Exploiting Emotion Information in Speaker Embeddings for Expressive Text-to-Speech

Zein Shaheen, *Tasnima Sadekova,* *Yulia Matveeva, Alexandra Shirshova, Mikhail Kudinov*

Huawei Noah's Ark Lab

shaheen.zein@huawei-partners.com, sadekova.tasnima@huawei.com

Abstract

Text-to-Speech (TTS) systems have recently seen great progress in synthesizing high-quality speech. However, the prosody of generated utterances often is not as diverse as prosody of the natural speech. In the case of multi-speaker or voice cloning systems, this problem becomes even worse as information about prosody may be present in the input text and the speaker embedding. In this paper, we study the phenomenon of the presence of emotional information in speaker embeddings recently revealed for i-vectors and x-vectors. We show that the produced embeddings may include devoted components encoding prosodic information. We further propose a technique for finding such components and generating emotional speaker embeddings by manipulating them. We then demonstrate that the emotional TTS system based on the proposed method shows good performance and has a smaller number of trained parameters compared to solutions based on fine-tuning.

Index Terms: text-to-speech, prosody control, speaker verification

1. Introduction

The recent advances in deep learning have resulted in great progress in TTS technologies [1, 2, 3, 4]. However, it remains a challenge to synthesize expressive speech that can accurately capture prosodic characteristics of speech [5]. Several approaches have been proposed to address this problem, including methods based on variational inference [2, 6, 7], explicit control of prosody [8, 9], and using an external prosody encoder [4, 1]. One of the main challenges in training a TTS system capable of fine-grained control of various speech parameters is to obtain sufficiently disentangled representations of input text and speaker and style embeddings. This problem becomes worse in the case of voice cloning (VC) systems [10, 7, 11]. Usually, such systems either are based on a few-shot model adaptation [11, 10, 12] or make use of a devoted speaker encoder module [13, 10]. In the latter case, such systems are sometimes called zero-shot because they do not need transcriptions of the adaptation phrases.

Since [13, 10] it has become a common practice to obtain speaker embeddings from a pre-trained speaker classifier trained on a sufficiently large dataset [14, 15]. The common problem of such systems is their tendency to generate *average* neutral prosody [16]. On the other hand, in some cases, the generated speech can have non-neutral but random and inadequate prosody. Recently, it has been demonstrated that i-vectors and x-vectors contain style and emotion-related information ([17], [18]). Technically, this is a problem of poor disentanglement which often leads to undesirable and unpredictable changes in the speaking style in the synthesized utterances but in many cases, the components corresponding to information about prosody and emotions may be effectively detected and manipulated.

In this paper, we analyze speaker embeddings obtained from a pre-trained speaker verification model and identify components that affect the emotional and prosodic features of the speech. We further propose a procedure for the identification of these components. Finally, we design a TTS system with controllable prosodic and emotional features based on manipulating the identified components. Subjective and objective tests show that the emotions present in the speech generated by means of such manipulations are recognized by humans and detected by emotion recognition models.

The paper structure is as follows: in Section 2 we describe the architecture of our TTS model and speaker encoder; in Section 3 we provide evidence that the speaker embedding contains components corresponding to emotions and prosody and describe our method of detecting these key components; in Section 4 we describe results of the subjective evaluation of the Emotional TTS based on manipulating the key emotional components; we conclude in Section 5.

2. Model

Our TTS model is a modified version of Non-Attentive Tacotron (NAT) [19] which is a feature generation model based on Tacotron2 [20] with duration predictor. We augment NAT model with two additional modules for prosody control: phoneme prosody predictor and pitch predictor.
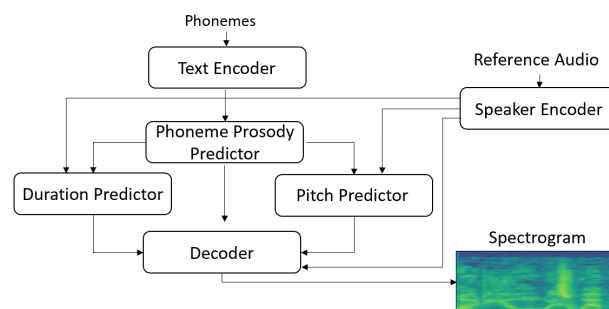

Figure 1: *Model Architecture. Speaker embedding are obtained from the same pre-trained model.*

Phoneme prosody predictor (see Fig.2) is implemented as two convolution layers followed by one LSTM layer and a linear layer. Each convolution layer has 1-d convolution with 256
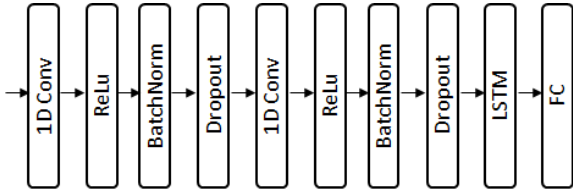
---

* Equal contribution

Figure 2: *Phoneme prosody predictor module*

channels and kernel sizes $k = 3$ followed by relu activation, batch normalization and dropout with rate 0.5. LSTM layer includes a uni-directional LSTM with 256 units and zone-out with rate of 0.1. The size of the final output projection is 4. For prosody predictor training we use the same method as [1] with the auxiliary prosody encoder module. They are trained jointly with the model with additional mean squared error (MSE) loss between predicted features. Pitch predictor is implemented as two Bi-LSTM with 256 units and zone-out rate 0.1 and one projection layer.

The overall architecture is depicted in Fig.1. We predict pitch frequencies directly. Duration predictor, decoder and encoder modules are the same as in NAT model. The model predicts mel-spectrogram features which are then fed to HiFi-GAN vocoder operating at 22050 Hz. For most hyperparameters we followed [19].

The model inputs phoneme sequence and reference audio sample used to obtain speaker embedding. We add speaker embedding as additional input to prosody, pitch, duration predictors and decoder.

We were inspired by the idea that some speaker representations such as i-vectors and x-vectors contain style and emotion-related information ([17], [18]) and may be applied to emotion recognition task. In our case, we studied 256 dimentional d-vectors, which were extracted from the pre-trained speaker verification model [1] (see [13]) trained on a combination of LibriSpeech [14] and VoxCeleb [15] datasets. Below we test the hypothesis that some components in the speaker embedding are responsible for emotions and speaker independent. Manipulation of such components allows to add emotions into generated speech. In the next section, we discuss how these components can be found and manipulated.

3. Method

Bearing in mind the simple truth that obtaining reliable disentangled features for a TTS system is a difficult problem [7] we started with a simple analysis of the speaker embeddings generated by our pre-trained speaker encoder. Our goal was to investigate how much information about emotions was stored in the embeddings by our speaker encoder. The Fig.3 demonstrates the results of LDA analysis of the embeddings calculated on the phrases from ESD [21] dataset. ESD contains 350 parallel utterances corresponding to one of 5 emotions (*neutral*, *sad*, *angry*, *happy*, *surprised*) produced by 10 English speakers.

We may see that while some classes corresponding to emotions in ESD may overlap all of them are separated from *neutral* style. This observation proves that the speaker embeddings contain information about emotions so we can try to find components of the embeddings having the best influence on emotions of the generated utterances.
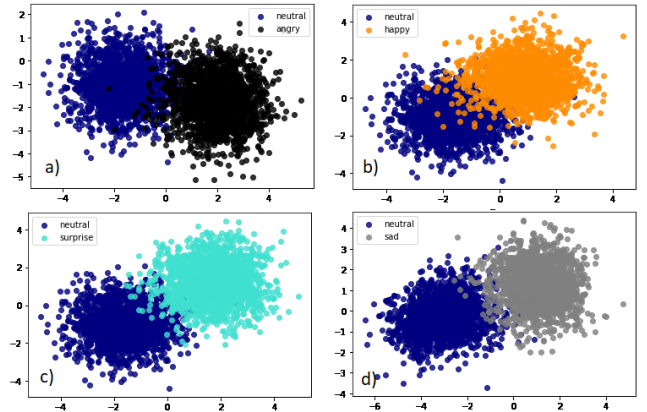

Figure 3: *LDA projections of speaker embeddings giving the best separation between classes of emotions in ESD. a) Neutral vs. Angry; b) Neutral vs. Happy; c) Neutral vs. Surprise; d) Neutral vs. Sad*

3.1. Components selection

Below we describe a method of finding the key emotional components based on selection of the most important features in the SVM emotion classifier:

For each emotion $e \in \{sad, angry, happy, surprised\}$:

1. for each speaker s from the set of ESD speakers we train a binary linear SVM classifier for each emotion e against *neutral* with weights W_{es}
2. Calculate importance $I_e = \sum_s W_{es}$
3. Take top K components by absolute value as key components for emotion e

For our experiments, we chose top-10 positive and top-20 negative components for each emotion. During our preliminary experiments, we found out that setting top negative components to zero worked as best as setting them to their optimal parameters, so we chose the first variant. The best values for top-10 positive components were searched in the range $(0, 2M_e)$ where M_e is the maximum value of the embedding component reached on the recordings corresponding to the target emotion e in ESD. The optimal values were found via Bayesian optimization [22].

Thus, for each emotion e we obtained a transformation $T_e(s)$ mapping a neutral speaker embedding s into an emotional speaker embedding s_e used as input to our TTS model. Each transformation had only 10 parameters found through an optimization process. Below we give a detailed description of the optimization procedure.

3.2. Hyperparameter optimization

Bayesian optimization is a popular method of parameter selection in cases when the optimized function $f(\theta)$ is expensive to evaluate. It utilizes a surrogate model to estimate value of the target function depending on the parameter values θ_t and then choose the next parameter combination θ_{t+1} to try. The surrogate model is usually a Gaussian process that provides a Bayesian posterior probability distribution $p(f(\theta))$ over potential values of the objective function. Each time we estimate a function at a new point θ_{t+1}, this posterior distribution is updated and used in a pre-selected acquisition function $\alpha(\theta)$ to

---

[1] https://github.com/CorentinJ/Real-Time-Voice-Cloning

define where to sample next.

An objective function in our task is a combination of three scores s_1, s_2, s_3 each of which controls one of the following characteristic:

- **Sound quality score.** Mean opinion score (MOS) is a subjective metric, so it cannot be used in the optimization scenario. For our objective function we used score s_1 obtained from automatic MOS estimation model NORESQA_MOS [23]. NORESQA_MOS is based on Non-Matching References (NMRs) approach and trained to predict a relative quality score compared to any provided reference without dependence on content and speaker's gender. It shows high generalization ability on out-of-domain datasets and has open-source realization and model[2]. Three records of different speakers from VCTK dataset [24] were utilized as references. It is a high-quality dataset with samples form 109 native English speakers.

- **Emotion score.** To ensure that speech synthesized with modified speaker embedding is expressive we trained a simple 5-class emotion classifier on the training part of ESD dataset following the architecture described in [21]: one LSTM and two dense layers with mel frequency cepstral coefficients (MFCCs) as input features. The final classifier accuracy on the dataset testing records reached 88.3%. Predicted probability of the target emotion was used as score s_2 for optimization.

- **Speaker similarity Score.** Component manipulation must influence only emotions, preserving speaker characteristics. For speaker similarity control we chose SOTA speaker representation model TitaNet [25] trained with speaker identification objective. We employ prediction of the largest pretrained speaker verification model as the score s_3. An embedding from synthesized speech is compared with the embedding from the ground truth record of the target speaker and emotion.

We run Bayesian optimization for 200 iterations for each emotion separately. At every step of optimization we generated 3 sentences for every 10 English speakers from ESD with speaker embeddings modified according to the chosen values of the current component. Described scores s_1, s_2, s_3 were averaged across all records and summed with weights $w_1 = 1$, $w_2 = 5$, $w_3 = 12$. For the experiments we used implementation from *scikit-optimize* library.

3.3. Model validation

To verify our hypothesis of the ability of emotion manipulation we conducted a subjective experiment, which was designed as a preference test against a baseline model without changes in speaker embedding. For testing purposes, we took the same 10 speakers from ESD and synthesized 4 sentences for 4 emotions: *angry*, *happy*, *sad*, and *surprise*. Every synthesized utterance was evaluated by 10 assessors on Amazon Mechanical Turk (AMT). The overall number of unique workers who participated in our tests was 46. To ensure the high quality of the obtained preference levels, only Master workers were allowed to complete tasks. Each task contained two variants of the same synthesized dialog with two different versions of the second phrase: one was synthesized by the baseline model and another was synthesized by the emotional model. Each worker was asked which recording sounded better in terms of expressed emotions.

Table 1: *Preference test on Amazon Mechanical Turk*

	Baseline Preferred	Emotional Preferred
Surprise	46.19%	**53.81%**
Happy	45.54%	**54.46%**
Sad	**56.37%**	43.63%
Angry	30.05%	**69.95%**
Overall	45.0%	**55.0%**

The results are shown in Table 1. We may see that for emotions *surprise*, *happy* and *angry* the assessors prefer the modified version. The emotion *sad* got a lower rate from the assessors. This emotion was less discernible and easily confused with non-emotional *neutral* class. In this case assessors preferred the baseline as the recordings generated by the baseline model tended to have better sound quality (see Section 4). Nevertheless, we see that the preference test and the emotional classifier performance clearly indicate that the proposed method allows generating emotional speech which proves that our method is valid.

In the next section, we test the model against other methods of controllable TTS in terms of emotion fidelity, speaker similarity, and sound quality.

4. Experiments

The goal of our main experiment was to understand the trade-off between the general features of the multi-speaker TTS system (i.e. sound quality and speaker similarity) and the capability of the model to generate the expressive speech as the results of the preference test on class *sad* suggested that manipulating the key emotional components of the speaker embedding could lead to degradation in sound quality or speaker similarity. Our second goal was to compare our method with one of the standard methods of emotional speech synthesis. For comparison we chose Tacotron with Global Style Tokens (GST) model [26].

For a more fair comparison we used a modified version of GST model. The sequence-to-sequence Tacotron part was replaced with Non-attentive Tacotron with the same architecture and parameters as in the main model including speaker embedding conditioning. The reference encoder and the style token layer were the same as in [26]: a stack of the 6 convolutional layers followed by a single layer unidirectional GRU and a multi-head attention module with 4 heads and 10 style tokens.

The baseline NAT and GST models were trained on LibriTTS dataset [27] which is a multi-speaker English corpus of approximately 585 hours of read English speech. As long as LibriTTS contained mostly neutral utterances GST model was unable to copy emotions from emotional reference recordings. For this reason, GST model was additionally fine-tuned on ESD dataset. One reference audio and speaker embedding averaged across all recordings made by the speaker with the corresponding emotion were passed as input. Thus, after fine-tuning each speaker had 4 speaker embeddings corresponding to $256 \times 4 = 1024$ parameters per speaker which gives 10.2k parameters for 10 English speakers used in our experiments.

The experiments were carried out on the Appen platform. We asked participants to estimate speaker similarity and sound quality of the synthesized speech on a 5-point scale. In addition, for each recording, the participant was asked which emotion dominated in the generated utterance with 5 options corresponding to the emotions from ESD. We synthesized 3 sentences for each of the 10 English speakers and 4 non-neutral

emotions from ESD, resulting in 120 records. In each task, the participant was given a neutral reference audio of the target speaker, one neutral ground-truth recording of the same speaker, one bad recording for quality control, and 3 synthesized recordings. The aim of adding the neutral ground-truth recording was to estimate an upper bound for similarity and sound quality because generally they may vary between groups of workers. Each audio was assessed by 10 participants to ensure reliability of our results. To ensure the high quality of the obtained preference levels, native English workers with level 2 were allowed to complete tasks. The total number of participants was 240.

We also needed to estimate the upper bound of emotion recognition accuracy for our workers. For this reason, we run additional test on ground truth emotional recordings. We asked participants to label emotions for the recordings from ESD. We synthesized 3 sentences for 10 speakers and 4 non-neutral emotions from ESD. Each audio was assessed by 10 participants.

Table 2: *Subjective tests on Appen: Speaker similarity (SMOS), Sound Quality (MOS) and Emotion recognition accuracy. Score with (*) was obtained in a separate test run with emotional recordings.*

	Similarity	Sound Quality	Recog. acc.
Baseline	**3.70 ± 0.05**	**3.96 ± 0.05**	9.8%
EM (Ours)	3.55 ± 0.08	3.76 ± 0.06	57.9%
GST	3.61 ± 0.08	3.75 ± 0.06	**58.1%**
GT	4.06 ± 0.05	4.09 ± 0.05	58.2*%

The results are summarized in the Table 2. The results support the assumption that there is a trade-off between similarity and expressiveness of the speech. The baseline model has shown the best performance in terms of sound quality and speaker similarity while for both emotional models we see a drop in these scores although the score ranges overlap. We consider that emotional recordings in LibriTTS are enough for the TTS model to learn how to detect and exploit the emotional components in the embeddings but not enough to learn how to synthesize emotional recordings for any voice.

We see that performance of both emotional models is similar but the number of additional parameters needed to support emotions in our method is much lower: only 40 versus over 10k for GST. Moreover, in fact the parameters of embedding transformations may also work in zero-shot settings for non-ESD speakers which is also an advantage of our approach compared to fine-tuning based ones. A subset of audio samples used in the human evaluation and examples of zero-shot emotional synthesis are available on our demo page `https://controllable-tts.github.io/`.

5. Conclusions

In this work we have investigated the possibility of controlling emotions of the generated speech by manipulating specific components of the speaker embedding. We showed that such components can be effectively detected and manipulated. We also proposed a method of speech synthesis making use of speaker embedding manipulation and showed that this method could demonstrate performance comparable to widely used models of controllable speech generation with no need for fine-tuning on emotional data. We believe that the analysis of the emotional components should be an important part of the design of multi-speaker TTS systems making use of pre-trained speaker encoder.

6. References

[1] Y. Liu, Z. Xu, G. Wang, K. Chen, B. Li, X. Tan, J. Li, L. He, and S. Zhao, "Delightfultts: The microsoft speech synthesis system for blizzard challenge 2021," 10 2021.

[2] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," *ArXiv*, vol. abs/2106.06103, 2021.

[3] V. Popov, I. Vovk, V. Gogoryan, T. Sadekova, and M. Kudinov, "Grad-tts: A diffusion probabilistic model for text-to-speech," *ArXiv*, vol. abs/2105.06337, 2021.

[4] A. Lancucki, "Fastpitch: Parallel text-to-speech with pitch prediction," in *ICASSP*, 2021.

[5] X. Tan, T. Qin, F. Soong, and T.-Y. Liu, "A survey on neural speech synthesis," 2021. [Online]. Available: https://arxiv.org/abs/2106.15561

[6] W.-N. Hsu, Y. Zhang, R. Weiss, H. Zen, Y. Wu, Y. Wang, Y. Cao, Y. Jia, Z. Chen, J. Shen, P. Nguyen, and R. Pang, "Hierarchical generative modeling for controllable speech synthesis," 10 2018.

[7] Y. Zhang, R. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," *Interspeech 2019*, pp. 2080–2084, Sep 2019.

[8] R. Valle, J. Li, R. J. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6189–6193, 2019.

[9] J. Lee, J. Y. Lee, H. Choi, S. Mun, S. Park, and C. Kim, "Into-tts : Intonation template based prosody control system," *ArXiv:2204.01271*, 2022.

[10] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural Voice Cloning with a Few Samples," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 10 019–10 029.

[11] M. Chen, X. Tan, B. Li, Y. Liu, T. Qin, sheng zhao, and T.-Y. Liu, "AdaSpeech: Adaptive Text to Speech for Custom Voice," in *International Conference on Learning Representations*, 2021.

[12] T. Sadekova, V. Gogoryan, I. Vovk, V. Popov, M. Kudinov, and J. Wei, "A Unified System for Voice Cloning and Voice Conversion through Diffusion Probabilistic Modeling," in *Proc. Interspeech 2022*, 2022, pp. 3003–3007.

[13] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, z. Chen, P. Nguyen, R. Pang, I. Lopez Moreno, and Y. Wu, "Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis," in *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 2018, pp. 4480–4490.

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[15] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017.

[16] S. Seshadri, T. Raitio, D. Castellani, and J. Li, "Emphasis Control for Parallel Neural TTS," in *Proc. Interspeech 2022*, 2022, pp. 3378–3382.

[17] R. Pappagari, T. Wang, J. Villalba, N. Chen, and N. Dehak, "X-vectors meet emotions: A study on dependencies between emotion and speaker recognition," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7169–7173, 2020.

[18] R. Xia and Y. Liu, "Using i-vector space model for emotion recognition," in *Interspeech*, 2012.

[19] J. Shen, Y. Jia, M. Chrzanowski, Y. Zhang, I. Elias, H. Zen, and Y. Wu, "Non-Attentive Tacotron: Robust and Controllable Neural TTS Synthesis Including Unsupervised Duration Modeling," *CoRR*, vol. abs/2010.04301, 2016. [Online]. Available: http://arxiv.org/abs/2010.04301

[20] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 4779–4783.

[21] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[22] P. I. Frazier, "A tutorial on bayesian optimization," *arXiv preprint arXiv:1807.02811*, 2018.

[23] P. Manocha and A. Kumar, "Speech Quality Assessment through MOS using Non-Matching References," in *Proc. Interspeech 2022*, 2022, pp. 654–658.

[24] J. Yamagishi, C. Veaux, and K. MacDonald, "CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92)," 2019.

[25] N. R. Koluguri, T. Park, and B. Ginsburg, "Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8102–8106.

[26] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*, 2018, p. 5180–5189. [Online]. Available: https://arxiv.org/abs/1803.09017

[27] H. Zen, R. Clark, R. J. Weiss, V. Dang, Y. Jia, Y. Wu, Y. Zhang, and Z. Chen, "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," in *Interspeech*, 2019.