INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Emotional Prosody Control for Speech Generation

Sarath Sivaprasad[1,2*], *Saiteja Kosgi*[1*], *Vineet Gandhi*[1]

[1]CVIT, KCIS, IIIT Hyderabad, India
[2]TCS Research, Pune, India

sarath.s@research.iiit.ac.in, saiteja.k@research.iiit.ac.in

Abstract

Machine-generated speech is characterized by its limited or unnatural emotional variation. Current text to speech systems generates speech with either a flat emotion, emotion selected from a predefined set, average variation learned from prosody sequences in training data or transferred from a source style. We propose a text to speech(TTS) system, where a user can choose the emotion of generated speech from a continuous and meaningful emotion space (Arousal-Valence space). The proposed TTS system can generate speech from the text in any speaker's style, with fine control of emotion. We show that the system works on emotion unseen during training and can scale to previously unseen speakers given his/her speech sample. Our work expands the horizon of the state-of-the-art FastSpeech2 backbone to a multi-speaker setting and gives it much-coveted continuous (and interpretable) affective control, without any observable degradation in the quality of the synthesized speech. Audio samples are available at https://researchweb.iiit.ac.in/~sarath.s/emotts/

Index Terms: speech generation, prosody control, human-computer interaction

1. Introduction

Text-to-speech(TTS) applications strive to synthesize 'human-like speech.' This task not only needs modeling of the human vocal system (to generate the frequencies given a sequence of phonemes), but also captures the prosody and intonation variations present in human speech. Neural network models have made significant improvements in enhancing the quality of generated speech, and most state-of-the-art TTS systems like Deep Voice[1], Tacotron[2], and Fastspeech2[3] generate natural sounding voice. However, high-level affective controllability remains a much-coveted property in these speech generation systems and is a problem of interest in the speech community for well over three decades [4, 5].

Controlling emotional prosody (affective control) is vital for many creative applications (like audiobook generation, virtual assistants) and desirable in almost all speech generation use cases. Affective control is a challenging task, and even with the significant improvements in recent years, TTS systems today do not have high-level interpretable emotion control. The existing systems are restricted to either transfer of prosody from source style[6] or learning prosody globally given a phoneme sequence[3]. Habib *et al.* [7] proposed a system to control affect; however, their method cannot incorporate fine control and is limited to six discrete emotional states.

We propose a TTS system based on FastSpeech2[3] and bring in fine-grain prosody control and multi-speaker control. The improvements are achieved without any observable degradation in the synthesized speech quality and without compromising its ultra-fast inference. Like Fastspeech2, our model predicts low-level features from the phoneme sequence (e.g., pitch, energy, and duration). However, the proposed model incorporates high-level and interpretable sentence-level control over the low-level intermediate predictions computed for each phoneme. Our approach has profound implications from a usability perspective because (a) for a human, a phoneme level control is difficult to interact with, and our model allows sentence level emotional control and (b) low-level features like pitch, energy, duration, etc. are difficult to interpret and by conditioning them on *arousal valence* values, our model allows an expressible emotional control. We condition the encoder to scale for multiple speakers and transform the encoded vector to incorporate the continuous *arousal-valence* values. Our core contributions are:

- We extend the FastSpeech2 architecture to scale for multiple speakers based on fixed-size speaker embeddings.
- We propose a novel Prosody Control(PC) block into FastSpeech2 architecture to incorporate high-level affective sentence level control. We use scalar *Arousal-Valence* values on the low level and phoneme level variance features like pitch, energy, and duration.
- The proposed architecture hence allows to generate speech with fine grain emotional control as they can be chosen from a continuous and interpretable *Arousal-Valence* space.

2. Related work

Neural TTS: Neural network-based TTS have changed the landscape of speech synthesis research and have significantly improved the speech quality over conventional concatenative and statistical parametric approaches [8, 9]. Some of the recent popular neural TTS systems are Tacotron [2], Tacotron 2 [10], Deep Voice 1,2,3 [1, 11] and ClariNet [12]. These approaches first generate Mel-spectrogram autoregressively from text input. The Mel-spectrogram is then synthesized into speech using vocoders like Griffin-Lim [13], WaveNet [14] and Parallel WaveNet [15]. More recently, the FastSpeech [16] and FastSpeech 2 [3] methods approach TTS in a non-autoregressive manner and show extremely high computational gains during training and inference. Despite synthesizing natural-sounding speech, the above-mentioned neural TTS models give little or no control over the emotional expression for a given sentence.

Multiple Speaker TTS: There has been a major focus on scaling TTS systems to multiple speakers. Early neural multi-speaker TTS models require tens of minutes of training data per speaker. Fan *et al.* [17] proposed a neural network model which uses a shared hidden state representation for multiple speakers and speaker-dependent output layers. Gibiansky *et*

* equal contribution

http://dx.doi.org/10.21437/Interspeech.2021-307

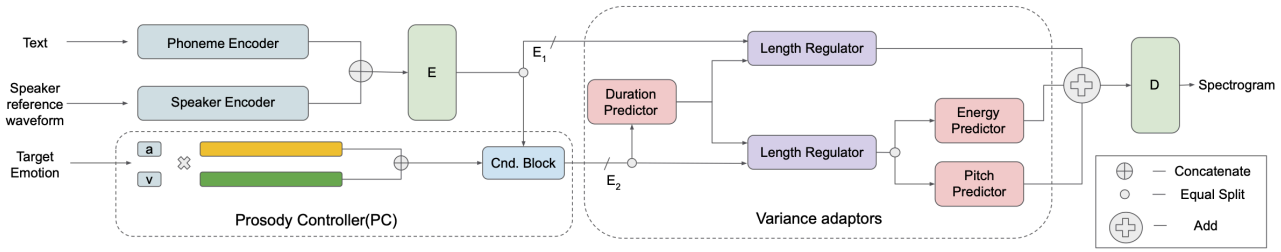

Figure 1: *Schematic diagram of the proposed model.*

al. [11] introduced a multi-speaker variation of Tacotron, which learned low-dimensional speaker embeddings for each training speaker. Their later work [18] scaled up to support over 2,400 speakers. Such systems [11, 18, 17] learn a fixed set of speaker embeddings and therefore only support the synthesis of voices seen during training. More recent approaches decouple speaker modeling from speech synthesis by independently training a speaker-discriminative embedding network [19]. The TTS models are then conditioned on these speaker-discriminative embedding obtainable from a few seconds of speech for the given speaker. Wan *et al.* [20] train speaker verification network, Jia *et al.* [21] condition the Tacotron 2 model on the embeddings of verification network. Our work extends such zero-shot multi-speaker support for a non-autoregressive model, FastSpeech2.

Prosody Control: Following enormous progress in neural TTS systems, the focus in recent years has shifted to modeling latent aspects of prosody. Humans speak with different styles and tonal variations, but there is an underlying pattern or constraint to these varying styles. The absence of an expected variation or the presence of an unexpected variation is easily detected as an uncanny speech by a human listener.

Wang *et al.* [22] proposed a framework to learn a bank of style embeddings called "Global Style Tokens" (GST) that are jointly trained within Tacotron (without any explicit supervision). A weighted combination of these vectors corresponds to a range of acoustic variations. Battenberg *et al.* [23] introduce a hierarchical latent variable model to separate style from prosody. Although such unsupervised methods [22, 23] can achieve prosodic variations, they can be hard to interpret and do not allow a straightforward control for varying the emotional prosody.

Skerry-Ryan *et al.* [24] proposed an end-to-end framework for prosody transfer, where the representation of prosody is learned from reference acoustic signals. The system transfers prosody from one speaker to another in a pitch-absolute manner. Karlapati *et al.* [6] proposed a framework for reference prosody by capturing aspects like rhythm, emphasis, melody, etc., from the source speaker. However, such reference-based methods cannot give a desired level of control as it requires a source reference for each different style of utterance. While such methods can work for scenarios like dubbing, they fall short on audiobook generation and other creative applications.

Habib *et al.* [7] proposed a generative TTS model with a semi-supervised latent variable that can control affect in discrete levels. Data collection involved recording reading text in either a happy, sad or angry voice at two levels of arousal. These six levels of arousal-valence combinations were used for partial supervision of latent variables. The model brings control only over discrete affective states (6 points), only representing a subset of emotions. Our work extends this idea by giving affect control over the continuous space of arousal and valence. Arousal(A) is a measure of intensity, whereas Valance (V) describes emotion's positivity or negativity. Russell *et al.* [25] show that these two parameters can represent various emotions in a 2D plane (Figure 2).

By conditioning TTS on AV values, our work allows fine-grained and interpretable control over the synthesized speech. We choose Fastspeech2[3] as the backbone due to its simplicity and ultra fast inference speed. Fastspeech2 predicts low level features like pitch, duration and energy for each phoneme and conditions the decoder on them. Our work facilitates a sentence level conditioning of these phoneme level features using scalar values for Arousal-Valence (AV).

3. Model

Our model uses Fastspeech2 as its backbone[3]. Unlike autoregressive models, Fastspeech2 does not depend on the previous frames to generate next frames, leading to faster synthesis. The model comprises of mainly three parts, namely: the encoder-decoder block, the prosody control block, and variance adaptor (Figure 1). The encoder block(E) and decoder block(D) are feed-Forward transformers with self-attention and 1-D convolution layers. The model has three inputs:

- Text: The text to be rendered as speech
- Speaker reference waveform: Audio sample of source speaker in whose voice the output will be rendered.
- Target Emotion: Arousal and valence values corresponding to the target emotion

Phoneme encoder gives a vector representation of fixed size for each of the phonemes in the input text. These embeddings are padded along sequence dimensions to match the number of phonemes in all the inputs across a batch. To incorporate speaker information into these embeddings, we condition our encoder(E) on speaker identity by using an embedding trained for speaker verification [20]. These embeddings capture the characteristics of different speakers, invariant to the content and background noise. Given a speaker reference waveform, using the pre-trained model, we generate 256 dimension speaker embedding. We concatenate the phoneme embedding with speaker embedding along the sequence dimension at the zeroth position. i.e., the speaker embedding appears as the first phoneme in the concatenated vector. This technique ensures the constant position of speaker embedding (irrespective of pad length of phonemes). The encoder(E) learns a representation for each phoneme attending to all other phonemes along with speaker embedding. We call this representation as E_1. We observed that conditioning the encoder with speaker embedding gives

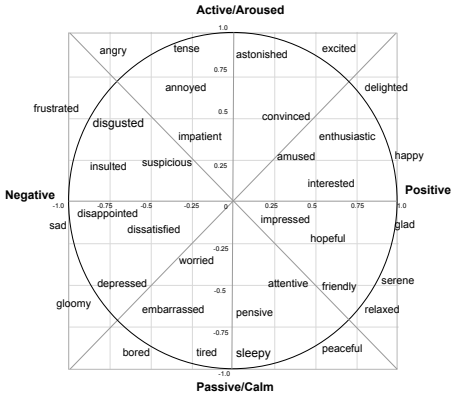

Figure 2: *The 2-D Emotion Wheel.*

better results than conditioning the decoder with speaker embedding. In our model, conditioning the decoder with speaker embedding did not capture the speaker's identity. We hypothesize this is because the variance predictions are dependent on speaker embeddings. The encoder's output and predicted variances (pitch, energy, duration) are decoded(at D) to obtain the Mel-spectrogram. The loss is computed between the generated Mel-spectrogram and the spectrogram of target speech(Mel-loss). This end-to-end structure forms the backbone of our system.

The Prosody Control(PC) block generates latent representation for each phoneme with affective cues from arousal and valence. We use two learnable vectors of length 256 to represent arousal and valance, respectively. The combined emotion is computed as the sum of these two vectors, weighted by arousal and valence inputs. The two vectors are trained with the loss computed at each of the variance predictors along with Mel-loss. The weighted sum is concatenated with E_1 and passed through a linear layer(condition block). The resulting representation is a phoneme embedding incorporating input emotions. We call this representation as E_2.

E_2 is passed through the duration predictor, which predicts a duration for each of the phonemes. Based on the duration (d) predicted for each phoneme using E_2, the length regulator expands the hidden states of the phoneme sequence d times for both E_1 and E_2. The total length of the hidden states in the two regulated embeddings now corresponds to the length of the output Mel-spectrograms. Pitch and energy are predicted at corresponding variance predictors using regulated E_2. Each variance predictor is trained with corresponding ground truth extracted from the speech wave. The energy and pitch computed are added to regulated E_1 and are passed to decoder block(D). The decoder outputs the Mel spectrogram. We use the MelGan vocoder [26] to generate raw speech from the spectrogram.

We use E_2 to predict the variances and use the resultant predictions to modify E_1. Decoder gets E_1 as input, which is not concatenated with affective cues. This strategy ensures that the emotion only modifies the pitch, energy, and duration, and the encoder-decoder module of the TTS can be trained, independent of the prosody control block. We propose this strategy to train the backbone and prosody controller block on LibriSpeech and MSP datasets independently. This ensures that errors incurred in transcribing MSP does not effect TTS quality. We train the prosody control block separately after training and freezing the encoder-decoder modules.

4. Experiments and Results

4.1. Dataset

We use two datasets to train our model. We train our backbone multi-speaker TTS model (leaving out Prosody Controller block) on LibriSpeech [27] dataset. LibriSpeech [27] contains transcripts and corresponding audio samples spoken by multiple speakers. Our model takes phoneme sequences as input. The text sequence is converted into phoneme sequence using the method proposed in [28]. We generate Mel spectrogram from the audio file following the work in [10]. This is used to compute loss with predicted Mel spectrogram. We compute energy, pitch, and duration from the speech to train the corresponding variance predictors. To train the duration predictor, we generate ground truth values of duration per phoneme using Montreal Forced aligner(MFA) [29]. MFA is a speech-text alignment tool used to generate time-aligned versions of audio files from the given transcript. LibriSpeech consists of two clean training sets comprising 436 hours of speech training data. We train on this data and use some of the speaker samples as a validation set. This dataset has no affective annotations.

We train our Prosody Controller(PC) block on MSP Podcast corpus [30] . MSP Podcast is a speech corpus annotated with emotions. It consisting of podcast segments annotated with emotion labels and valance arousal values ranging from 1 to 7. The corpus consists of 73K segments comprising 100 hours of speech, split into training and validation data. MSP Podcast corpus does not contain transcripts for the audio segments. To generate transcripts, we use Google speech-to-text API. We use Montreal-Forced-aligner(MFA) [29] to achieve alignment, and if MFA does not find proper alignment for the text and audio pair, the sample is discarded. This accounts for the inaccuracies of the speech-to-text API and background noise in audio samples. After applying MFA and discarding the wrongly transcribed samples, we are left with 55k samples comprising roughly 71 hours of speech. We use this data to train our prosody control module.

4.2. Training

We train our model in two stages.

- We first train our multi-speaker model barring Prosody Controller block on LibriSpeech [27] dataset. The encoder-decoder model with variance adaptors is trained together. The total loss consists of Mel loss(computed between predicted spectrogram and the spectrogram of corresponding ground truth audio), pitch loss, energy loss, and duration loss (each of which is computed directly from the ground truth audio). In this phase, E_1 is directly used as input to variance predictors. The model is trained on 4 GPUs with a batch size of 16. We use Adam optimizer to train the model. The training takes around 200k steps until convergence.
- In the second phase, we train our Prosody Controller block using MSP Podcast[30] corpus. We bring the emotion control by conditioning variance predictors on arousal valance values along with phoneme sequences(E_2). In this phase, we freeze the weights of the encoder-decoder model trained on LibriSpeech and only train the PC and variance adaptors. The model is

Table 1: *The MOS with 95% confidence intervals.*

Model	**Mean Opinion Score(MOS)**
Fastspeech2	3.65 $\pm$0.09
Our Model	3.62 $\pm$0.13
Speaker Similarity	**Mean Opinion Score(MOS)**
Same speaker set	3.6 $\pm$0.08
Same gender speaker set	2.55 $\pm$0.09
Different gender speaker set	1.2 $\pm$0.04
Affect control	**Avg. rater score in %**
Superlative emotion match	86.0

trained on 4 GPUs with a batch size of 16, and it takes 150k steps until convergence.

4.3. Model Performance

We measure the naturalness of generated speech, speaker sensitivity, and emotion control of our model through three user studies. We assess the voice's naturalness, speaker similarity using the Mean Opinion Score (MOS) collected from subjective listening tests. We use a Likert scale, with a range of 1 to 5 in 1.0 point increments. We evaluate emotion control using the average rater score. The results are reported in Table 1. The qualitative audio samples are available at `https://researchweb.iiit.ac.in/~sarath.s/emotts/`

Naturalness of generated speech: To evaluate the naturalness of the generated speech, we use a set of 30 phrases that do not appear in training set of either MSP or LibriSpeech and synthesize audio using our model. To compare the MOS of our model, we also synthesize the same phrases using Fastspeech2 [3]. A collection of samples from both these models are provided to users. Twenty proficient English speakers are asked to make quality judgments about the naturalness of the synthesized speech samples and asked to rate on a Likert scale of range 1 to 5 where 1 being 'completely unnatural' and 5 being 'completely natural'. The results in Table 1 show that similar scores are obtained for the two models. The results demonstrate that our model does not bring any noticeable distortions in terms of the naturalness of generated speech compared to the Fastspeech2 backbone.

Capturing reference speaker voice : Speaker similarity is evaluated in a similar fashion using MOS. We validate the speaker similarity on three different sets.

- Same speaker set: This set consists of sample pairs synthesized from the same speaker. The pair consists of either a ground truth speech and a synthesized sample or both synthesized samples.
- Same-gender speaker set: Here, we synthesize phrases for a set of speakers of the same gender. We form pairs of samples with the same gender but different speaker.
- Different gender speaker set: This set is curated by pairing synthesized audio samples generated for speakers from opposite genders.

Given a pair of samples, participants were asked to rate the similarity score of how close the voices sound on a Likert scale of 1 to 5. Where 1 corresponds to 'Not at all similar' and 5 corresponds to 'extremely similar'. For the same speaker set, we obtained a MOS of 3.6. This shows that our model can synthesize voices that sound close to a given target speaker. The MOS of 2.55 for the same gender set shows that audio generated from different speakers of the same gender has a certain degree of similarity. Furthermore, the low MOS of 1.2 for samples from different genders shows that our model's synthesized speech can be discriminated based on gender.

Affective control: Interpreting affect in rendition is subjective, challenging, and highly correlated with the content. We use user ratings to evaluate affect control. The model being conditioned on the continuous and meaningful space of emotion, user can change the level of emotion like happy to delighted, sad to depressed, etc., superlatively, during synthesis. We synthesize a set of phrases with different arousal valence(AV) values to evaluate the control obtained by changing AV values.

For our survey, we choose samples consisting of different levels of four emotions: happy, sad, angry, and excited. We provide a pair of samples for each of the above emotions, with one sample corresponding to the lower level of the emotion and the other corresponding to the higher level of respective emotion (e.g., happy to delighted). We choose appropriate AV values such that particular emotion is expressed in two degrees. Raters are asked not to judge the content and choose the sample expressing the particular emotion strongly (e.g., which one is angrier or happier). Every rater is shown eight different pairs of samples. We choose two pairs from each of the aforementioned emotions. The reported score shows the average percentage score obtained by raters in choosing the stronger emotion.

Compiling results from all the users, we observe that 86% of the raters can correctly choose the sample strongly expressing a particular emotion. The study shows the ability of the model to control prosody using arousal valance values.

5. Conclusions

Our work addresses the problem of emotional prosody control in machine-generated speech. In contrast to previous prosody control methods, which are either difficult to interpret by humans, require reference audio or allow selection only among a discrete set of emotions, our method allows a continuous and interpretable variation. We use the FastSpeech2 TTS model as a backbone and add a novel Prosody Control (PC) block. The PC blocks conditions the phoneme level variational parameters on sentence-level Arousal Valance values. We also extend the FastSpeech2 framework to support multiple speakers by conditioning it on a discriminative speaker embedding. Our user study results demonstrate the efficacy of the proposed framework and show that it can synthesize natural-sounding speech, mimic reference speakers, and allow interpretable emotional prosody control.

6. Acknowledgement

We want to thank Anil Nelakanti for the initial comments and discussions. The author would also like to thank Carlos Busso and others at The University of Texas at Dallas for sharing the MSP-Podcast database. The work would not have been possible without this valuable resource. We would also like to thank K L Bhanu Moorthy for the discussions.

7. References

[1] S. Ö. Arık, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *ICML*, 2017.

[2] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," *arXiv preprint arXiv:1703.10135*, 2017.

[3] Y. Ren, C. Hu, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text-to-speech," *arXiv preprint arXiv:2006.04558*, 2020.

[4] J. E. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, no. 1, pp. 1–1, 1990.

[5] M. Schröder, "Emotional speech synthesis: A review," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[6] S. Karlapati, A. Moinet, A. Joly, V. Klimkov, D. Sáez-Trigueros, and T. Drugman, "Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech," *arXiv preprint arXiv:2004.14617*, 2020.

[7] R. Habib, S. Mariooryad, M. Shannon, E. Battenberg, R. Skerry-Ryan, D. Stanton, D. Kao, and T. Bagby, "Semi-supervised generative modeling for controllable speech synthesis," *arXiv preprint arXiv:1910.01709*, 2019.

[8] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *ICASSP*, 1996.

[9] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system." in *SSW*, 2016, pp. 202–207.

[10] 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 *ICASSP*, 2018.

[11] A. Gibiansky, S. Ö. Arik, G. F. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech." in *NIPS*, 2017.

[12] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *arXiv preprint arXiv:1807.07281*, 2018.

[13] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 2, pp. 236–243, 1984.

[14] A. v. d. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[15] A. Oord, Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, E. Lockhart, L. Cobo, F. Stimberg *et al.*, "Parallel wavenet: Fast high-fidelity speech synthesis," in *ICML*, 2018.

[16] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech: Fast, robust and controllable text to speech," *arXiv preprint arXiv:1905.09263*, 2019.

[17] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for dnn-based tts synthesis," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 4475–4479.

[18] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: 2000-speaker neural text-to-speech," *ICLR*, 2018.

[19] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[20] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[21] Y. Jia, Y. Zhang, R. J. Weiss, Q. Wang, J. Shen, F. Ren, Z. Chen, P. Nguyen, R. Pang, I. L. Moreno *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *arXiv preprint arXiv:1806.04558*, 2018.

[22] 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 *ICML*, 2018.

[23] E. Battenberg, S. Mariooryad, D. Stanton, R. Skerry-Ryan, M. Shannon, D. Kao, and T. Bagby, "Effective use of variational embedding capacity in expressive end-to-end speech synthesis," *arXiv preprint arXiv:1906.03402*, 2019.

[24] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *ICML*, 2018.

[25] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[26] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019.

[27] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *ICASSP*, 2015.

[28] H. Sun, X. Tan, J.-W. Gan, H. Liu, S. Zhao, T. Qin, and T.-Y. Liu, "Token-level ensemble distillation for grapheme-to-phoneme conversion," *arXiv preprint arXiv:1904.03446*, 2019.

[29] M. S. S. M. M. W. McAuliffe, Michael and M. Sonderegger, "Montreal forced aligner: trainable text-speech alignment using kaldi," *In Proceedings of the 18th Conference of the International Speech Communication Association*, 2017.

[30] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, October-December 2019.