Interspeech 2024
1-5 September 2024, Kos, Greece

Controlling Emotion in Text-to-Speech with Natural Language Prompts

Thomas Bott, Florian Lux, Ngoc Thang Vu

University of Stuttgart, Germany

[thomas.bott, florian.lux, thang.vu]@ims.uni-stuttgart.de

Abstract

In recent years, prompting has quickly become one of the standard ways of steering the outputs of generative machine learning models, due to its intuitive use of natural language. In this work, we propose a system conditioned on embeddings derived from an emotionally rich text that serves as prompt. Thereby, a joint representation of speaker and prompt embeddings is integrated at several points within a transformer-based architecture. Our approach is trained on merged emotional speech and text datasets and varies prompts in each training iteration to increase the generalization capabilities of the model. Objective and subjective evaluation results demonstrate the ability of the conditioned synthesis system to accurately transfer the emotions present in a prompt to speech. At the same time, precise tractability of speaker identities as well as overall high speech quality and intelligibility are maintained.

Index Terms: speech synthesis, prompting, emotional speech

1. Introduction

With rapid advancements in text-to-speech (TTS) systems in recent years [1, 2, 3], speech can be synthesized with naturalness and intelligibility comparable to human speakers [4, 5]. However, the one-to-many-mapping problem remains as one of the fundamental challenges. This refers to the fact that for a given input text there are infinitely many valid realizations which can differ in their prosody, including speaking style, intonation, stress or rhythm. A frequently used approach for mitigating this problem is to enrich the input side, i.e. the text to be encoded, with auxiliary prosodic information to alleviate the mismatch in the mapping [1, 6]. These additional prosodic inputs can often be controlled during inference time. Many previous approaches rely on reference audios for transferring the desired speaking style [7, 8, 9, 10, 11]. However, such methods oblige users to provide reference speech with the desired criteria during inference, which may not always be available. Addressing this issue, recent work has focused on using natural language descriptions to guide prosodic aspects in TTS systems, trained on speech datasets augmented with style descriptions [12, 13, 14, 15, 16]. Style tagging TTS [12] introduces a specialized loss, allowing to either provide reference speech or style tag during inference. PromptTTS [14] fine-tunes style embeddings on pre-defined labels such as gender, pitch, speaking speed, volume and emotion. PromptStyle [16] and InstructTTS [15] introduce a cross-modal style encoder which learns a shared embedding space of prompt and style embedding from speech. These approaches however require datasets with style descriptions which are expensive to create. Moreover, manually provided style descriptions are limited, as they typically follow similar patterns. PromptTTS 2 [17] tries to overcome this issue by labeling voice characteristics such as gender and speed from audio and generating descriptive prompts based on these properties automatically, which however limits the granularity of their control. Since emotional states are one of the most obvious aspects that can be expressed by varying prosodic features [18, 19, 20], emotional TTS is an important subfield within controllable TTS. In this vein, [21] automatically extract prompts from an emotional text dataset and match them to speech samples annotated with emotion labels.

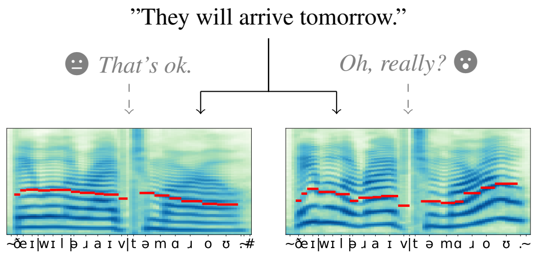

Figure 1: *Spectrograms with pitch contour for the same text, synthesized by our proposed system given two different emotional prompts. On the left the underlying emotion is neutral ("That's ok.") and on the right it is surprise ("Oh, really?").*

Our approach follows a similar strategy, combining publicly available emotional speech and text datasets and obtaining strong dependency between prosodic properties of audio and prompts. Furthermore, during each training iteration, prompts are selected randomly from a large pool, which increases the generalization capabilities of the TTS system and reduces the risk of learning connections that are too specific. In contrast to [21], who model speaker identities within the prompt, our method effectively combines prompt and speaker embeddings, allowing for precise prosodic and timbral controllability.

Summarizing our contributions, we propose 1) an architecture that allows for separate modeling of a speaker's voice and the prosody of an utterance, using a natural language prompt for the latter, 2) a training strategy to learn a strongly generalized prompt conditioning, and 3) a pipeline that allows users to generate speech with fitting prosody without manually selecting the emotion by simply using the text to be read as the prompt. We evaluate our contributions objectively and subjectively, finding that the emotion present in the prompt can be accurately transferred to speech while maintaining precise tractability of speaker identities and high speech quality. All of our code and models are available under an open source license[1].

[1] https://github.com/DigitalPhonetics/IMS-Toucan/tree/ToucanTTS_Prompting

10.21437/Interspeech.2024-1337

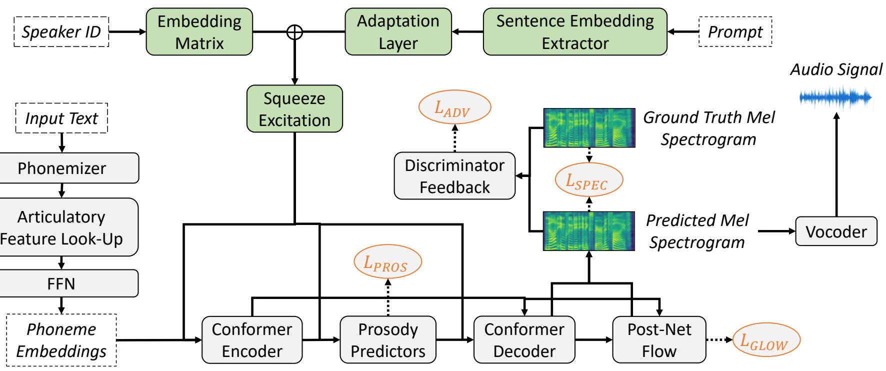

Figure 2: *Architecture of the prompt conditioned TTS system. Green components handle the integration of speaker and prompt embedding. + indicates concatenation. The loss functions with which the components in this system are optimized are marked in orange.*

Table 1: *Properties of the emotional speech datasets used for training. Each dataset is annotated with emotion categories.*

	ESD	RAVDESS	TESS
Speakers	10	24	2
Emotion Categories	5	8	7
Utterances per Emotion	350	8	200
Utterances in Total	17,500	1,440	2,800

2. Resources

2.1. Speech Datasets

We select two datasets without emotion annotation, LJSpeech[2] (single speaker, 24 hours) and LibriTTS-R [22, 23] (2,456 speakers, 585 hours), as well as three datasets which contain an emotion label along with each utterance: Emotional Speech Database (ESD) [24], Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [25] and Toronto Emotional Speech Set (TESS) [26]. All used datasets are publicly available, an overview of the properties of the emotional speech datasets can be seen in Table 1.

2.2. Text Datasets

In order to obtain non-descriptive emotional prompts, we extract sentences from text datasets labels with emotions. For training, a subset of the Yelp open dataset [27] is used, which contains 650k reviews for businesses and restaurants that contain highly emotional text from various different authors. We extract 10k prompts for each emotion category and confirm the emotion label with an auxiliary emotion classification model [28]. For the evaluation, we extract prompts that will remain unseen during training from DailyDialog [29] (containing 13,118 annotated sentences) and Tales-Emotion [30, 31] (containing 15,292 annotated sentences). From both datasets, 50 prompts per emotion category are extracted while ensuring quality with help of the emotion classification model. The conversational nature of the sentences in DailyDialog and the story telling style in Tales-Emotion demonstrate the usefulness of the TTS model for both dialog systems and reading stories. All used datasets are publicly available.

[2]`https://keithito.com/LJ-Speech-Dataset`

3. Methods

3.1. Prompt Conditioned TTS Architecture

An overview of the system is shown in Figure 2. We base our implementation on the IMS Toucan toolkit [32, 33] and extend it in order to condition the model on the emotional content of the textual prompt. The input text is converted into a sequence of phonemes using a phonemizer[3] with eSpeak-NG[4] as backend. Each phoneme is further transformed into an articulatory feature vector, following [34]. Spectrogram frames are generated by a FastSpeech-2-like system [1] comprising a Conformer encoder and decoder [35] as well as prosody predictors for duration and pitch and energy per phoneme as in FastPitch [6] using a small self-contained aligner following [36]. As proposed in PortaSpeech [37] a normalizing-flow-based post-net is used to improve details in high frequencies. Finally, the system is trained with discriminator feedback coming from an adversarial network optimized to distinguish real and generated spectrograms as proposed in [38].

Natural language prompts are fed into a sentence embedding extractor based on a DistilRoBERTa [39] model fine-tuned on the task of emotion classification [28]. Embeddings are obtained from the 756 dimensional hidden representation of the *[CLS]* token. Since the emotion classification is based on the embedding of this token, it is expected to effectively capture relevant information about the emotional content of the input. These prompt embeddings are further passed through a linear layer to enable them to adapt for TTS purposes, as the prompt encoder is not updated during TTS training. On the contrary, speaker embeddings are obtained from an embedding matrix, which is updated jointly during TTS training to capture the different speaker identities. To allow for zero-shot voice adaptation, a pretrained speaker embedding function could be used instead [40], which we choose to leave out for simplicity in this study. Prompt and speaker embeddings are concatenated and passed through a squeeze and excitation block [41]. This component models inter-dependencies between features of both sources and projects into the hidden dimensionality of the system. The use of the squeeze and excitation block is motivated by an internal pilot study, in which we compare the effectiveness of using various forms of conditioning mechanisms for this, such as concatenation followed by projection [40], addition [6], conditional layernorm [42], and the squeeze and excitation block

[3]`https://github.com/bootphon/phonemizer`
[4]`https://github.com/espeak-ng/espeak-ng`

Table 2: *Results for speaker similarity in terms of cosine similarity between speaker embeddings of ground truth and synthesized audio across all speaker ids of ESD.*

System	μ	σ
Baseline	0.950	±0.0044
Prompt Conditioned	0.953	±0.0025
EmoSpeech	0.861	±0.0115

[41]. Despite the differences being minor, we decided to go forwards with the squeeze and excitation block due to its slightly better performance in picking up fine nuances in a conditioning signal perceptually. The output of this block is a representation that contains information about the speaker identity and semantics of the prompt. This representation is integrated in the TTS system's pipeline by providing it as auxiliary input to the encoder and decoder as well as to the prosody predictors. In these places, it is integrated using conditional layernorm [42], which is proven to work well in a TTS pipeline [42, 33]. Adding the conditioning signals in multiple places is motivated by StyleTTS [43] who argue that a model quickly forgets about conditioning signals and needs to be reminded of them for more accurate conditioning. Finally, the spectrograms are converted to waveforms using the HiFi-GAN [44] generator with the Avocodo [45] discriminator set. During inference, this pipeline achieves a real-time-factor of 0.07 on a Nvidia GeForce RTX 2080 Ti GPU and 0.16 on a AMD EPYC 7542 CPU without the use of batching.

3.2. Prompt Inducing Training Procedure

The training of the TTS system is carried out by curriculum learning with two stages. Although conditioning prompts are still used in the first stage, its main purpose is to obtain a robust and high quality system. Therefore this stage includes LJSpeech and LibriTTS-R in addition to the emotional speech datasets. The large amount of training samples and huge variety of speakers is beneficial for a high speech quality and makes the system more robust against mispronunciations. Since LJSpeech and LibriTTS-R do not contain emotion labels, prompt embeddings are extracted from the corresponding utterances themselves. During the second stage, the model is trained using only the emotional speech datasets, allowing it to focus on learning the connection between prompt embeddings and speech emotion. For each training sample, a prompt embedding is selected randomly from the 10k available ones based on the emotion label. This ensures a high correspondence between prompt and speech emotion and further has the advantage that a large number of different prompts is seen which reduces the risk of overfitting and increases the generalization capabilities of the system such that during inference arbitrary prompts can be used. The whole system is trained for 120k steps during the first stage and an additional 80k steps in the second stage on a single Nvidia GeForce RTX A6000 GPU.

4. Experiments

4.1. Experimental Setup

We compare the proposed TTS system to a baseline following the exact same architecture except for the missing conditioning on prompt embeddings. Additionally we include EmoSpeech [46] [5] in our objective evaluation which conditions a FastSpeech 2 architecture on the discrete emotion labels of ESD. For our prompt conditioned system, the test sentences from the emotional text datasets are synthesized using the sentences themselves as prompts as well as using sentences annotated with a different emotion as prompts. This allows us to assess if the generated speech emotion relies on the provided prompt embedding throughout all experiments we are conducting. We also pass all ground truth speech samples through the vocoder of the TTS system, in order to allow a fair comparison to synthesized speech. Speaker identities from ESD are used for evaluation purposes, including the emotion categories *anger*, *joy*, *neutral*, *sadness* and *surprise*.

Table 3: *Cramér's V values indicating association strength between predicted emotions and underlying labels. For Prompt Conditioned Same input text is the same as prompt while Prompt Conditioned Other uses a different prompt from the text. All values are statistically significant with respect to the χ^2-test between predicted and underlying emotion labels with $\alpha = 0.005$.*

System	Cramér's V
Ground Truth	0.85
Baseline	0.06
Prompt Conditioned Same	0.80
Prompt Conditioned Other	0.80
EmoSpeech	0.96

Table 4: *Mean opinion score on a 5-point scale indicating naturalness, fluency and intelligibility as perceived by human raters.*

System	MOS	σ
Ground Truth	3.95	±0.42
Baseline	3.30	±0.45
Prompt Conditioned	3.37	±0.31

4.2. Objective Evaluation

4.2.1. Multi-Speaker Capabilities

We calculate the speaker similarity as the cosine similarity between speaker embeddings of ground truth speech samples and synthesized ones. Thereby speaker embeddings are extracted using a pre-trained speaker verification model [47]. The results in Table 2 show a high overall speaker similarity across all speakers of ESD, demonstrating that speaker identities are almost perfectly preserved during synthesis and not affected by the integration of prompt embeddings. Comparing our results to EmoSpeech, both our proposed system and our baseline perform significantly better. This is likely caused by the multi-speaker training phase in our curriculum learning procedure.

4.2.2. Prosody Controllability

We use an auxiliary speech emotion recognition model [48] trained on ESD to predict emotion labels of synthesized speech

[5] `https://github.com/deepvk/emospeech`

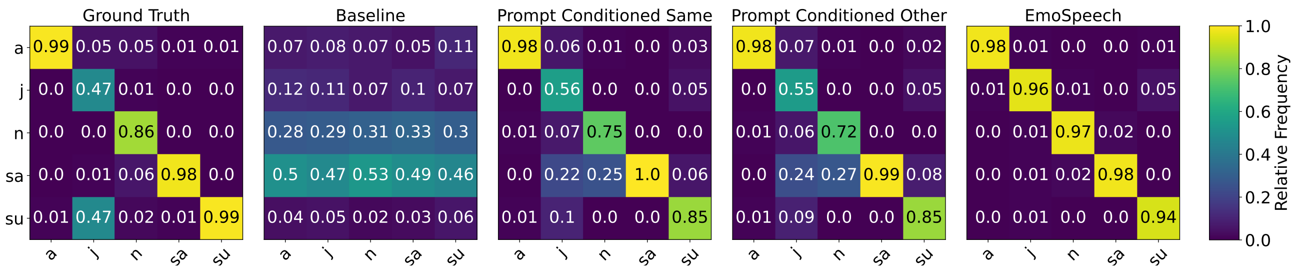

Figure 3: *Results of speech emotion recognition in terms of relative frequency for predicted emotion labels opposed to underlying ones. For Prompt Conditioned Same input text is the same as prompt while Prompt Conditioned Other uses different prompts. Emotion labels are abbreviated as follows: a(nger), j(oy), n(eutral), sa(dness), su(rprise).*

and compare those to the ground truth labels of the applied prompts. The confusion matrices in Figure 3 illustrate the relative frequencies of predicted emotion labels opposed to underlying ones. Furthermore, as a measure of association strength between emotion labels, Cramér's V [49] is calculated with results shown in Table 3. The emotion recognition model achieves an overall high accuracy and high association strength for ground truth speech, demonstrating that the emotion can generally be recognized reliably. Considering this, the strong alignment between underlying and predicted emotion labels for the prompt conditioned system indicates that the emotional content of the prompt is accurately transferred to speech. Moreover, the speech prosody exclusively relies on the provided prompt and is not influenced by the input text of the synthesized utterance as revealed by the high accuracy when combining prompts and input texts that stem from different emotion categories ("Prompt Conditioned Other"). In contrast, for the baseline, the predicted emotion categories are mostly sadness and neutral, showing that there is hardly any prosodic variation in generated speech irrespective of the emotional content of the input text. These observations are further confirmed by the Cramér's V values for the prompt conditioned system which are on par with ground truth following a student's t-test with $\alpha = 0.005$. EmoSpeech yields very strong results, outperforming even the ground truth. It is, however, restricted to discrete emotion labels, while our system captures a continuous space that doesn't require the manual selection of an appropriate emotion. This offers a great advantage over the state-of-the art of specialized systems, such as EmoSpeech, at the cost of a slightly degraded emotion accuracy.

4.3. Subjective Evaluation

Due to EmoSpeech's large variance in quality and intelligibility, which we noticed in a small pilot study, we chose to exclude it from the subjective evaluation, to prevent ceiling effects. Hence we compare our proposed system only with the baseline and human recordings in the following. We conducted a listening study with 82 participants, utilizing test sentences generated with both a female and a male speaker identity from ESD and varying prompts.

4.3.1. Speech Quality

We ask participants to rate speech quality on a 5 point scale, considering naturalness, fluency and intelligibility. The results of this mean opinion score (MOS) study based on 656 ratings (Table 4) indicate that synthesized speech from both the baseline as well as the proposed system is only slightly but significantly degraded compared to ground truth speech, yet not significantly different from one another (following a student's t-test with $\alpha = 0.005$). We conclude that the addition of prompt conditioning does not hinder the perceived naturalness of the TTS system.

Emotion	Female (speaker 15)	Male (speaker 14)
anger	4.61 ± 1.63	4.12 ± 1.73
joy	3.30 ± 1.60	3.63 ± 1.66
neutral	4.52 ± 1.67	4.41 ± 1.65
sadness	4.35 ± 1.68	4.66 ± 1.60
surprise	4.22 ± 1.42	4.55 ± 1.66
Overall	4.19 ± 0.94	4.27 ± 0.89

Table 5: *Mean similarity ratings (5-point scale) across all emotions and two speakers. Utterances were produced with the same prompt but different input text. The speaker IDs refer to the ones in ESD.*

4.3.2. Emotional Style Transfer

Finally, the participants are presented synthesized speech from the prompt conditioned system where the same prompt is used for multiple utterances with mismatching emotional content and are asked to rate the similarity of the speech samples to the prompt with respect to their prosodic realization on a 5-point scale. We received 320 prosody similarity ratings. The results are displayed in Table 5. The overall strong ratings for both speakers demonstrate that the model accurately follows the prompt for realizing speech emotion and that this emotion can effectively be transferred to arbitrary utterances, even those with different emotional content, by using the same prompt.

5. Conclusion

In this work, we propose a text-to-speech system that is conditioned on embeddings extracted from natural language prompts which makes the prosodic parameters of generated speech controllable in an intuitive and effective way. Prompt embeddings are concatenated with speaker embeddings and provided as input to the model's encoder, decoder and prosody predictors. Furthermore the proposed training strategy merges emotional speech and text datasets to obtain relevant prompts which are varied in each iteration, increasing the generalization capability and reducing the risk of overfitting. Evaluation results confirm the prosodic controllability through prompting while maintaining high speech quality and multi-speaker capability.

6. References

[1] Y. Ren, C. Hu, X. Tan, T. Qin *et al.*, "FastSpeech 2: Fast and High-Quality End-to-End Text to Speech," *arXiv:2006.04558*, 2020.

[2] E. Kharitonov, D. Vincent, Z. Borsos, R. Marinier *et al.*, "Speak, read and prompt: High-fidelity text-to-speech with minimal supervision," *arXiv:2302.03540*, 2023.

[3] C. Wang, S. Chen, Y. Wu, Z. Zhang *et al.*, "Neural Codec Language Models Are Zero-Shot Text to Speech Synthesizers, 2023," *URL: https://arxiv. org/abs/2301.02111. doi: doi*, 2023.

[4] Y. Liu, Z. Xu, G. Wang, K. Chen *et al.*, "DelightfulTTS: The Microsoft speech synthesis system for Blizzard Challenge 2021," *arXiv:2110.12612*, 2021.

[5] X. Tan, J. Chen, H. Liu, J. Cong *et al.*, "NaturalSpeech: End-to-End Text-to-Speech Synthesis with Human-Level Quality," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[6] A. Lancucki, "Fastpitch: Parallel Text-to-Speech with Pitch Prediction," in *ICASSP*. IEEE, 2021.

[7] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang *et al.*, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *ICML*. PMLR, 2018.

[8] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan *et al.*, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *ICML*. PMLR, 2018.

[9] Y. Yan, X. Tan, B. Li, T. Qin *et al.*, "Adaspeech 2: Adaptive Text to Speech with Untranscribed Data," in *ICASSP*, 2021.

[10] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior *et al.*, "Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone," in *ICML*. PMLR, 2022.

[11] F. Lux, J. Koch, and N. T. Vu, "Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech," in *SLT*. IEEE, 2023.

[12] M. Kim, S. J. Cheon, B. J. Choi, J. J. Kim *et al.*, "Expressive Text-to-Speech Using Style Tag," in *Interspeech*. ISCA, 2021.

[13] Y. Shin, Y. Lee, S. Jo, Y. Hwang *et al.*, "Text-driven Emotional Style Control and Cross-speaker Style Transfer in Neural TTS," in *Interspeech*. ISCA, 2022.

[14] Z. Guo, Y. Leng, Y. Wu, S. Zhao *et al.*, "Prompttts: Controllable Text-To-Speech With Text Descriptions," in *ICASSP*, 2023.

[15] D. Yang, S. Liu, R. Huang, G. Lei *et al.*, "InstructTTS: Modelling Expressive TTS in Discrete Latent Space with Natural Language Style Prompt," *arXiv*, 2023.

[16] G. Liu, Y. Zhang, Y. Lei, Y. Chen *et al.*, "PromptStyle: Controllable Style Transfer for Text-to-Speech with Natural Language Descriptions," *arXiv:2305.19522*, 2023.

[17] Y. Leng, Z. Guo, K. Shen, X. Tan *et al.*, "Prompttts 2: Describing and generating voices with text prompt," *arXiv:2309.02285*, 2023.

[18] A. F. G. Leentjens, S. M. Wielaert, F. van Harskamp, and F. W. Wilmink, "Disturbances of affective prosody in patients with schizophrenia, a cross sectional study," *J Neurol Neurosurg Psychiatry*, 1998.

[19] D. A. Sauter, F. Eisner, A. J. Calder, and S. K. Scott, "Perceptual Cues in Nonverbal Vocal Expressions of Emotion," *Quarterly Journal of Experimental Psychology*, 2010.

[20] M. D. Pell and S. A. Kotz, "On the Time Course of Vocal Emotion Recognition," *PLoS ONE*, 2011.

[21] J. Tu, Z. Cui, X. Zhou, S. Zheng *et al.*, "Contextual Expressive Text-to-Speech," *arXiv:2211.14548*, 2022.

[22] Y. Koizumi, H. Zen, S. Karita, Y. Ding *et al.*, "Miipher: A Robust Speech Restoration Model Integrating Self-Supervised Speech and Text Representations," *arXiv:2303.01664*, 2023.

[23] H. Zen, R. Clark, R. J. Weiss, V. Dang *et al.*, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech*, 2019.

[24] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset," in *ICASSP*. IEEE, 2021.

[25] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PloS one*, 2018.

[26] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020.

[27] N. Asghar, "Yelp dataset challenge: Review rating prediction," *arXiv:1605.05362*, 2016.

[28] J. Hartmann, "Emotion English DistilRoBERTa-base," https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[29] Y. Li, H. Su, X. Shen, W. Li *et al.*, "DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset," in *IJCNLP*. Asian Federation of Natural Language Processing, 2017.

[30] C. O. Alm, D. Roth, and R. Sproat, "Emotions from Text: Machine Learning for Text-based Emotion Prediction," in *EMNLP*, 2005.

[31] C. O. Alm and R. Sproat, "Perceptions of emotions in expressive storytelling," in *Speech Communication and Technology*, 2005.

[32] F. Lux, J. Koch, A. Schweitzer, and T. Vu, "The IMS Toucan system for the Blizzard Challenge 2021," in *Blizzard Challenge 2021*, 2021.

[33] F. Lux, J. Koch, S. Meyer, T. Bott *et al.*, "The IMS Toucan system for the Blizzard Challenge 2023," in *Proc. Blizzard Challenge Workshop*. Speech Synthesis SIG, 2023.

[34] F. Lux, J. Koch, and N. T. Vu, "Low-Resource Multilingual and Zero-Shot Multispeaker TTS," in *AACL*, 2022.

[35] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," *arXiv:2005.08100*, 2020.

[36] F. Lux, J. Koch, and N. T. Vu, "Exact Prosody Cloning in Zero-Shot Multispeaker Text-to-Speech," in *SLT*. IEEE, 2022.

[37] Y. Ren, J. Liu, and Z. Zhao, "PortaSpeech: Portable and High-Quality Generative Text-to-Speech," *NeurIPS*, 2021.

[38] P. Sani, J. Bauer, F. Zalkow, E. A. P. Habets *et al.*, "Improving the Naturalness of Synthesized Spectrograms for TTS Using GAN-Based Post-Processing," in *Speech Communication; 15th ITG Conference*, 2023.

[39] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," *ArXiv*, 2019.

[40] Y. Jia, Y. Zhang, R. Weiss, Q. Wang *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *NeurIPS*, 2018.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *CVPR*, 2018.

[42] Y. Wu, X. Tan, B. Li, L. He, S. Zhao *et al.*, "AdaSpeech 4: Adaptive Text to Speech in Zero-Shot Scenarios," in *Interspeech*, 2022.

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

[44] J. Kong, J. Kim, and J. Bae, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," in *NeurIPS*. Curran Associates, Inc., 2020.

[45] T. Bak, J. Lee, H. Bae, J. Yang *et al.*, "Avocodo: Generative adversarial network for artifact-free vocoder," *arXiv:2206.13404*, 2022.

[46] D. Diatlova and V. Shutov, "EmoSpeech: Guiding FastSpeech2 Towards Emotional Text to Speech," *arXiv:2307.00024*, 2023.

[47] D. Snyder, D. Garcia-Romero, A. McCree, G. Sell *et al.*, "Spoken language recognition using x-vectors." in *Odyssey*, 2018.

[48] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe *et al.*, "SpeechBrain: A General-Purpose Speech Toolkit," 2021, arXiv:2106.04624.

[49] H. Cramér, *Mathematical methods of statistics*. Princeton university press, 1999.