Interspeech 2024
1-5 September 2024, Kos, Greece

# Contextual Interactive Evaluation of TTS Models in Dialogue Systems

*Siyang Wang, Éva Székely, Joakim Gustafson*

Department of Speech, Music, and Hearing, KTH Royal Institute of Technology
{siyangw},{szekely},{jkgu}@kth.se

## Abstract

Evaluation of text-to-speech (TTS) models is currently dominated by Mean-Opinion-Score (MOS) listening test, but MOS has been increasingly questioned for its validity. MOS tests place listeners in a passive setup, in which they do not actively interact with the TTS and usually evaluate isolated utterances without context. Thus it gives no indication for how well a TTS model suits an interactive application like spoken dialogue system, in which the capability of generating appropriate speech in the dialogue context is paramount. We aim to take a first step towards addressing this shortcoming by evaluating several state-of-the-art neural TTS models, including one that adapts to dialogue context, in a custom-built spoken dialogue system. We present system design, experiment setup, and results. Our work is the first to evaluate TTS in contextual dialogue system interactions. We also discuss the shortcomings and future opportunities of the proposed evaluation paradigm.

**Index Terms**: text-to-speech, spoken dialogue system, evaluation methodology, human-computer interaction

## 1. Introduction

Text-to-Speech (TTS) technology has advanced drastically after the introduction of deep learning methods [1]. State-of-the-art neural TTS models have achieved human-level naturalness and prosody variation, albeit only in read-speech utterances, or with limited contextual input. At the same time, the amount of spoken dialogue applications built with large language models (LLMs) as the text backend is increasing quickly [2]. These systems demand TTS to be capable of synthesizing human-like conversational speech in context, however such capability is a major shortcoming of current TTS models. A limiting factor in developing better contextual conversational TTS systems is the lack of contextual and interactive evaluation methods [3]. Current TTS evaluation is dominated by Mean-Opinion-Score (MOS) listening test on isolated utterances, which is neither contextual nor interactive. There have been some efforts to infuse contextual information to conventional MOS tests [4, 5, 6], or using longer contexts in evaluation of TTS for long read texts [7, 8]. Another work evaluated TTS in human-controlled (WoZ) dialogue systems as alternative to traditional, non-interactive MOS tests [9], but it lacks evaluation in an autonomous dialogue system and contextual appropriateness evaluation.

In our study, we aim to establish an experiment setup where participants engage with a singular AI entity, but through different TTS models acting as the speech synthesizer as seen in Figure 1. Participants are agnostic to the fact that only TTS models are different between systems. The models are trained on the same speech corpus thus there is minimal speaker timbre difference between them. Moreover, we also examine whether user perceptions vary when the dialogue system manipulates prosodic aspects of the speech synthesis according to the context of interaction. We measure and evaluate interactions at three levels: 1) questionnaires (Godspeed [10] and MOSx [11]), 2) behavioral metrics (measuring prosody, turn-taking times, and no of words in the user input), and 3) overall preference (through post-experiment interviews). We provide discussions of pros and cons the proposed evaluation paradigm at the end. Conversational TTS samples are found here: https://www.speech.kth.se/tts-demos/interspeech2024-contextual.

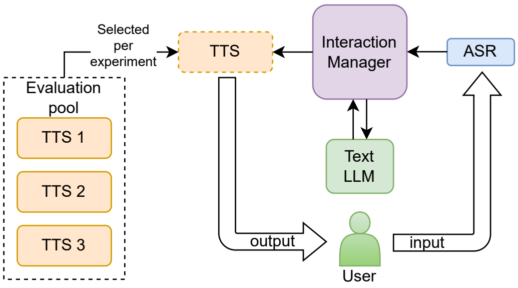

Figure 1: *Dialogue system for TTS evaluation.*

## 2. Related work

In the TTS community, there is an established consensus on the necessity of evolving beyond conventional sentence-level listening tests, such as Mean Opinion Score (MOS) and Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) scores [3]. For evaluation of read speech synthesis, there have been innovations such as evaluating in longer reading contexts that more accurately reflect real-world use [12], evaluating with more comprehensive criteria including personality, narrative distinction, and extended listenability [13], as well as adherence to an additional prompt input when it is available to the system [14]. Moreover, evaluations are beginning to assess systems' capacities for generating nuanced, emotionally resonant character dialogue, albeit still with listening tests [15, 16].

The accelerated development of conversational TTS has led to models being trained using actor-read chatbot transcripts and role-play dialogues to deliver natural and context-appropriate speech [17, 18, 19, 20]. It has also led to increased effort to innovate conversational TTS evaluation. Mitsu et al., created a conversational TTS from dialogues between two actors, and evaluated the resulting model by re-synthesizing part of the training dialogues and asked listeners to rate naturalness and appropriateness [21]. A similar study [22] used both tra-

10.21437/Interspeech.2024-1008

ditional objective metrics (e.g. duration, mel-cepstral distortion and mel-spectral distortion) and more novel subjective metrics like emotional expression accuracy and conversational fluidity. Broadening the evaluation criteria beyond conventional MOS metrics, recent studies have also delved into the proficiency of TTS systems in supporting fluid turn-taking scenarios [23, 24].

Evaluations involving real dialogue system interactions are scarce. One particular study involved 15 participants and two tourist guide robots, one with a charismatic speaking style featuring higher pitch and faster rate, found a strong preference for the charismatic robot, perceived as more enthusiastic and charming [25]. Similarly, an experiment with the Amazon Alexa Prize socialbot Gunrock tested the effect of using interjections and fillers in responses. Engaging 5,000 users with four system variants, the study found that versions incorporating these verbal cues were rated highest [26]. A study more similar to ours is [9], but the conversational system in that study was not autonomous, thus the TTS prompts were pre-synthesized.

In this study, our objective is to examine TTS models in a real interaction setting in a smart home lab. We developed a fully autonomous dialogue system within the context of playing a guessing game. Participants played four sessions with each TTS model as the speech synthesizer in the dialogue system, allowing for comprehensive exposure and robust evaluation. Additionally, our analysis combines objective measures of user utterances and subjective measures in the form of questionnaires and post-interviews.

# 3. System

## 3.1. Corpus and TTS models

### *3.1.1. Corpus of spontaneous conversations*

Developers of conversational systems should use TTS voices trained on ecologically valid data [27]. Our goal is to build effective conversational systems, concequently we trained our models on our AptSpeech corpus, a multimodal, multi-party interaction dataset [28]. This corpus includes 15 one-hour sessions where a male American moderator guides two participants in decorating an apartment. The moderator engage in a range of spontaneous speaking styles like small talk, instructions, advice and self-directed speech [6]. The moderator's speech was used for TTS training data. The moderator also recorded a set of read speech utterances in a separate session. We used both the conversational and read speech corpora to build a 2-speaker TTS where the "speakers" are the two speaking styles. By controlling the mix of these speakers we get explicit style control from read to conversational.

### *3.1.2. Tacotron2 and context adaptation*

The first TTS model we trained and evaluated is the popular neural TTS model *Tacotron2* [29]. Tacotron2 is representative of neural acoustic models that have significantly increased naturalness compared to prior modeling paradigms. We also trained a modified model that is capable of controlling the output prosody and synthesizing from multi-style corpus introduced and made publicly available in [30]. We connect the prosody and speaking style controls with the dialogue system output to adapt synthesis output to the dialogue context. We call this model *Tacotron2-context-adapted*. Both Tactoron2 models have 28.2M parameters, are each trained on a single NVIDIA 1080 until convergence. Utterance inference time is 0.5-1.0s.

### *3.1.3. Matcha-TTS*

The second model we evaluated is *Matcha-TTS* [31], a recently proposed non-autoregressive neural TTS engine applying latest development in flow-matching diffusion to achieve state-of-the-art naturalness while enabling fast inference. A multi-speaker architecture using different speaker embeddings for the read and spontaneous speech parts of the corpus similar to that in Tacotron2 has been applied. We use a HifiGAN vocoder in all TTS systems [32]. The model has 18.2M parameters, is trained on a single NVIDIA 3090 GPU until convergence. Its inference time per utterance is 0.5-1.0s.

## 3.2. Dialogue system

To evaluate different TTS systems in a conversational context, we designed a dialogue system that combines the strong generative capabilities of a large language model (GPT-4) with a structured interaction context to ensure both robustness and expressivity. We chose the ”20 Questions” guessing game as the context, elaborated in Section 4.1. The architecture of our dialogue system, depicted in Figure 1, features the Interaction Manager as the core component. This manager handles turn-taking, task progression, and the flow of information between: *speech recognition* (Google ASR), *dialogue management* (GPT-4 [33]), and *speech synthesis* (the evaluated TTS models). The system operates in two modes, corresponding to the roles in the guessing game: *system-as-answerer* and *system-as-questioner*, that share a persona prompt for GPT-4. This prompt outlines an expressive chatbot characterized by fillers and emotional reactions such as self-reproach and humorous remarks. To optimize response times, the Interaction Manager generates filler phrases while awaiting GPT-4's next response, reducing the average response time from 3 seconds to 1 second and enhancing turn-taking dynamics. These feedback tokens are chosen based on the difficulty of the user's question or the response to a question posed by the system in the previous turn. The latter are generated by GPT-4 together with the question.

The Interaction Manager begins by synthesizing instructions and submitting prompts and instructions to the GPT-4 server. It then sends the chatbot response to the local TTS server and activates the microphone, streaming audio to Google ASR for voice activity detection and speech recognition. Subsequently, it combines the ASR results with the current dialogue context and sends them to GPT-4 for processing. During this period, it generates suitable turn-taking feedback utterances. The Interaction Manager also controls the speaking style and prosodic features of the Tacotron2-context-adapted TTS model to adapt its output to the dialogue context. It achieves this in a rule-based manner. It monitors the progress of the game, and when it encounters a series of correct guesses, the system gradually adopts an enthusiastic, rapid, and high-pitched tone. Conversely, for incorrect guesses, the system gradually reduces these levels slightly, to diversify the prosody and align it more closely with the content's emotional tone and the game's evolving context. All queries, replies, and comments are delivered in a conversational, but clear speaking style, whereas turn-taking fillers have fast speaking rate and low pitch. This distinction signals that the filler speech is self-directed, serving as a placeholder while the system prepares its response. In contrast, the systems using Matcha and base Tacotron TTS lack prosody control, making them less aligned to the game progression. Their speaking styles are closer to read than conversational. This way they are optimized for intelligibility, but with a conversational tone that exceeds systems trained solely on read speech.

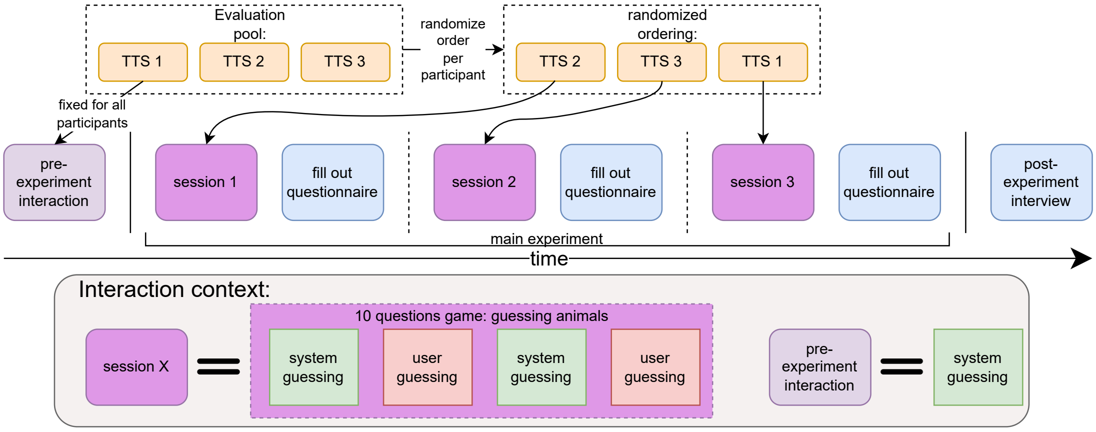

Figure 2: *Proposed experiment protocol: spoken dialogue interaction in guessing game context for TTS evaluation.*

# 4. Experiment

## 4.1. Interaction context: “20 questions” guessing game

We ground the interaction in a shared context. The domain we have chosen in this study is guessing games, where the participants play the classic guessing game “20 questions” with the system. In this game, there are two players, the questioner and the answerer. The answerer thinks about an object, and the questioner must guess the object by asking up to 20 yes-no questions to the answerer. In our setup, the questioner can ask up to 10 questions and only animals are guessed. Furthermore, since the speaker roles during a “20 questions” game are different: the questioner produces more speech while the answerer can simply answer “yes” or “no” without saying much else; we configured the system to play both roles to thoroughly assess capability of the TTS to generate speech in different contexts.

## 4.2. Experiment setup

As seen in Figure 2, the main experiment consists of 3 sessions. A selected TTS is used in the dialogue system for each session. During pilot studies, we observed that the user’s interaction pattern shifts over the course of the experiment potentially due to that the user adapts to interacting with the system. Thus, we randomize ordering of the 3 TTS models to counter this adaptation bias in the evaluation process. Each session consists of 4 games with the system as the questioner in round 1, 3 and the user as questioner in round 2, 4. We also installed a pre-experiment where the user play one game with the system to further decrease adaptation bias in the main experiment and to check for equipment malfunction.

## 4.3. Multi-dimensional interaction data collection

We want to evaluate the user’s experience holistically. This necessitates multi-dimensional data collection shown in Table 1. Firstly, we collect audio and video of the entire experiment process for analysis of the interaction itself, including aspects such as prosody of the user’s utterances, emotional state of the user at various points, and turn-taking delay. Secondly, we ask the user to rate their interaction with the system after each session. We chose the Godspeed questionnaire [10] commonly used in human-robot interaction studies. We eliminated 2 scales related to the embodiment and movement of the robot from the original questionnaire. Additionally, we added 3 questions to assess aspects specific to the voice inspired by MOS-X questionnaire [11] , a) Overall naturalness, b) Intelligibility, c) Prosody appropriateness. Both Godspeed and voice questionnaire items are rated on a discrete 5-point scale. Lastly, to understand the user’s preference over different systems, we conduct a loosely structured interview at the end of the experiment, where we ask the following questions: 1. *Do you think the three systems are same or different?* 2. *If you can choose one of the systems to play more games with, which one would you choose?*. We also ask the reasons behind their choices as well as any comments.

| **data type** | **level** | **perspective from user** |
|---|---|---|
| audio, video | utterance | third-person |
| questionnaire | session | self-reported |
| overall preference and comments | experiment | self-reported |

Table 1: *Multi-dimensional interaction data collection.*

# 5. Results

## 5.1. User preference

We recruited 21 self-reported English speakers for this study. The dialogue system worked reasonable well, where about a fifth of the interactions contained a few problems due to bad ASR, turntaking or GPT-4 responses. Overall, the users guessed the animal correctly half of the time, while the system did it every fourth. User-reported overall preference for the TTS systems is shown in Figure 3. A clear preference towards Tacotron2-context-adapted model is shown. After that, no preference is chosen by second most participants, while base Tacotron2 and Matcha-TTS received lowest preference counts. The majority of users who preferred Tacotron2-context-adapted reported that they did not perceive differences between the other two TTS systems, i.e. Tacotron2 and Matcha-TTS. This, added to the fact that both models received low preference, shows that the model architecture is not a determining factor for user preference. Without being prompted, participants used the following words to describe Tacotron2-context-adapted, “human-like” (n=5), “interactive” (n=2), “less robotic” (n=2).

## 5.2. Questionnaire

We hypothesized that both TTS and session are significant variables for questionnaire items. Both are proven wrong. Neither are significant factors in any of the questionnaire items by

| | Godspeed questionnaire | | | | | Voice questionnaire | | | User utterances | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | anthropomorphism | animacy | likeability | perceived intelligence | perceived safety | naturalness | intelligibility | prosody | pitch mean | pitch var | speech rate | turntaking delay | num words |
| C(TTS) | 0.285 | 0.870 | 0.146 | **0.024** | 0.843 | 0.308 | 0.972 | 0.407 | **0.030** | 0.703 | 0.218 | 0.067 | 0.266 |
| C(session) | 0.424 | 0.985 | 0.939 | 0.344 | 0.334 | 0.728 | 0.418 | 0.346 | 0.465 | **0.011** | 0.370 | **0.028** | 0.869 |
| C(user) | **0.000** | **0.006** | **0.000** | **0.000** | **0.044** | **0.019** | **0.038** | **0.005** | **0.000** | **0.000** | 0.656 | **0.000** | **0.000** |
| C(role) | - | - | - | - | - | - | - | - | 0.675 | 0.682 | 0.149 | **0.000** | **0.000** |

Table 2: *ANOVA analysis (Type II) p-values of the questionnaire and user utterance metrics. Significant p-values $<$ 0.05 are highlighted.*

ANOVA analysis (Type II) with p=0.05 on fitted ordinary linear squares as seen in Table 2, except for TTS in Godspeed questionnaire's "perceived intelligence" category. In that category, the fitted linear model indicates that Tacotron2 is 0.78 higher than Matcha-TTS with p=0.007 and Tacotron2-context-adapted is higher for 0.629 with p=0.026. On the other hand, the user factor is significant in all questionnaire items, indicating that users rate differently. This shows that questionnaire may not be an effective measurement instrument in our evaluation setup.

### 5.3. User utterance analysis

We analyzed the following features in user utterances: pitch mean, pitch variance, speech rate, turn-taking delay, and number of words spoken. As shown in Figure 2, ANOVA analysis (Type II) with p=0.05 reveals that TTS is only a significant factor in pitch mean of user utterances, with Tacotron2 12.14 Hz lower than others with p=0.010 while other differences are insignificant. This shows that TTS did not significantly alter interaction style of the users, thus it is difficult to effectively evaluate TTS from user utterance analysis in this experiment design. Moreover, session is a significant factor for 2 features, pitch mean and turn-taking delay, consistent with our hypothesis that users interact differently over the course of experiment, potentially due to that they become more accustomed at interacting with the system. User is again a highly significant factor while speaker role is also a moderately significant factor. Session and speaker role are part of experiment design. The fact that they significantly altered user utterance features shows the importance of experiment design when evaluating TTS in interaction.

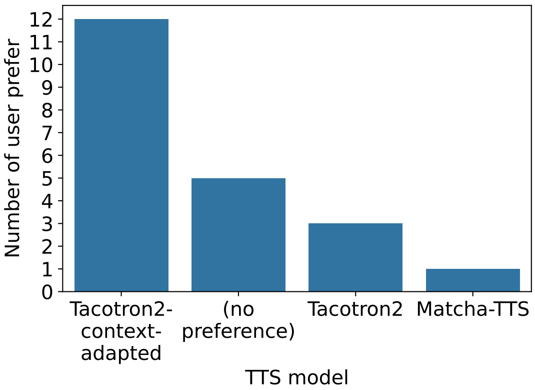

Figure 3: *Overall user preference for the same dialogue system with different TTS models.*

## 6. Discussion

Evaluating TTS within dialogue system interactions presents multiple challenges. One concern is the potential interference from non-TTS factors that can influence user impressions: domain, user preferences, understanding and reasoning capabilities of the system, and response times, etc. We observed that ASR and turn-taking failures significantly impacted interaction quality. Several users mentioned these issues as key factors influencing their ratings of the systems. However, we can minimize these noise factors by increasing dialogue system robustness through iterative development and testing, while incorporating latest advancement in spoken dialogue system technology, such as better turn-taking predictors [34]. Another potential limitation is scalability, primarily because these studies are conducted in-person with participants individually, in contrast to MOS listening tests which can be conducted in parallel through online crowdsourcing platforms. However, we contend that the quality of both the participant engagement and the data collected is substantially higher in the context of in-person interaction studies. The rich interaction recording in both audio and video can be further rated and analyzed by third-person observers to provide detailed understanding of the impact of different TTS models. Moreover, our analysis of user utterances indicates that speaker role within the guessing game context (questioner vs. answerer) significantly influenced the results, becoming yet another non-TTS noise factor. Future research could explore different interaction contexts, including shorter interactions and more balanced speaker roles. Another future direction is to analyse the prosody of the user, and increase the engagement and trust through prosodic entrainment [35].

## 7. Conclusion

We advocate for contextual and interactive evaluation of TTS models in dialogue system interaction. To this end, we built a fully autonomous dialogue system for the context of a "20 questions" guessing game and proposed a novel user study setup. The dialogue system is novel in that its interaction manager is able to control the TTS speaking style (from read to spontaneous) and prosodic realisation, given the content of its output (turn-taking filler, question/answers and entertaining remarks), and the current progression of the game. Three TTS models, Tacotron2, Matcha-TTS, and Tacotron2-context-adapted, were evaluated as part of the dialogue system. We collected multi-dimensional user data at utterance-, session- and experiment-level. Results show that Tacotron2-context-adapted is significantly preferred overall by users. However, there is little significance found in session-level questionnaire and user utterance-level prosody features, suggesting that these two measuring instruments are less effective in our evaluation setup. Future studies could explore other interaction contexts as well as other measuring instruments to increase evaluation efficacy.

# 8. Acknowledgement

This work is supported by Digital Futures project Advanced Adaptive Intelligent Systems (AAIS), and the Swedish Research Council projects Perception of speaker stance (VR-2020-02396) and Connected (2019-05003).

# 9. References

[1] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] S. Schöbel, A. Schmitt, D. Benner, M. Saqr, A. Janson, and J. M. Leimeister, "Charting the evolution and future of conversational agents: A research agenda along five waves and new frontiers," *Information Systems Frontiers*, pp. 1–26, 2023.

[3] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, É. Székely, C. Tånnander *et al.*, "Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program," in *Proc. SSW*, 2019.

[4] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors affecting the evaluation of synthetic speech in context," *Proc. SSW*, 2021.

[5] H. Lameris, J. Gustafsson, and É. Székely, "Beyond style: Synthesizing speech with pragmatic functions," in *Proc. Interspeech*, 2023, pp. 3382–3386.

[6] H. Lameris, A. Kirkland, J. Gustafson, and E. Székely, "Situating speech synthesis: Investigating contextual factors in the evaluation of conversational tts," in *12th Speech Synthesis Workshop (SSW) 2023*, 2023.

[7] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," in *Proc. SSW*, 2019.

[8] J. Edlund, C. Tånnander, and J. Gustafson, "Audience response system-based assessment for analysis-by-synthesis," in *Proc. ICPhS*, 2015.

[9] J. Mendelson and M. P. Aylett, "Beyond the listening test: An interactive approach to tts evaluation." in *Proc. Interspeech*, 2017, pp. 249–253.

[10] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *International Journal of Social Robotics*, vol. 1, pp. 71–81, 2009.

[11] J. R. Lewis, "Investigating mos-x ratings of synthetic and human voices," *Voice Interaction Design*, vol. 2, no. 1, p. 22, 2018.

[12] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language*, vol. 84, p. 101577, 2024.

[13] W. Zhang, C.-C. Yeh, W. Beckman, T. Raitio, R. Rasipuram, L. Golipour, and D. Winarsky, "Audiobook synthesis with long-form neural text-to-speech," in *Proc. SSW*, 2023.

[14] Y. Leng, Z. Guo, K. Shen, X. Tan, Z. Ju, Y. Liu, Y. Liu, D. Yang, L. Zhang, K. Song *et al.*, "PromptTTS 2: Describing and generating voices with text prompt," *arXiv preprint arXiv:2309.02285*, 2023.

[15] M. Lajszczak, G. C. Ruiz, Y. Li, F. Beyhan, A. van Korlaar, F. Yang, A. Joly, Álvaro Martín Cortinas, A. Abbas, A. Michalski, A. Moinet, S. Karlapati, E. Muszynska, H. Guo, B. Putrycz, S. L. Gambino, K. Yoo, E. Sokolova, and T. Drugman, "BASE TTS: Lessons from building a billion-parameter text-to-speech model on 100K hours of data," *arXiv*, 2024.

[16] Z. Borsos, M. Sharifi, D. Vincent, E. Kharitonov, N. Zeghidour, and M. Tagliasacchi, "Soundstorm: Efficient parallel audio generation," *arXiv preprint arXiv:2305.09636*, 2023.

[17] K. Georgila, A. W. Black, K. Sagae, and D. R. Traum, "Practical evaluation of human and synthesized speech for virtual human dialogue systems." in *Proc. LREC*, 2012, pp. 3519–3526.

[18] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "RyanSpeech: A Corpus for Conversational Text-to-Speech Synthesis," in *Proc. Interspeech 2021*, 2021, pp. 2751–2755.

[19] K. Lee, K. Park, and D. Kim, "Dailytalk: Spoken dialogue dataset for conversational text-to-speech," in *Proc. ICASSP*, 2023, pp. 1–5.

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

[21] K. Mitsui, T. Zhao, K. Sawada, Y. Hono, Y. Nankaku, and K. Tokuda, "End-to-end text-to-speech based on latent representation of speaking styles using spontaneous dialogue," in *Proc. Interspeech*, 2022.

[22] 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*, 2022.

[23] E. Ekstedt, S. Wang, É. Székely, J. Gustafson, and G. Skantze, "Automatic evaluation of turn-taking cues in conversational speech synthesis," in *Proc. Interspeech*, 2023.

[24] J. O'Mahony, C. Lai, and S. King, "Synthesising turn-taking cues using natural conversational data," in *Proc. SSW*, 2023.

[25] K. Fischer, O. Niebuhr, L. C. Jensen, and L. Bodenhagen, "Speech melody matters—how robots profit from using charismatic speech," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–21, 2019.

[26] M. Cohn, C.-Y. Chen, and Z. Yu, "A large-scale user study of an alexa prize chatbot: Effect of TTS dynamism on perceived quality of social dialog," in *Proc. SIGdial*, 2019, pp. 293–306.

[27] M. P. Aylett, B. R. Cowan, and L. Clark, "Siri, echo and performance: You have to suffer darling," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–10.

[28] D. Kontogiorgos, V. Avramova, S. Alexanderson, P. Jonell, C. Oertel, J. Beskow, G. Skantze, and J. Gustafson, "A multimodal corpus for mutual gaze and joint attention in multiparty situated interaction," in *Proc. LREC*, 2018, pp. 119–127.

[29] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*, 2018, pp. 4779–4783.

[30] É. Székely, S. Wang, and J. Gustafson, "So-to-speak: an exploratory platform for investigating the interplay between style and prosody in tts," in *Proc. Interspeech*, 2023, pp. 2016–2017.

[31] S. Mehta, R. Tu, J. Beskow, É. Székely, and G. Eje Henter, "Matcha-TTS: A fast TTS architecture with conditional flow matching," in *Proc. ICASSP*, 2024.

[32] 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.

[33] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[34] E. Ekstedt and G. Skantze, "Voice activity projection: Self-supervised learning of turn-taking events," *INTERSPEECH 2022*, pp. 5190–5194, 2022.

[35] Š. Beňuš, M. Trnka, E. Kuric, L. Marták, A. Gravano, J. Hirschberg, and R. Levitan, "Prosodic entrainment and trust in human-computer interaction."