded from 71 participants, including 19 women. The dialogues include interactions both between pairs meeting for the first time and pairs already familiar with each other. For our experiments, we selected a total of 95 dialogues recorded from 19 participants. Utterances from conversations between participants meeting for the first time were classified as “low-intimacy,” while those from conversations between acquaintances with subjective intimacy scores of 4 or 5 were classified as “high-intimacy” [3]. Utterances with extremely long or short durations were excluded from the experiments. As a result, we obtained a total of 4,700 utterances, comprising 2,563 high-intimacy utterances and 2,137 low-intimacy utterances. In addition, we selected 100 utterances (55 with high-intimacy and 45 with low-intimacy levels) for the test set.

Since the SMOC data consisted of real conversational speech from multiple speakers, there was considerable variation in phoneme distribution across them. To reduce speaker variability and facilitate effective model training, we applied voice conversion to standardize the utterances to a single speaker’s speech. The target speaker for the voice conversion was a female speaker from the JSUT corpus [22], which contains

Japanese single-speaker reading speech. For voice conversion, we utilized the Retrieval-based Voice Conversion (RVC) tool¹. A total of 4,700 utterances, selected from the BASIC5000 set of the JSUT corpus, were used as training data for the RVC. The model was trained for 30 epochs. If the original speech was from a male speaker, the fundamental frequency was increased by one octave after conversion. Finally, all samples were down-sampled to 16 kHz.

4.2. Training speech synthesis models

As the neural vocoder, we used HiFi-GAN [23], which is implemented in ESPnet2 [24]. The Mel-spectrogram generation component was constructed using a GitHub repository compatible with ESPnet2². The intimacy label embedding vector was applied to the encoder output. The dimensionality of the intimacy embedding network was set to 512.

For training Tacotron 2, we used the Adam optimizer having a learning rate of 10^{-3} and trained for 100K steps. The batch type during training was set to `numel`, with `batch_bins` set to 3,750,000. HiFi-GAN training followed its default discriminator and generator architectures. Here, the Adam optimizer had a learning rate of 10^{-4} . The training process consisted of 2.5M steps with a batch size of 16.

4.3. Construction of dialogue system

The experimental system was built using the Real-Time Multimodal Dialogue System Toolkit (Remdis) [25]. This toolkit natively supports response generation using ChatGPT. For the response generation model, we used `gpt-4`. The entire dialogue history up to the target utterance on that day was provided for the model. This study does not use the Voice Activity Projection (VAP) module [26, 27]; instead, the speech recognition confirmation timing is used as the turn-shift point.

To reduce discomfort during conversations with the system, a CG agent was rendered on a monitor. MMDAgent-EX³ was used for rendering the CG agent, with Gene⁴ selected as the agent character. Although MMDAgent-EX supports controlling the emotions and actions of the CG agent, these were fixed to “neutral” and “idle,” respectively, to eliminate their influence on user evaluations. Captions displaying both user and system utterances were shown at the bottom of the screen.

4.4. Experimental setup

The experiments were designed based on the previous study by Kageyama et al. [9]. Participants were asked to engage in spoken dialogue with the system, completing 10 exchanges of utterances per day for three consecutive days. During the experiment, the proposed system gradually adjusted the intimacy level of its speech, transitioning from polite to casual over the three-day period. On the first day, all responses were presented in a polite tone. On the second day, responses were presented in a 50-50 ratio of polite and casual tones. On the third day, all responses were delivered in a casual tone. The conditions for changing expressions followed the approach described in the previous study [9]. Changes to the linguistic expressions were achieved by switching the prompts used by the system. Changes

¹<https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI/releases>

²<https://github.com/kan-bayashi/ParallelWaveGAN>

³<https://github.com/mmdagent-ex/MMDAgent-EX>

⁴<https://github.com/mmdagent-ex/gene>

Table 1: Subjective evaluation items.

Index	Question
Q1	Did you feel satisfied with the dialogue? (<i>Satisfaction</i>)
Q2	Did you feel a sense of familiarity with the system? (<i>Friendliness</i>)
Q3	Did you have a positive impression of how the system spoke when listening to its responses? (<i>Impression</i>)
Q4	Would you like to talk with the system again? (<i>Intention of Talk</i>)
Q5	Were the system’s responses natural? (<i>Naturalness</i>)
Q6	Did the system’s responses align with what you said? (<i>Engagement</i>)
Q7	Were the system’s responses consistent? (<i>Consistency</i>)
Q8	Were the system’s responses accurate? (<i>Accuracy</i>)
Q9	Did you regard the system as favorable? (<i>Likeability</i>)
Q10	Was your conversation with the system frustrating? (<i>Annoyance</i>)
Q11	Did continuing the conversation with the system require concentration? (<i>Cognitive Demand</i>)

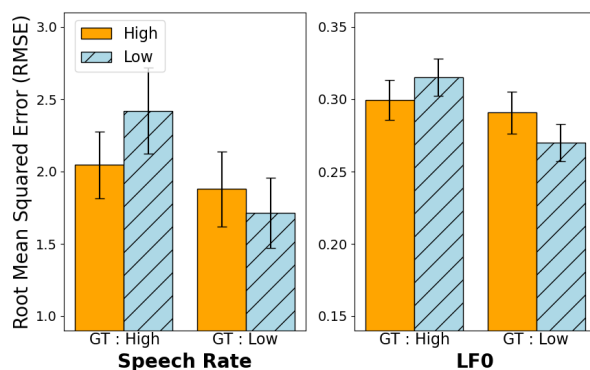


Figure 2: RMSE of speech rate and log F0 (mean \pm SE).

to the para-linguistic expressions were made by switching the intimacy labels that were input to the speech synthesis system. To investigate the effectiveness of changing to para-linguistic expressions as the number of interactions increases, we compared the following three conditions:

- Casual:** Consistently responds with speech synthesized at a high-intimacy level;
- Polite:** Consistently responds with speech synthesized at a low-intimacy level;
- Proposed:** Adjusts the intimacy level of synthesized speech based on the number of utterances, where the rate of change matches that of the linguistic expression.

The experiment involved 11 participants in each condition, for a total of 33 participants (including 2 females). After each dialogue session with the system, participants completed a subjective evaluation survey. The survey items were selected from commonly used evaluation criteria for dialogue systems (i.e., *satisfaction*, *naturalness*, *consistency*, and *accuracy*), as well as scales used in previous studies [9] and items of the Subjective Assessment of Speech System Interfaces (SASSI) [28]. The selected items are listed in Table 1. Participants rated each item on a 5-point scale (1: lowest, 5: highest).

5. Experimental Results

5.1. Objective evaluation of speech synthesis

To verify whether the trained speech synthesis model could generate speech reflecting intimacy, we first performed an objective

Table 2: Subjective evaluation results. Underlined scores represent the best-rated day for each condition. Bold texts indicate the scores of the best-rated condition on the last day.

Item	Description	Casual			Polite			Proposed		
		Day1	Day2	Day3	Day1	Day2	Day3	Day1	Day2	Day3
Q1	<i>Satisfaction</i> (↑)	3.18	<u>3.45</u>	3.27	2.82	<u>3.45</u>	<u>3.45</u>	2.82	3.55	4.00
Q2	<i>Friendliness</i> (↑)	2.82	3.55	<u>3.91</u>	3.27	3.82	<u>4.00</u>	2.73	3.55	4.09
Q3	<i>Impression</i> (↑)	2.55	3.27	3.55	2.73	3.09	3.55	2.45	3.18	3.55
Q4	<i>Intention of Talk</i> (↑)	3.45	3.91	4.09	3.45	<u>4.00</u>	<u>4.00</u>	3.09	3.64	<u>4.00</u>
Q5	<i>Naturalness</i> (↑)	3.00	<u>3.36</u>	3.09	3.00	3.00	3.73	2.73	<u>3.45</u>	3.36
Q6	<i>Engagement</i> (↑)	3.27	<u>3.36</u>	3.00	3.27	3.09	<u>3.45</u>	2.82	3.55	3.73
Q7	<i>Consistency</i> (↑)	<u>4.36</u>	4.00	3.82	3.64	<u>3.73</u>	3.64	3.73	3.64	4.18
Q8	<i>Accuracy</i> (↑)	3.64	<u>3.82</u>	3.27	3.09	3.09	<u>3.27</u>	3.09	<u>3.73</u>	3.45
Q9	<i>Likeability</i> (↑)	3.36	<u>3.64</u>	3.45	3.27	3.36	3.91	3.00	3.64	<u>3.82</u>
Q10	<i>Annoyance</i> (↓)	2.27	1.91	1.82	2.82	2.27	<u>2.09</u>	2.82	2.27	<u>2.00</u>
Q11	<i>Cognitive Demand</i> (↓)	<u>2.64</u>	2.82	2.73	3.82	3.45	<u>3.27</u>	3.91	3.00	2.18

evaluation. We compared prosodic features between synthesized speech and reference speech for the SMOC test set. The speech was generated with the same text as the reference for each intimacy level. Here, we focused on speech rate and log F0 as the prosodic features and calculated the Root Mean Squared Error (RMSE). The previous study [3] found that these prosodic features differed according to intimacy level. For F0, we calculated RMSE frame-by-frame. In cases where the lengths of the original and synthesized speech differed, the shorter sequence was linearly interpolated to match the length of the longer one.

Figure 2 shows the results, where “GT: High” and “GT: Low” represent the ground truth intimacy labels of the reference speech. From the graph, it can be observed that the RMSE values are lower for the groups of synthesized speech generated with the same intimacy level as the ground truth labels, in both speaking rate and log F0. This suggests that the constructed speech synthesis system can generate audio that reflects the level of intimacy.

5.2. Results of subjective evaluation obtained through dialogue experiments

The results of the subjective evaluation indicate that, regardless of condition, user evaluations tend to improve as the experiments progress (Table 2). This finding aligns with previous research [9], suggesting that gradually adjusting the system’s utterance style based on the number of interactions is effective for enhancing user evaluation. In particular, although the Proposed condition initially received lower scores than the Casual condition for many items on the first day, its scores improved over time, ultimately achieving the highest score in most items on the final day. This outcome is likely due to the effect of gradually adjusting both linguistic and para-linguistic expressions. In the following discussion, we focus on Q1, Q2, Q5, and Q11.

First, regarding Q1 (*Satisfaction*), which reflects an overall evaluation of the dialogue system, the Proposed condition showed an increase in scores over time, achieving the highest rating on the final day (i.e., 4.00). Similarly, the Casual and Polite conditions also exhibited improved scores as the days progressed; nevertheless, their scores remained around 3.5. These results suggest that, in addition to linguistic information, adjusting para-linguistic information, specifically the level of intimacy in speech, based on the number of interactions is effective for enhancing the overall evaluation of the dialogue system.

Next, regarding Q11 (*Cognitive Demand*), the Casual condition received the best user rating on the first day. This suggests that a system that consistently responds with high-intimacy speech has a certain effect in reducing cognitive load. How-

ever, in the Casual condition, the Q11 score remained relatively stable with little variation. On the other hand, in the Proposed condition, the score decreased over time, eventually falling below that of the Casual condition on the final day. This indicates that gradually adjusting para-linguistic expressions is effective in reducing cognitive load. In the Polite condition, a decrease in score was also observed, which may be attributed to users becoming more familiar with the system. However, the decrease was not as pronounced as in the Proposed condition, and the final score did not reach the level of the Casual condition.

Regarding Q5 (*Naturalness*), which evaluates the naturalness of system responses, the ratings were generally comparable across all conditions. Since no changes other than style were instructed for response generation, this result seems reasonable. This suggests that, at least in terms of naturalness, all systems provided responses with comparable content. Additionally, Q2 (*Friendliness*) was expected to show a significant effect in controlling para-linguistic expressions based on intimacy, as examined in this study. Although the Proposed condition received the highest rating on the third day, the improvement was slight. This suggests that intimacy may be less effectively conveyed through para-linguistic expressions of speech than through linguistic expressions.

In addition, in the experiments conducted in this study, no significant differences among conditions were observed on Day 3 for any of the evaluation items. However, in 7 out of the 11 items, the Proposed condition received higher ratings than did the other conditions on the third day, suggesting that the effect of the proposed method on user evaluations may become clear with continued interactions. In the future, we plan to conduct longer-term dialogue experiments to further investigate the impact on user evaluations.

6. Conclusions

In this paper, we proposed a dialogue system that adapts both linguistic and para-linguistic expressions as the number of interactions increases. To develop this system, we constructed a speech synthesis system capable of adjusting the level of intimacy. The results of dialogue experiments revealed that a gradual adjustment of the intimacy level in para-linguistic expressions further improved user evaluations, such as satisfaction, compared to only changing linguistic expressions.

In the future, we plan to expand para-linguistic expressions based on intimacy to include multimodal behaviors, such as gestures and facial expressions of the agent. Additionally, it will be necessary to evaluate the impact on user evaluations by conducting continuous dialogue experiments over an extended period.

7. References

- [1] R. Brown and M. Ford, "Address in American English." *The Journal of Abnormal and Social Psychology*, vol. 62, no. 2, pp. 375–385, 1961.
- [2] S. Ikuta, "Speech level shift and conversational strategy in Japanese discourse," *Language Sciences*, vol. 5, no. 1, pp. 37–53, 1983.
- [3] Y. Chiba and A. Ito, "Speaker intimacy estimation in chat-talks based on verbal and non-verbal information," *IEEE Access*, pp. 184 592–184 606, 2024.
- [4] G. Hornstein, "Intimacy in conversational style as a function of the degree of closeness between members of a dyad," *Journal of Personality and Social Psychology*, vol. 49, no. 3, pp. 671–681, 1985.
- [5] S. Planalp, "Friends' and acquaintances' conversations II: Coded differences," *Journal of Social and Personal Relationships*, vol. 10, no. 3, pp. 339–354, 1993.
- [6] M. Rands and G. Levinger, "Implicit theories of relationship: An intergenerational study," *Journal of Personality and Social Psychology*, vol. 37, no. 5, pp. 645–661, 1979.
- [7] B. Reeves and C. Nass, "The media equation: How people treat computers, television, and new media like real people," *Cambridge, UK*, vol. 10, no. 10, pp. 19–36, 1996.
- [8] T. W. Bickmore and R. W. Picard, "Establishing and maintaining long-term human-computer relationships," *ACM Transactions on Computer-Human Interaction*, vol. 12, no. 2, pp. 293–327, 2005.
- [9] Y. Kageyama, Y. Chiba, T. Nose, and A. Ito, "Improving user impression in spoken dialog system with gradual speech form control," in *Proc. SIGDIAL*, 2018, pp. 235–240.
- [10] T. Bickmore, L. Caruso, and K. Clough-Gorr, "Acceptance and usability of a relational agent interface by urban older adults," in *Proc. CHI*, 2005, pp. 1212–1215.
- [11] T. Kanda, M. Shiomi, Z. Miyashita, H. Ishiguro, and N. Hagita, "An affective guide robot in a shopping mall," in *Proc. HRI*, 2009, pp. 173–180.
- [12] Y. Kim, S. Kwak, and M.-S. Kim, "Am I acceptable to you? Effect of a robot's verbal language forms on people's social distance from robots," *Computers in Human Behavior*, vol. 29, no. 3, pp. 1091–1101, 2013.
- [13] H. Kim, J. Hessel, L. Jiang, P. West, X. Lu, Y. Yu, P. Zhou, R. Bras, M. Alikhani, G. Kim, M. Sap, and Y. Choi, "SODA: Million-scale dialogue distillation with social commonsense contextualization," in *Proc. EMNLP*, 2023, pp. 12 930–12 949.
- [14] N. Chen, H. Li, J. Chang, J. Huang, B. Wang, and J. Li, "Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations," in *Proc. COLING*, 2025, pp. 755–773.
- [15] T. Iizuka and H. Mori, "How does a spontaneously speaking conversational agent affect user behavior?" *IEEE Access*, vol. 10, pp. 111 042–111 051, 2022.
- [16] T. A. Nguyen, E. Kharitonov, J. Copet, Y. Adi, W.-N. Hsu, A. Elkahky, P. Tomasello, R. Algayres, B. Sagot, A. Mohamed *et al.*, "Generative spoken dialogue language modeling," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 250–266, 2023.
- [17] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baeovski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.
- [18] A. Défossez, L. Mazaré, M. Orsini, A. Royer, P. Pérez, H. Jégou, E. Grave, and N. Zeghidour, "Moshi: a speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, pp. 1–67, 2024.
- [19] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *Proc. ICASSP*, 2020, pp. 6184–6188.
- [20] 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*, 2018, pp. 4779–4783.
- [21] Y. Yamazaki, Y. Chiba, T. Nose, and A. Ito, "SMOC corpus: A large-scale Japanese spontaneous multimodal one-on-one chat-talk corpus for dialog systems," *Acoustical Science and Technology*, vol. 42, no. 4, pp. 210–213, 2021.
- [22] 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.
- [23] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.
- [24] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, "ESPnet2-TTS: Extending the edge of TTS research," *arXiv preprint arXiv:2110.07840*, 2021.
- [25] Y. Chiba, K. Mitsuda, A. Lee, and R. Higashinaka, "The Remdis toolkit: Building advanced real-time multimodal dialogue systems with incremental processing and large language models," in *Proc. IWSDS*, 2024, pp. 1–6.
- [26] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," in *Proc. INTERSPEECH*, 2022, pp. 5190–5194.
- [27] —, "How much does prosody help turn-taking? Investigations using voice activity projection models," in *Proc. SIGDIAL*, 2022, p. 541–551.
- [28] K. S. Hone and R. Graham, "Towards a tool for the subjective assessment of speech system interfaces (SASSI)," *Natural Language Engineering*, vol. 6, no. 3–4, pp. 287–303, 2000.