INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Beyond Style: Synthesizing Speech with Pragmatic Functions

Harm Lameris, Joakim Gustafson, Éva Székely

Division of Speech, Music & Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
{lameris, jkgu, szekely}@kth.se

Abstract

With recent advances in generative modelling, conversational systems are becoming more lifelike and capable of long, nuanced interactions. Text-to-Speech (TTS) is being tested in territories requiring natural-sounding speech that can mimic the complexities of human conversation. Hyper-realistic speech generation has been achieved, but a gap remains between the verbal behavior required for upscaled conversation, such as paralinguistic information and pragmatic functions, and comprehension of the acoustic prosodic correlates underlying these. Without this knowledge, reproducing these functions in speech has little value. We use prosodic correlates including spectral peaks, spectral tilt, and creak percentage for speech synthesis with the pragmatic functions of small talk, self-directed speech, advice, and instructions. We perform a MOS evaluation, and a suitability experiment in which our system outperforms a read-speech and conversational baseline.

Index Terms: speech synthesis, pragmatic functions, conversational TTS

1. Introduction

Voice assistants have become ubiquitous in recent years, providing users with a convenient way to interact with technology using spoken language. Typically, these assistants use text-to-speech (TTS) voices with neutral and warm speaking styles that are easy to comprehend and suitable for command-based interactions. However, as conversational AI assistants and social robots become more prevalent in social and guiding scenarios, there is a growing need for these systems to display their attitude towards what they are saying through a range of speaking styles and prosodic realisations [1]. It is essential that the TTS systems used for these kinds of applications allow for an extended pragmatic repertoire, in order to effectively convey the intended meaning and provide a more natural and engaging interaction for the user [2]. Prosodic realization has been found to be a key factor in the interpretation of the pragmatic implications of a phrase [3]. In conversational settings, it is also important to generate appropriate turn-taking cues [4, 5] and indicate the speaker's level of certainty [6]. Different pragmatic functions require different pronunciations, prosodic realizations, and voice qualities. Instructions and advice need to be delivered with perfect enunciation and pronunciation in a clear voice, while small talk is characterized by more informal, colloquial tone with a more expressive and varied prosody, and self-directed speech is characterized by a less varied prosody and a softer voice quality [7].

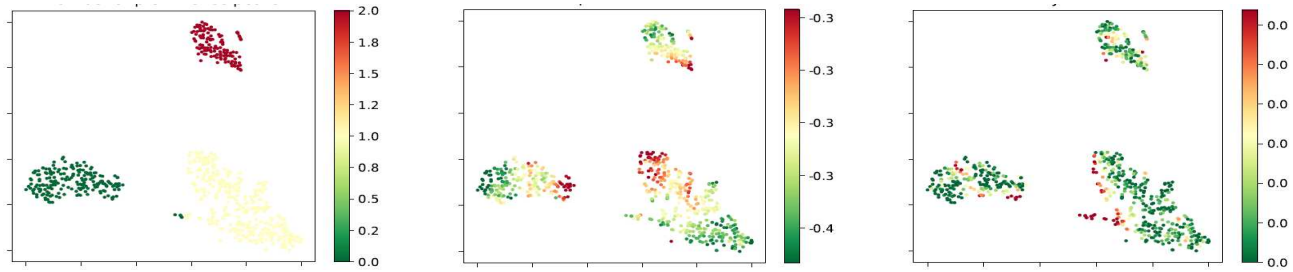
There are several neural TTS systems available that provide different ways to control the speaking style and prosodic realization. Some of these systems allow for control of speaking rate [8, 9], pitch [10] or both [11, 12]. Some systems also allow for control of the voice quality [13], and insertion of hesitations [14, 15, 16]. One way of controlling the prosodic realisation in neural TTS is to train it on a large corpus that contains a varied manner of speaking and then automatically detect a given number of speaking styles (Global Style Tokens) [17]. By listening to the generated speech, the style tokens could then be categorized according to expressive speech acts they seem to convey. Style tokens have also been used for emotional TTS trained on a corpus of acted emotions [18]. Mellotron combines GSTs and explicit f0 values to guide a Tacotron 2 decoder in order to control the prosodic realization [19]. There are also systems that combine GSTs with speaker embeddings in multi-speaker TTS [20]. A potential problem with global style tokens is that it is not obvious whether it is possible to extend these from book reading or acted emotions to the kinds of communicative functions needed in real interactions [2].

There have been some notable efforts in recording conversational TTS corpora where actors read scripts from chat-bots and task-oriented dialogues [21, 22]. Specially recorded conversational data has also been combined with general purpose data in multi-speaker TTS systems. In [23] a female voice actor read utterances that were tagged for 10 different speech acts, e.g. greeting, instruction, surprise and uncertainty. Given text and speech act as input, they were able to transplant the prosody acquired from training on the conversational speech on to two general-purpose TTS voices. An issue with using acted conversational corpora is that it is hard, even for a voice actor, to sound spontaneous with believable emotions when reading a dialogue script. It has been shown that spontaneous conversational speech is more varied than scripted conversational speech in terms of variation in pitch and speaking rate [13]. A potential solution is to train the TTS system using a dataset of many speakers speaking spontaneously in various contexts, and then control the generated speaking style by providing a reference audio alongside the input text [24, 25]. However, this approach has limitations as only a single example of the desired speaking style is provided, and matching a situation with the appropriate reference file is challenging. Additionally, modelling long dependency-based prosodic differences in speaking styles becomes difficult using this method.

In this paper, we propose that conversational neural TTS systems should be trained on ecologically valid speech corpora that have been specifically recorded for purpose of developing the target dialogue application. This is in line with [27], which

This research was supported by the Swedish Research Council projects Connected (VR-2019-05003), Perception of speaker stance (VR-2020-02396), the Riksbankens Jubileumsfond project CAPTivating (P20-0298).

10.21437/Interspeech.2023-2072



(a) *Prominence peaks* (b) *Spectral tilt* (c) *Creak percentage*

Figure 1: *The Starmap [26] t-SNE plot for prosodic features*

suggests that developers of conversational systems should ensure that the voices they use are trained on speech data that contain the type of verbal expression they want their conversational agents to be able to deliver.

We visualized the utterances in the conversational TTS corpus based on prosodic and voice quality features using a human-in-the-loop annotation tool [26], and annotated a subset of the data for its pragmatic function. In order to identify pragmatic speaking styles beyond pitch and speaking rate we used five speech features: mean energy, number of prominence peaks, speech rate, spectral tilt and creakiness. Using this human-in-the-loop method we selected four pragmatic functions that were both important for the intended use case, and that actually sounded different: small talk, instructions, advice and self-directed speech. We finetuned a speech synthesis model trained on read and spontaneous speech with separate embeddings for each pragmatic function. In a perceptual MOS evaluation the finetuned system performed similarly to a read-speech and spontaneous baseline. Additionally we conducted a suitability listening experiment in which listeners were asked to rate suitability of the speaking style to the content and provided context, for which the finetuned system outperformed both baselines. Samples may be found at `speech.kth.se/tts-demos/beyond_style/`

2. Method

2.1. Data

The data was obtained from a publicly available multimodal corpus of 15 interactions between a human moderator and two users that were given the task of decorating an apartment using a GUI on a large touch screen [28]. In all recordings the same person acted as mediator, order to use the multimodal interactional data to develop a social robot that could be used as a moderator in similar collaborative tasks. The moderator was provided with a general outline of the topics to cover in each interaction, but was not given specific instructions on what to say. This allowed the moderator to engage in spontaneous yet pre-planned extemporaneous conversations during the interactions.

During the first phase, the moderator engaged in *small talk* with participants about living situations with roommates, discussing topics such as tidiness and conflict. In the second phase, the moderator *instructed* the participants about the setup of the experiment, in which the participants collaborate in designing a living space where they would hypothetically cohabit for three months while being recorded for a reality television series. During the third phase, the moderator plays the role of interior decorator who *advised* the participants on their design choices. In the final phase, the moderator comments on the participants' final choices as engages in debriefing. Lastly, the corpus also contains *self-directed speech* occurring as the moderator was contemplating different design options or commenting on how the users progressed while moving an object he had suggested them to select next. As expected, the mediator performed different pragmatic functions in each phase, and these functions were expressed using a variety of speaking styles.

Figure 2: *Data collection photo, courtesy of the authors [28]*

We extracted the speech data from the moderator, a male speaker of General American English, which were automatically segmented into breath groups of 1 to 10 seconds. The corpus was initially transcribed using ASR and subsequently manually corrected. The final orthographic transcription includes tokens for filled pauses, semi-colons for audible breaths, commas for turn-internal pauses and full stops or question marks at turn-endings. By adding these in the text input to the TTS system dialogue system, designers get explicit control of the manner of speaking of their conversational agents. These spontaneous speech data were supplemented with read-speech audio of the mediator reading 1129 the CMU Arctic sentences [29] and 1132 sentences from online news paper texts. The total TTS corpus has a duration of approximately 8 hours (2h 26min of reading and 5h 40min of spontaneous speech). The data will be made publicly available in the future once privacy concerns have been addressed.

2.2. Data annotation

A subset of 490 utterances of the corpus was selected based on length (4-10 seconds) and subsequently annotated by the first author for their pragmatic function using Starmap, a human-in-the-loop annotation tool [26] that uses t-SNE dimensionality reduction [30] on utterance-level prosodic features to aid in the navigation of corpora. We selected five features on Starmap that were hypothesized to aid in the distinction of the pragmatic functions. The *mean energy*, the *number of prominence peaks*, and the estimated *speech rate* were extracted using Continu-

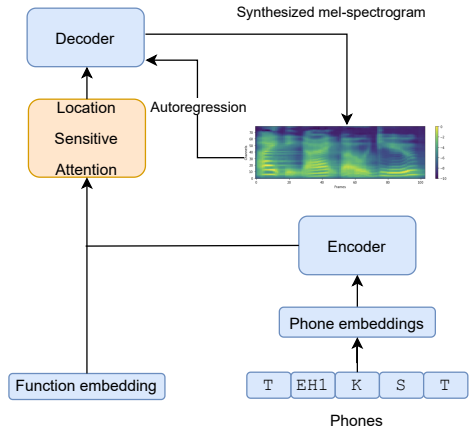


Figure 3: *The model architecture.*

ous Wavelet Transformation-based hierarchical prosody representation as described in [31]. The prominence peaks were extracted using the high-level hierarchical scales as in [26, 31].We supplemented these with per-utterance *spectral tilt* over voiced segments and *creak percentage*. The spectral tilt was calculated using the Python package parselmouth [32]. For the creak percentage, the duration of creaky voice per file was extracted using DeepFry [33], and the creak percentage was calculated with the following formula:

$$\text{creak percentage} = \frac{\text{total creak duration}}{\text{total duration}} \quad (1)$$

The t-SNE dimensionality reduction [30], as implemented in Starmap, was used to create an overview of the prosodic features in order to enable selection of utterances based on the pragmatic function. The t-SNE plot can be found in Figure 1. We used the prompt selection tool in Starmap to annotate 490 utterances for their pragmatic function based on both the prosodic features and semantic content. The pragmatic functions used for annotations are: small talk (99 utterances), instructions (130 utterances), advice/guidance (163 utterances), and self-directed speech during decision making (98 utterances).

2.3. Model architecture

We used a modified PyTorch implementation of the Tacotron2[1] [34] to which we added pragmatic style conditioning. The pragmatic style is identified in the model through an 8-dimensional speaker-like embedding, set up after [19]. This embedding is appended to each utterance's encoded text and passed to the attention and decoder blocks from the model, increasing the number of parameters to 28.26M from 28.19M in the base implementation. When a model is transfer learning from a model with a different embedding, the embedding is reinitialized, irrespective of whether the number of embeddings is identical. A model is first trained on the whole corpus with two embeddings, indicating whether the utterance is from the read or spontaneous part of the corpus. This model is trained for 70k iterations on 4 NVIDIA GeForce RTX 3090 12 GB GPUs with batch size 28 and with 5% of the data withheld as validation set. We used a HiFi-GAN [35] vocoder fine-tuned on the same corpus for 383k iterations on the top of the published model.[2]. At inference, denoiser strength is set at 0.04.

[1] https://github.com/NVIDIA/tacotron2
[2] https://github.com/jik876/hifi-gan

A model is then fine-tuned on the part of the training corpus where a pragmatic function was identified (439 utterances). The embedding from the original model is dropped and a new one is initialized with four embeddings representing the four pragmatic functions. In order to realize the pragmatic functions, the model is then trained for a further 4000 iterations, with checkpoints saved every 500 iterations. In the process of changing the embedding, some of the speech quality is lost, and the stopping point is chosen through informal listening tests where speech quality sufficiently recovered, yet the pragmatic function can still be effectively generated. At inference the pragmatic styles can be generated by putting additional weight (2.5×) on a particular embedding compared to in the training, while setting the weight on the other embeddings to 0.

3. Experiments

We conducted two listening experiments, a Mean Opinion Score (MOS) subjective listening evaluation and a suitability listening test in which we compared the pragmatic function setup to two baselines: a read-speech baseline trained, which was trained on the complete corpus for 74k iterations, and that was conditioned with the read-speech embedding at synthesis, and a spontaneous baseline which was trained identically to the read-speech baseline and conditioned on the spontaneous-speech embedding at synthesis.

In order to evaluate the quality of the synthesized speech, a subjective listening experiment was conducted using a MOS quality assessment for each of the baseline models and the pragmatic synthesis. As this evaluation was designed with the sole purpose of comparing speech quality across these systems, we synthesized in-domain sentences that were representative of the respective speech contexts. Specifically, the synthesized sentences for the read-speech baseline were derived from popular science style prose, whereas those for the two spontaneous setups were derived from conversational language. For the read-speech baseline 40 sentences were synthesized that contained animal facts, while for the spontaneous baseline and the pragmatic function setup we synthesized 48 sentences that were evenly divided over the four pragmatic functions.

In addition to the MOS evaluation, we conducted a suitability listening experiment. For this evaluation, we presented the participants one stimulus each of the pragmatic function setup and the two baselines. Participants were asked to rate *How well does the speaking style match the content and context?* on a scale from *1. Very poorly* to *5. Very well*.

The following context was provided for the respective pragmatic functions:

1. Someone is mumbling something to themselves.
2. Someone is giving advice to you.
3. Someone is chatting to you.
4. Someone is giving you instructions.

All contexts matched the intended speaking style for the pragmatic function setup. Participants were presented with a total of 32 stimuli evenly divided over each pragmatic function with the semantic content matching the function.

4. Results

4.1. Subjective MOS evaluation

For the MOS subjective listening evaluation, we recruited 3 groups of 20 native English speakers from the US who were balanced for self-identified gender. Participants were presented

Table 1: *Mean Opinion Scores for each setup*

System	MOS
Pragmatic	3.73 ± 1.23
Spontaneous	3.60 ± 1.04
Read Speech	3.51 ± 1.03

with one stimulus per page. The stimuli were loudness normalized at -18.0 LUFS. The results of the subjective listening test can be found in Table 1. A one-way ANOVA showed no differences between the pragmatic function synthesis and the read-speech and spontaneous baseline. The MOS for the individual pragmatic functions were 3.23 for self-directed speech, 3.88 for advice, 3.93 for small talk, and 3.89 for instruction.

4.2. Suitability listening test

Table 2: *The means and standard deviations for the suitability test. Italics indicate improvement over read speech, bold indicates significant improvement over read- and spontaneous speech*

System	Self-directed	Advice	Small Talk	Instructions
Pragm.	*3.47*±1.09	**3.89**±1.06	**4.18**±0.88	**3.87**±0.99
Spont.	*3.31*±1.01	3.50±1.06	*3.86*±0.92	3.45±1.09
Read	2.28±1.21	3.25±1.15	3.08±1.19	3.20±1.14

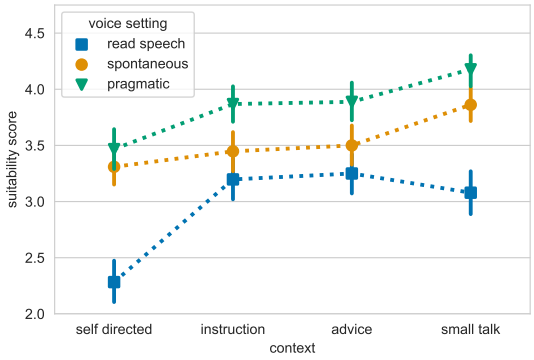


Figure 4: *Suitability test results*

We recruited 20 native English speakers on Prolific [3] who were balanced for self-identified gender from any majority English-speaking country or territory. The results of the suitability listening test can be found in Table 2 and Figure 4. The synthesis according to the pragmatic function was found significantly more suitable than the read speech for each of the pragmatic functions ($p < 0.0001$), and was rated as more suitable than the spontaneous speech for the pragmatic functions of advice ($p = 0.006$), small talk ($p = 0.02$), and instructions ($p = 0.002$) on a one-way ANOVA with post-hoc Tukey. Additionally, spontaneous self-directed speech and small talk significantly outperform read-speech synthesis.

5. Discussion

In our study, we followed the suggestions of [36] and evaluated both the quality and appropriateness of synthesized speech samples based on their pragmatic function in mediated interactions. Our results showed discrepancies between MOS and suitability listening tests, indicating the potential of in-context evaluation as an improvement to standard MOS evaluation, especially for expressive and spontaneous synthesis. While read-speech often receives higher quality ratings in MOS evaluations due to optimal recording conditions, most applications of TTS require speech that goes beyond high quality and necessitate appropriate speech and dialogue behaviors not obtainable from read-speech.

We propose that modelling prosody at the pragmatic function level is meaningful, as it captures long dependency-based prosodic differences and enables synthesis with limited labelled in-domain data. Our work diverges from the style-based focus, prosodic modification, and voice conversion commonly seen in speech synthesis. Unlike style-based approaches like GST [17], we provide a more concrete designation of pragmatic functions that can be integrated into applications requiring specific communicative functions. Moreover, pragmatic functions can be used when there is insufficient data for GST. Compared to prosodic control, pragmatic function synthesis offers benefits by addressing the challenge of implementing prosodic control throughout an entire user-based interaction, which would also apply to voice conversion.

Additionally, we aim to enhance our understanding of the relationship between prosodic correlates of voice quality and speech variation specific to dialogue acts. To achieve this, we manually annotated a subset of sentences from the corpus using the human-in-the-loop tool Starmap [26]. This tool, employing prosodic correlate measurements, Continuous Wavelet Transformations, and dimensionality reduction, helps identify the prosodic aspects that distinguish speech typical for each function: small talk, instruction, advice, and self-directed speech.

One limitation of our current implementation is the heuristic selection of the checkpoint for pragmatic function synthesis. To address this, an automated checkpoint selection method proposed by [23] could be employed, which objectively assesses a model checkpoint using a text-audio pair from a held-out set and evaluates the synthesized intonation pattern. Furthermore, improvements in this area could be achieved through additional annotated data, which would allow training with the pragmatic function embedding without requiring fine-tuning.

6. Conclusion

In this paper, we propose a spontaneous TTS system which leverages data that was recorded for the development of a robot guide application. A human-in-the-loop tool was used annotate a subset of the corpus for the pragmatic function of the speech, based on prosodic and voice quality features. The data were supplemented with read speech for a base voice that was fine-tuned with separate embeddings for pragmatic functions. The proposed system was evaluated using MOS, in which it was rated comparable to read-speech and spontaneous baselines, and a suitability evaluation, the results of which demonstrate that the system outperforms the baselines in terms of suitability of the speaking style to the content and provided context.

Our results demonstrate the importance of using appropriate speech data to train conversational systems to ensure that they can deliver the desired verbal expression, and the shortcomings of MOS when using speech synthesis in applied settings. We emphasize that this paper is a preliminary undertaking that lays the groundwork for further in-context evaluation, which we hope will see increased usage in TTS evaluations.

[3] https://www.prolific.co

7. References

[1] N. G. Ward, *Prosodic patterns in English conversation*. Cambridge University Press, 2019.

[2] M. Marge, C. Espy-Wilson, N. G. Ward, A. Alwan, Y. Artzi, M. Bansal, G. Blankenship, J. Chai, H. Daumé III, D. Dey *et al.*, "Spoken language interaction with robots: Recommendations for future research," *Computer Speech & Language*, vol. 71, p. 101255, 2022.

[3] S. Herment and L. Leonarduzzi, "The pragmatic functions of prosody in English cleft sentences," in *Speech Prosody 2012*, 2012.

[4] M. P. Aylett, A. Carmantini, and D. A. Braude, "Why is my social robot so slow? How a conversational listener can revolutionize turn-taking," in *Proc. IROS*.

[5] Y. Li and C. Lai, "Robotic speech synthesis: Perspectives on interactions, scenarios, and ethics," in *Robo-Identity: Exploring Artificial Identity and Emotion via Speech Interactions*, 2022.

[6] A. Kirkland, H. Lameris, E. Szekely, and J. Gustafson, "Where's the uh, hesitation? The interplay between filled pause location, speech rate and fundamental frequency in perception of confidence," in *Proc. Interspeech 2022*, 2022, pp. 4990–4994.

[7] N. Campbell and P. Mokhtari, "Voice quality: the 4th prosodic dimension," in *Proc. ICPhS*, 2003, pp. 2417–2420.

[8] J. Park, K. Han, Y. Jeong, and S. W. Lee, "Phonemic-level duration control using attention alignment for natural speech synthesis," in *Proc. ICASSP*, 2019, pp. 5896–5900.

[9] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *Advances in neural information processing systems*, vol. 32, 2019.

[10] A. Łańcucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*, 2021, pp. 6588–6592.

[11] T. Raitio, J. Li, and S. Seshadri, "Hierarchical prosody modeling and control in non-autoregressive parallel neural TTS," in *Proc. ICASSP*, 2022, pp. 7587–7591.

[12] N. Ellinas, M. Christidou, A. Vioni, J. S. Sung, A. Chalamandaris, P. Tsiakoulis, and P. Mastorocostas, "Controllable speech synthesis by learning discrete phoneme-level prosodic representations," *Speech Communication*, vol. 146, pp. 22–31, 2023.

[13] H. Lameris, S. Mehta, G. Henter, J. Gustafson, and É. Székely, "Prosody-controllable spontaneous TTS with neural HMMs," in *Proc. ICASSP*, 2023.

[14] É. Székely, G.-E. Henter, J. Beskow, and J. Gustafson, "How to train your fillers: uh and um in spontaneous speech synthesis," in *Proc. SSW*, pp. 245–250.

[15] H. Guo, S. Zhang, F. K. Soong, L. He, and L. Xie, "Conversational end-to-end tts for voice agents," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 403–409.

[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] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 5180–5189.

[18] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *Proc. ICASSP*, 2020, pp. 7254–7258.

[19] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proc. ICASSP*, 2020, pp. 6189–6193.

[20] L.-W. Chen and A. Rudnicky, "Fine-grained style control in transformer-based text-to-speech synthesis," in *Proc. ICASSP*, 2022, pp. 7907–7911.

[21] R. Zandie, M. H. Mahoor, J. Madsen, and E. S. Emamian, "Ryanspeech: A corpus for conversational text-to-speech synthesis," in *Proc. Interspeech*, 2021, pp. 2751–2755.

[22] K. Lee, K. Park, and D. Kim, "Dailytalk: Spoken dialogue dataset for conversational text-to-speech," *Proc. ICASSP*, 2023.

[23] R. Fernandez, D. Haws, G. Lorberbom, S. Shechtman, and A. Sorin, "Transplantation of conversational speaking style with interjections in sequence-to-sequence speech synthesis," in *Proc. of Interspeech*, 2022, pp. 5488–5492.

[24] L.-W. Chen, S. Watanabe, and A. Rudnicky, "A vector quantized approach for text to speech synthesis on real-world spontaneous speech," *arXiv preprint arXiv:2302.04215*, 2023.

[25] Y. A. Li, C. Han, and N. Mesgarani, "StyleTTS: A style-based generative model for natural and diverse text-to-speech synthesis," *arXiv preprint arXiv:2205.15439*, 2022.

[26] É. Székely, J. Edlund, and J. Gustafson, "Augmented prompt selection for evaluation of spontaneous speech synthesis," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 6368–6374.

[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. Kominek and A. W. Black, "The CMU arctic speech databases," in *Proc. SSW*, 2004, pp. 223–224.

[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[31] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using Continuous Wavelet Transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[32] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[33] B. Chernyak, T. Simon, Y. Segal, J. Steffman, E. Chodroff, J. Cole, and J. Keshet, "DeepFry: Identifying vocal fry using deep neural networks," in *Proc. Interspeech*, 2022, pp. 3578–3582.

[34] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen *et al.*, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *Proc. ICASSP*, pp. 4779–4783.

[35] 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.

[36] 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, pp. 105–110.