DiVA
http://www.diva-portal.org

This is the published version of a paper presented at *Speech Synthesis Workshop (SSW11), Budapest, Hungary August 26-28, 2021.*

Citation for the original published paper:

Kirkland, A., Włodarczak, M., Gustafsson, J., Székely, É. (2021)
Perception of smiling voice in spontaneous speech synthesis
In: *Proceedings of Speech Synthesis Workshop (SSW11)* (pp. 108-112).
https://doi.org/10.21437/SSW.2021-19

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-329143

11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Perception of smiling voice in spontaneous speech synthesis

Ambika Kirkland[1], *Marcin Włodarczak*[2], *Joakim Gustafson*[1], *Éva Székely*[1]

[1]Division of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden
[2]Department of Linguistics, Stockholm University, Sweden

kirkland@kth.se, wlodarczak@ling.su.se, jkgu@kth.se, szekely@kth.se

Abstract

Smiling during speech production has been shown to result in perceptible acoustic differences compared to non-smiling speech. However, there is a scarcity of research on the perception of "smiling voice" in synthesized spontaneous speech. In this study, we used a sequence-to-sequence neural text-to-speech system built on conversational data to produce utterances with the characteristics of spontaneous speech. Segments of speech following laughter, and the same utterances not preceded by laughter, were compared in a perceptual experiment after removing laughter and/or breaths from the beginning of the utterance to determine whether participants perceive the utterances preceded by laughter as sounding as if they were produced while smiling. The results showed that participants identified the post-laughter speech as smiling at a rate significantly greater than chance. Furthermore, the effect of content (positive/neutral/negative) was investigated. These results show that laughter, a spontaneous, non-elicited phenomenon in our model's training data, can be used to synthesize expressive speech with the perceptual characteristics of smiling.

Index Terms: speech synthesis, text-to-speech, smiling voice, smiled speech

1. Introduction

There are many well-documented functions of smiling in interpersonal communication. A smile can influence a speaker's perceived desire for cooperation[1] as well as their perceived trustworthiness [2], competence [3], extroversion, sympathy, kindness, and attractiveness [4]. And smiling is not merely a visual phenomenon–it creates changes in speech that can be perceived by listeners. The features associated with smiling voice include greater pitch height and pitch range [5, 6], and higher formant frequencies for some vowels [7, 8]. These audible characteristics allow smiling voice to mirror at least some of the social functions of smiling even in the absence of visual cues (e.g., conveying trustworthiness in virtual agents [9]).

With the advancements of conversational AI allowing for more nuanced interactions than ever before [10], synthetic voices of conversational agents need to become more realistic and versatile, displaying character, and complex conversational capabilities. In the area of expressive speech synthesis, there has been a relatively recent shift of research interest from synthesizing speech reflecting specific emotion categories or dimensions, towards unsupervised approaches of synthesizing "speaking styles" [11, 12]. These data-driven approaches have benefited from the availability of audiobooks, which contain a higher variability of speaking styles than traditional TTS corpora, but are lacking explicit annotation found in corpora of specifically recorded emotional speech. Less attention has been given to the interpretability and perceptual effect of synthesized styles using unsupervised methods. On the other end of the spectrum, various systems have been proposed for interpretable and intuitive control of prosodic features such as melody, rhythm, [13], pitch range, phone duration and spectral tilt [14]. However, for more stylistic characteristics, the gap between controllability and interpretability still remains to be closed. Thanks to recent advances in deep learning which have resulted in more robust systems both in text-to-speech and in speech processing tools for annotation and segmentation, spontaneous speech synthesis has made a leap forward in terms of naturalness and appropriateness for certain contexts [15]. As corpora of spontaneous speech have become available targets for text-to-speech, we are no longer restricted to modeling speaking styles in audiobooks, which are mostly a result of colorful reading, such as the speaker imitating characters. Real-world spontaneous speech data contains a myriad of speech phenomena that reveal the speaker's cognitive state, attitude stance, etc., which are represented in a variety of acoustic-prosodic and segmental features. Much of the research in spontaneous speech synthesis to date has been focused on modeling and understanding the use of hesitations such as *uh* and *um* [16, 17] and breathing [18], with many styles and phenomena left to be explored, both in terms of synthesis and perception.

This paper focuses on synthesizing a specific voice style, namely amused speech following laughter in a spontaneous monologue, which we refer to here as "smiling voice". We propose a context-driven method for synthesizing speech following laughter, using state-of-the art neural TTS built entirely from spontaneous conversational speech. In the training data, laughter (short affect burst) is not explicitly elicited, emerging as part of the spontaneous delivery contributing to the narrative. The perceptual effect of smiling voice is explored in different contexts, using sentences with positive, negative and neutral sentiment.

2. Related work

While much of the research on smiling voice has involved naturally produced speech, there have been a few investigations of smiling voice in synthesized speech. Lasarcyk and Trouvain [19], for example, synthesized four different German vowels using articulatory synthesis, and applied combinations of three different parameters to these vowels which correspond to the effects of smiling on articulation: raised f0, spread lips, and raised larynx. They found that higher f0 resulted in a greater degree of perceived smiling for all vowels. Both spread lips and a raised larynx influenced vowel formant frequencies as well as the perception of smiling, but this effect was different for different vowels. The vowels /a:/ and /y:/ were perceived as more smiley when synthesized with spread lips, while /i:/ showed no difference and /u:/ was considered less smiley with spread lips. The vowels /a:/ and /i:/ were perceived as more smiley with a raised larynx, but this parameter had no effect on the other vow-

10.21437/SSW.2021-19

els.

Another approach [20] used HMM-based synthesis to allow for a controllable degree of smiling in synthesized speech. Two models were created using recordings of neutral and smiling speech from one actor. For the recordings of smiling speech, the actor was instructed to smile and "sound happy" but not to laugh. A new model with controllable degrees of smiling was created by using a weighted-sum interpolation between the neutral and smiling models, with a degree of smile that varied according to the weights used. The evaluation showed that higher weights resulted in synthesized speech that was perceived as smiling to a greater degree, but also less natural.

In terms of synthesizing amused or happy-sounding speech, generating laughter is another important issue. Some previous approaches attempting to combine laughter and smiling voice have synthesized these two components independently from one another and then combined them. An HMM-based approach in [21], for example, inserted vowels produced while laughing into smiling speech generated with a different method, and the approach of [22] inserted phrase-sized "affect bursts" using concatenative speech synthesis. A more recent effort to synthesize laughter [23] employed a sequence-to-sequence neural text-to-speech system, with the goal was to create natural-sounding laughter which could then be integrated with a model for smiling speech.

In contrast to the approaches described above, our method of producing smiling voice neither explicitly manipulates acoustic parameters, nor does it use data that was explicitly elicited while smiling. Rather, we employ a context-driven approach on spontaneous data, generating smiling voice by synthesizing speech following laughter in one integrated model.

3. Database and synthesis

3.1. Spontaneous speech corpus

The TTS corpus was created from the audio recordings of the Trinity Speech-Gesture Dataset (TSGD) [24], which is comprised of 25 impromptu monologues by a male actor, on average 10.6 minutes long. The recordings were performed over multiple recording sessions by a male speaker of Irish English. The actor is speaking in a colloquial style, spontaneously and without interruption on topics such as hobbies, daily activities, and interests. During the monologues, he addresses a person seated behind the cameras who is giving visual, but no verbal feedback. Because a large part of the monologues involve story-telling, the actor often engages in retelling entertaining anecdotes, which naturally elicit laughter followed by the impression of amused, smiling voice, the synthesis of which is the focus of the current paper.

3.2. Annotation

To create a TTS corpus, the recording was transcribed using ASR and subsequently manually corrected to contain as few errors as possible, and to ensure that all filler words are accurately transcribed. In order to maximize the utterance length in the corpora and to enable insertion of inhalation breaths in the TTS, we used a data augmentation method called *breathgroup bigrams*, which essentially consists of segmenting a speech corpus into stretches of speech delineated by breath events, and then combining these breath groups in an overlapping fashion to form utterances no longer than 11 seconds [18] (see Figure 1). This method also makes it possible to learn contextual information beyond respiratory cycles during TTS training. Aside

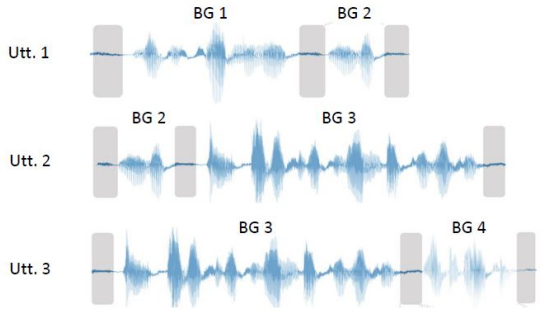

Figure 1: *Illustration of the breathgroup-bigram utterance structure [18] applied to create the TTS corpus from continuous recordings of spontaneous speech. Breath events are highlighted in grey.*

from filled pauses such as *uh* and *um*, the ASR transcription was enhanced with manual annotation of laughter, style breaks and silent pauses, the latter indicated with a comma. Both the filled pauses and laughter were transcribed using ARPABET phones. No new characters were introduced outside the standard. If a laughter involved ingressive airstream and was directly followed by more speech, the last voiced inhalation was annotated as breath event.

3.3. Systems

Two systems were trained using the sequence-to-sequence neural TTS engine Tacotron 2 [25]. The first system uses the standard Tacotron 2 architecture. The second system implements an utterance-level prosody control method, similar to [14], to be able to direct f0 and speech rate at inference. Speech rate (syllables/second) over the utterance and mean f0 are normalised, aligning the 1st and the 99th percentile points of the data to -1 and 1 respectively, and allowing outliers to go outside of that range. Normalized values for both features are appended to each utterance's encoded text and passed to the attention and decoder blocks from the pre-trained model. In order to fit the additional features, the input dimension to the attention, LSTM, projection and gate layers in the decoder are expanded. The additional weights added to the model are initialized with zero values. As such, at the start of the training the model evaluates as the pre-trained model. This method allows for directing mean f0 and speech rate on utterance level based on the natural distribution of these features in the corpus, as opposed to direct manipulation.

We used a PyTorch implementation of Tacotron 2[1], training each voice using transfer learning for 200k iterations on top of a pre-trained model trained on the LJ speech corpus [26]. Transfer learning based on a model trained on a large read-speech corpus has been shown to improve the quality of spontaneous speech synthesis [15]. For vocoding, the pre-trained universal model of WaveGlow [27] was fine-tuned for 290k iterations.

3.4. Synthesis of smiling voice

Our hypothesis is that due to the natural occurrence of laughter in the spontaneous speech corpus, synthesizing a laughter token followed by a breath event will result in an amused speaking style, characteristic of smiling voice in the subsequent speech.

[1] https://github.com/NVIDIA/tacotron2

Our reasoning is that the presence of smiling in speech that follows laughter introduces acoustic differences from comparable speech sounds not preceded by laughter, and accurately reproducing these differences will reduce the loss function (MSE) used to train the synthesizer. The synthesizer's ability to achieve these loss-function reductions and "remember" when to produce smiling speech also across a breath likely relies primarily on the encoder (rather than the acoustic memory offered by the autoregression and the LSTM in the decoder), since the encoder contains several CNN layers ideal for learning short-range dependencies and operates on the phone level, where the smiling token and the next speech sound are adjacent. To gain an insight into the perceptual effect of this method, the two systems were used to create two different conditions of synthesizing smiling voice. The baseline original Tacotron 2 architecture was used in the first condition, which we call *unconstrained*, because it allows the system to use the proximity of laughter to influence the rendering without any further constraints. The second condition employs our prosody-controllable architecture. During inference, we set both the normalized mean f0 and speech rate values to 0, in order to assess whether smiling voice can still be elicited while directing the system to render a realization close to the median of the distribution in the corpus for these two prosodic features. Hence, we call this condition *constrained*. We propose this method to help isolate other acoustic-prosodic features characteristic to smiling speech, to be able to assess their perceptual impact.

4. Evaluation

4.1. Stimuli

The samples for this experiment were synthesized from 36 utterances that stated an opinion. Twelve of each type of statement was used: positive (e.g., "I agree with that"), negative (e.g., "I don't really agree with that idea") and neutral (e.g., "It's fine with me either way"). These utterances were then synthesized with the constrained system and the unconstrained system, both preceded by laughter and without laughter. This resulted in a total of 144 stimuli with combinations of 3 different parameters: model (constrained/unconstrained), context (laughter/no laughter) and content (positive/negative/neutral). Laughter and inhalation breaths were removed from the beginning of each utterance, as we were interested in whether the utterances themselves would carry the perceptual characteristics of smiling and did not want the participants to base their judgments on whether or not they heard laughter.

4.2. Acoustic-prosodic analysis

The evaluation samples were analyzed for a number of acoustic and prosodic features to determine whether they differed between model (constrained vs. unconstrained), context (post-laughter vs. no laughter), and/or content (positive/negative/neutral). Speech rate (syl/sec), mean f0, and f0 variation were measured and compared for the four different combinations of model and context. In addition, to compare the conditions in terms of breathiness, we calculated median smoothed cepstral peak prominence (CPPS, [28]) of all voiced frames in an utterance. CPPS quantifies strength of the first harmonic relative to the regression line over the power cepstrum, with high values corresponding to more modal voice and low values indicating breathiness. No significant difference was found in mean f0. However, analysis of variance showed a significant main effect of context on f0 variation. f0 variation (as measured by the standard deviation of f0 per utterance) was higher for speech following laughter (M=14.49, SD=5.56) than for speech synthesized without laughter (M=12.27, SD=5.61), $F(1,33) = 4.61$, $p < 0.05$. There was also a main effect of model on speech rate. The samples synthesized with the unconstrained model had a higher speech rate (M=5.15, SD=0.96) than samples synthesized with the constrained model (M=5.08, SD=1.01), $F(1,33) = 7.67$, $p < 0.05$. Finally, analysis of variance showed a significant effect of model on CPPS in voiced segments. Samples synthesized with the constrained model had a higher CPPS (M=12.35, SD=1.70) than samples synthesized with the unconstrained model (M=11.94, SD=1.57), $F(1,33) = 6.86$, $p < 0.05$.

4.3. Naturalness test

The systems were assessed for naturalness based on a web-based MUSHRA-like listening test. The test involved four versions of each utterance side by side (post-laughter/constrained, post-laughter/unconstrained, no laughter/constrained and no laughter/constrained) in randomized order with a scale for each item that ranged from 1 (very unnatural) to 5 (very natural).

4.4. Pairwise listening test

The extent to which post-laughter speech sounded like smiling was evaluated with a web-based forced-choice audio discrimination task. In one version of the test, stimuli synthesized with the constrained model were used. The other version used stimuli synthesized with the unconstrained model. Otherwise the setup was identical: smiling and non-smiling versions of each of the 36 utterances were presented side by side and the task was to choose which of the two versions sounded the most as if the speaker was smiling. The samples could be played as many times as needed. The order in which the two versions were displayed was randomized, as was the order of the utterances. The TTS samples used in the experiments are available here: https://www.speech.kth.se/tts-demos/ssw2021smiling

5. Results

5.1. Naturalness test

Twenty-one participants recruited online via Prolific completed the test. 54.38% were female and 47.62% were male. A within-subjects factorial analysis of variance showed that there was no main effect of content (positive/negative/neutral), context (post-laughter/no laughter) or model (constrained/unconstrained) on how natural-sounding participants rated the stimuli. The interaction between model and content was significant, $F(2,19) = 5.62$, $p < 0.05$, however, simple main effects of content were not significant for either the constrained or the unconstrained model. Results are summarized in Figure 2.

5.2. Pairwise listening test

A total of 60 participants were recruited via Prolific, of which 55.9% were female and 44.1% were male. All participants were native speakers of English. Half (30) received the unconstrained version of the task while the other half received the constrained version. One participant from the unconstrained group was excluded from the final analysis because their completion time was over 4 standard deviations above the mean.

Participants who heard stimuli synthesized with the unconstrained model identified the post-laughter synthesized speech as smiling 67.62% of the time, while participants who heard

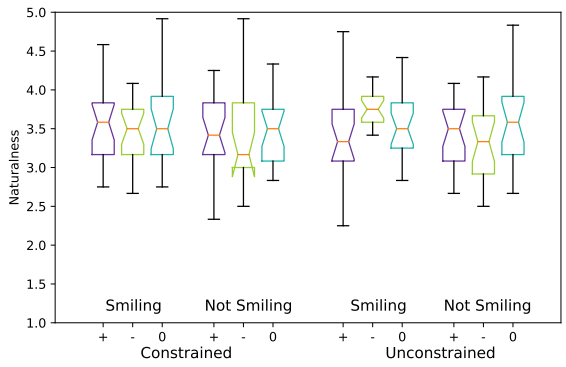

Figure 2: *Results of MUSHRA-like naturalness test on conversational utterances with positive (+), negative (-) and neutral (0) linguistic content.*

stimuli created with the constrained model identified post-laughter speech as smiling at a rate of 62.55%. Single-sample t-tests showed that this rate was significantly higher than chance for both participants who heard stimuli synthesized with the prosody-constrained model, $t(29) = 5.01$, $p < 0.001$, and those who rated stimuli synthesized with the unconstrained model, $t(28) = 10.26$, $p < 0.001$.

A mixed factorial analysis of variance was carried out to investigate the effect of content (positive/negative/neutral) and model (constrained/unconstrained) on the rate of identifying post-laughter speech as smiling. There were significant main effects of both content ($F(2,56) = 44.25$, $p < 0.001$) and model ($F(1,57) = 6.97$, $p < 0.05$). Participants who heard the prosodically unconstrained samples rated the post-laughter speech as smiling more often (M=67.62, SD=14.29) than those who heard the prosodically constrained samples (M=60.46, SD=17.33). In addition, participants were more likely to rate utterances that stated positive opinions as smiling (M=74.44,SD=12.17) compared to negative (M=65.96,SD=16.76) and neutral statements (M=51.55,SD=17.40). Post-hoc tests with the Bonferroni correction showed that the differences between these means were all significant ($p < 0.01$).

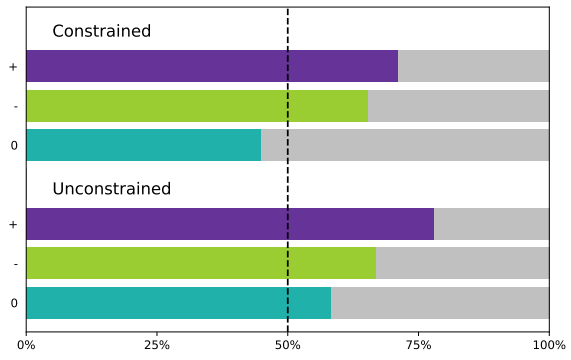

Figure 3: *Results of pairwise listening test, with coloured bars representing correct identification of smiling voice for utterances of positive (+), negative (-) or neutral (0) linguistic content produced by each model.*

6. Discussion

As hypothesized, it appears that speech following synthesized laughter is perceived as smiling, showing that the voice style we referred to as "smiling voice" conveys some of the perceptual aspects of smiling. The content of the utterances (whether they stated a positive, negative or neutral opinion) seemed to play a role in how participants performed at classifying post-laughter speech as smiling: listeners found it easier to discriminate between smiling voice and non-smiling voice when the content of the utterance was positive. Note that due to the use of a forced choice test in our evaluation, this does not mean that positive linguistic content increases the likelihood of perceived amusement in speech, but rather that it improves discrimination between two utterances on the basis of perceived amusement. This may indicate that, as a consequence of the context-driven approach, the TTS system was better at generating utterances that sounded like smiling when the content was positive. An alternative explanation would be that there is an effect of congruence between content and perceived emotional valence, whereby participants had an easier time distinguishing between smiling and non-smiling speech when the content and expressive characteristics of the synthesized smiling speech matched. However, participants had the most difficult time distinguishing between smiling and non-smiling speech when the linguistic content was neutral, which makes the first possibility more plausible.

Unlike in some previous studies, smiling speech in this case was not perceived as less natural than non-smiling speech. There appears to have been some joint effect of model and content on perceived naturalness, but since the differences in naturalness between positive, negative and neutral content were not significant with either model, this is difficult to interpret. The takeaway is that smiling voice synthesized with our method did not sound less natural.

In terms of acoustic/prosodic features, there were some differences between the constrained and unconstrained model. The unconstrained model produced breathier speech with a higher speech rate. However, these differences did not, in turn, seem to affect discrimination between smiling and non-smiling voice. Although participants who listened to samples from the unconstrained model did have an easier time with the discrimination task, there was no association between their performance and the parameters on which the models differed. Rather, f0 variation seems to have had the largest impact on performance at the discrimination task independent of model, consistent with previous findings that f0 variation is higher in naturally produced smiling speech [5, 6].

7. Conclusions

By synthesizing speech following laughter, we were able to exploit a spontaneous phenomenon in our models' training data to create the impression of smiling, without affecting the naturalness of the speech signal. This was the case even in a prosody-constrained model that restricted f0 and speech rate variation towards the median in the corpus, although listeners found the discrimination task more challenging with this model. Due to the context-driven nature of our method it seems that the linguistic content of the utterances affected the ease of discriminating between smiling and non-smiling speech. It is not entirely clear, whether this is due to the smiling speech sounding more like smiling when synthesized from utterances that suggest agreement, the non-smiling speech sounding less like smil-

ing in this context, both, or some other difference between the stimuli that made the discrimination task easier by making the stimuli sound more dissimilar.

The fact that the mere proximity of synthesized speech to synthesized laughter can create an impression of smiling means that it may not be necessary to synthesize laughter and smiling speech independently, as previous approaches suggest. Integrated into a conversational system equipped with voice style management modules, this approach could both create smiling voice that emerges in the context of laughter, and standalone amused speech (where the synthesized laughter is masked in the output), thereby improving the dialogue systems' capability to engage in informal social interactions.

8. Acknowledgements

This research is supported by the Swedish Research Council projects: Perception of speaker stance – using spontaneous speech synthesis to explore the contribution of prosody, context and speaker (VR-2020-02396), Connected: context-aware speech synthesis for conversational AI (VR-2019-05003), Prosodic functions of voice quality dynamics (VR-2019-02932), and the Riksbankens Jubileumsfond project CAPTivating – Comparative Analysis of Public speaking with Text-to-speech (P20-0298).

9. References

[1] L. Johnston, L. Miles, and C. N. Macrae, "Why are you smiling at me? social functions of enjoyment and non-enjoyment smiles," *British Journal of Social Psychology*, vol. 49, no. 1, pp. 107–127, 2010.

[2] K. Schmidt, R. Levenstein, and Z. Ambadar, "Intensity of smiling and attractiveness as facial signals of trustworthiness in women," *Perceptual and motor skills*, vol. 114, no. 3, pp. 964–978, 2012.

[3] K. Krys, C.-M. Vauclair, C. A. Capaldi, V. M.-C. Lun, M. H. Bond, A. Domínguez-Espinosa, C. Torres, O. V. Lipp, L. S. S. Manickam, C. Xing *et al.*, "Be careful where you smile: Culture shapes judgments of intelligence and honesty of smiling individuals," *Journal of nonverbal behavior*, vol. 40, no. 2, pp. 101–116, 2016.

[4] E. Otta, F. F. E. Abrosio, and R. L. Hoshino, "Reading a smiling face: Messages conveyed by various forms of smiling," *Perceptual and motor skills*, vol. 82, no. 3_suppl, pp. 1111–1121, 1996.

[5] C. Émond, L. Ménard, M. Laforest, F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, and L. Lamel, "Perceived prosodic correlates of smiled speech in spontaneous data." in *INTERSPEECH*, 2013, pp. 1380–1383.

[6] V. C. Tartter, "Happy talk: Perceptual and acoustic effects of smiling on speech," *Perception & psychophysics*, vol. 27, no. 1, pp. 24–27, 1980.

[7] I. Torre, "Production and perception of smiling voice," in *Proceedings of the First Postgraduate and Academic Researchers in Linguistics at York (PARLAY 2013) Conference. York, UK.*, 2014.

[8] M. Keough, A. Ozburn, E. K. McClay, M. D. Schwan, M. Schellenberg, S. Akinbo, and B. Gick, "Acoustic and articulatory qualities of smiled speech," *Canadian Acoustics*, vol. 43, no. 3, 2015.

[9] I. Torre, J. Goslin, and L. White, "If your device could smile: People trust happy-sounding artificial agents more," *Computers in Human Behavior*, vol. 105, p. 106215, 2020.

[10] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "Dialogpt: Large-scale generative pre-training for conversational response generation," *arXiv preprint arXiv:1911.00536*, 2019.

[11] 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*. PMLR, 2018, pp. 5180–5189.

[12] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, "Learning latent representations for style control and transfer in end-to-end speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6945–6949.

[13] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6189–6193.

[14] T. Raitio, R. Rasipuram, and D. Castellani, "Controllable neural text-to-speech synthesis using intuitive prosodic features," in *Proc. Interspeech*, 2020, pp. 4432–4436.

[15] É. Székely, G. E. Henter, J. Beskow, and J. Gustafson, "Spontaneous conversational speech synthesis from found data," in *Interspeech*, 2019, pp. 4435–4439.

[16] R. Dall, "Statistical parametric speech synthesis using conversational data and phenomena," Ph.D. dissertation, School of Informatics, The University of Edinburgh, Edinburgh, UK, 2017.

[17] 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*, vol. 10, 2019, pp. 245–250.

[18] ——, "Breathing and speech planning in spontaneous speech synthesis," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7649–7653.

[19] E. Lasarcyk and J. Trouvain, "Spread lips+ raised larynx+ higher f0= smiled speech?-an articulatory synthesis approach," *Proceedings of ISSP*, pp. 43–48, 2008.

[20] K. El Haddad, H. Cakmak, A. Moinet, S. Dupont, and T. Dutoit, "An HMM approach for synthesizing amused speech with a controllable intensity of smile," in *2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*. IEEE, 2015, pp. 7–11.

[21] K. El Haddad, S. Dupont, J. Urbain, and T. Dutoit, "Speech-laughs: an hmm-based approach for amused speech synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4939–4943.

[22] N. Campbell, "Conversational speech synthesis and the need for some laughter," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1171–1178, 2006.

[23] N. Tits, K. E. Haddad, and T. Dutoit, "Laughter synthesis: Combining seq2seq modeling with transfer learning," *arXiv preprint arXiv:2008.09483*, 2020.

[24] Y. Ferstl and R. McDonnell, "Investigating the use of recurrent motion modelling for speech gesture generation," in *Proc. IVA*, 2018, pp. 93–98. [Online]. Available: https://trinityspeechgesture.scss.tcd.ie

[25] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[26] K. Ito, "The LJ Speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[27] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3617–3621.

[28] J. Hillenbrand and R. A. Houde, "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," *Journal of Speech Language and Hearing Research*, vol. 39, no. 2, 1996.