CROSS-SPEAKER EMOTION TRANSFER BASED ON SPEAKER CONDITION LAYER NORMALIZATION AND SEMI-SUPERVISED TRAINING IN TEXT-TO-SPEECH

Pengfei Wu, Junjie Pan, Chenchang Xu, Junhui Zhang, Lin Wu, Xiang Yin, Zejun Ma

AI Lab, ByteDance
{wupengfei.ganyue, panjunjie.jeff, xuchenchang}@bytedance.com

ABSTRACT
In expressive speech synthesis, there are high requirements for emotion interpretation. However, it is time-consuming to acquire emotional audio corpus for arbitrary speakers due to their deduction ability. In response to this problem, this paper proposes a cross-speaker emotion transfer method that can realize the transfer of emotions from source speaker to target speaker. A set of emotion tokens is firstly defined to represent various categories of emotions. They are trained to be highly correlated with corresponding emotions for controllable synthesis by cross-entropy loss and semi-supervised training strategy. Meanwhile, to eliminate the down-gradation to the timbre similarity from cross-speaker emotion transfer, speaker condition layer normalization is implemented to model speaker characteristics. Experimental results show that the proposed method outperforms the multi-reference based baseline in terms of timbre similarity, stability and emotion perceive evaluations.

Index Terms— emotion transfer, text-to-speech, global style tokens, conditional layer normalization

1. INTRODUCTION
Text-to-speech (TTS) aims to produce natural-sounding speech. In recent years, various deep learning based TTS acoustic methods [1-5] and vocoder methods [6-8] are proposed, to generate high quality speech. However, human speech contains a lot more super-segmental information beyond texts, such as prosody and emotion, which are essential to further improve the naturalness of the synthesized speech.

Several style modeling methods [9-19] are proposed to model these non-textual information. Part of them highly depend on data with extra annotations [9,10], which are complicated and lack generality and consistency, making it impractical in commercial production. On the other hand, some propose unsupervised methods [11-14] with encoder-decoder architectures, where utterance level representations are extracted by a style or prosody encoder, to boost the expressiveness of synthesized speech. However, the learned representations usually lack interpretability and controllability. Semi-supervised methods [16,18] are therefore proposed to improve the interpretability of the learned representations by providing partial supervision. Wu et al. [16] propose an emotion control method called Semi-GST with 5% supervision data which heuristically turns style token weights into one-hot vectors by introducing a semi-supervised cross-entropy loss. In this manner, a specific physical interpretation (a single emotion) can be assigned to one style token. Recently, multi-reference methods [15,17] are proposed for better performance and interpretation. They tried to learn independent speaker and style representations through multiple reference encoders. By introducing strategies such as inter-cross training, paired-unpaired triplets and adversarial cycle consistency, those methods achieve the purpose of learning independent speaker and style representations.

This paper aims to transfer emotions from source speaker with multi-emotional speech corpus to the target speaker without emotional annotations. Two main issues exist in this kind of emotion transfer task, that in the target speaker synthesis, the emotion perception and pronunciation stability should be guaranteed, and the timbre similarity should be kept. To resolve that, we propose a parallel Tacotron [20] based model, with the variational residual encoder replaced by a global style tokens (GSTs) module because we are aiming at controllable cross-speaker emotion transfer. The contributions of this paper include the following. Firstly, we propose to use GSTs and semi-supervised training strategy for controllable cross-speaker emotion transfer. Secondly, we introduce speaker condition layer normalization (SCLN) into cross-speaker emotion transfer task.

We notice that recent work by Li et al. [19] which achieves cross-speaker emotion transfer based on Tacotron 2. The proposed work differs from Li’s as follows. Firstly, their backbone is an autoregressive model, while our back-bone is non-autoregressive. We believe that non-autoregressive models perform better in feature decoupling because they do not directly take the previous frame as input when predicting the next one, resulting in less feature leakage. Secondly, they get speaker-independent emotion embeddings by explicitly constraining speaker and emotion embeddings, while we learn speaker embeddings by SCLN blocks, and emotion embeddings by emotion tokens and semi-supervised strategy.
2. PROPOSED METHOD

The architecture of the proposed cross-speaker emotion transfer model is illustrated in Fig. 1(a). It mainly consists of a Transformer-based text encoder, a Transformer-based duration predictor, an upsampling block, a spectrogram decoder stack, a reference encoder, and an emotion token layer.

2.1. Duration Predictor

The duration predictor takes the encoder outputs added by the emotion embedding as input, and outputs the predicted phoneme durations. We extract the ground-truth phoneme durations, by an external hidden Markov model (HMM)-based aligner, to train the duration predictor. As shown in Fig. 1(b), it consists of a Transformer block, a 3×1 LConv block, and a projection layer. A sinusoidal positional embedding is added to the inputs and then fed into the Transformer block. The outputs of the Transformer block are used for upsampling and added to speaker embedding as the input of the LConv block. At last, the speaker and emotion-related phoneme durations are predicted by the LConv block and the projection layer.

2.2. SCLN-LConv Block

The spectrogram decoder consists of six SCLN-LConv blocks, and the architecture of SCLN-LConv block is shown in Fig. 1(c). We insert the SCLN module after the LConv block and FF layer in each block. The SCLN module is a conditional layer normalization [21], which takes the speaker embedding as inputs and predicts the scale and bias parameters of layer normalization.

2.3. Emotion Token Layer and Semi-Supervised Training

We model emotion properties by introducing utterance-level emotion embeddings, which are extracted as follows. Firstly, the target mel-spectrogram is fed into a reference encoder, which encodes the reference mel-spectrogram into a fixed-length vector called reference embedding. Thereafter, the reference embedding is used as query to calculate a set of weights with pre-defined emotion tokens using a single-head attention module. Finally, the emotion embedding is generated by the weighted sum of the emotion tokens.

Similar to Wu et al. [16], we add an emotion classifier loss between token weights and one-hot emotion ID to ensure that the trained tokens have a one-to-one correspondence with emotions at the training stage. In this way, the emotion embedding can be generated by multiplying one-hot emotion ID and emotion tokens during the inference stage. Since this paper focuses on cross-speaker emotion transfer based on disjoint databases, it is worth considering how to deal with the situation where the target speaker has no emotion annotations. Instead of regarding all emotions of target speaker’s speech as neutral [17], we treat it as a semi-supervised learning problem. The emotion classifier loss of target speaker is not calculate and the model will softly determine what emotions each speech contains.

Overall, the training objective of the proposed method are shown in Eq. 1~Eq. 2.

\[
L_{cc} = - \sum_{i \in s_s} e_i \log(\hat{e}_i) \tag{1}
\]

\[
L = \sum_{i \in K} L_{reco}^i + \alpha L_{cc} + \beta L_{dur} \tag{2}
\]

where \(L_{reco}^i\) is the reconstruction loss of the \(i\)-th decoder stack, \(K\) is the number of decoder stacks, \(L_{dur}\) is the duration loss, \(L_{cc}\) is the emotion classifier loss, \(s_s\) denotes to source speaker, \(\alpha\) and \(\beta\) is the loss weight of emotion classifier loss and duration loss respectively.
3. EXPERIMENTS

3.1. Experimental Setup

Two internal Chinese speech databases from two male speakers are utilized in our experiments. One is a multi-emotion speech database with 7-emotion annotations (refer as source speaker), and the other is an audio-book database (refer as target speaker). The source speaker database contains 7-emotion annotations (800 utterances in each) and is 8.32 hours in total. The target speaker database contains 6778 utterances, and the total duration is 8.61 hours. 80-dimensional mel-frequency spectrograms are extracted with 10ms frame shift and 50ms frame length. We split 50 utterances in each corpus for the test. To evaluate the performance of the proposed method, the following models are constructed for further comparisons.

- **baseline**: A parallel tacotron-based multi-reference emotion transfer model using the paired-unpaired training strategy and adversarial cycle consistency scheme proposed by Whitehill et al. [17]. The emotion embeddings and the speaker embeddings generated by reference encoders are concatenated with encoder outputs.

- **proposed**: Proposed model describe in Sec. 2

- **M1**: An ablation model which removes the SCLN module in decoder LConv blocks and the speaker embeddings are added to encoder outputs.

- **M2**: An ablation model which removes emotion classifier loss in the training stage, multi-head attention is utilized in this model.

In the proposed and M1, seven 256-dimentional emotion tokens are pre-defined, and $\alpha$ and $\beta$ in Eq. 2 are both set to 0.1. As for M2, tokens and heads are set to 10 and 4, respectively. Speaker IDs are mapped into 64-dimensional vectors with a speaker lookup table. The baseline is implemented following the setup of [17] except the change of back-bone. All models are trained with 32 batch size for 200k steps. WaveRNN [8] is used in our experiments as the vocoder.

3.2. Results and Analysis

In this paper, we conduct three types of subjective evaluations to compare the cross-speaker emotion transfer performance of different models. All the samples are synthesized using unseen texts.

**Timbre similarity**: Participants are given synthesized speech of target speaker and two original recordings from source speaker and target speaker respectively. They are asked to give a 1~5 score with 0.5 interval for the timbre similarity between each synthesized speech and recordings. Higher score means higher similarities with the target speaker’s timbre. In our experiments, 70 utterances (10 for each emotion) are synthesized for each model, and 15 participants conduct this evaluation. Furthermore, utterances in the same group are shuffled and the model information is invisible to participants.

**Stability comparison feedback**: Given the synthesized speech from different models, participants are asked to give feedback if there are stability problems such as missing words, speaking rate problem, blurred speech, and pronunciation defect. The results are the percentages of synthesized speech with stability problems in sentence level. In this evaluation, 300 utterances (60 for neutral and 40 for the rest of emotions) are synthesized from each model and one linguistic expert conducts this evaluation. The model information is invisible to participant.

**Emotion perceive preference**: It is carried out in the form of ABX test. Speech are synthesized from both models with the same emotion labels and texts. Participants are asked to determine which utterance is perceived closer to the description of the emotion label. The two utterances with the same text and emotion label are scored in parallel with the random order, and model information is invisible to participants. The set of synthesized speech used in this evaluation are the same as in the timbre similarity evaluation.

3.2.1. Comparison with baseline

In order to compare with the baseline, three subjective evaluations aforementioned and an objective timbre similarity evaluation are conducted. The results are shown in Table 1~Table 3 and Fig. 2.

Table 1 shows that the proposed method outperforms the baseline in terms of subjective timbre similarity. Especially, the baseline gets extremely lower scores in sad, angry, surprise and scare. To objectively compare the timbre similarity of baseline and the proposed method, we randomly select 300 recordings from target speaker and source speaker, and then extract the 1024-dimensional utterance-level speaker verification (SV) embeddings of both recordings and synthesized speech using a pre-trained SV model. The SV embeddings are reduced to 2-dimensional vectors using t-SNE [22] and are plotted in Fig. 2(a) and Fig. 2(b). As shown in these figures, the synthesized clusters of the proposed method are closer to the target speaker than baseline, indicating a better objective performance in timbre similarity. Fig. 2(a) shows that the neutral cluster is closer to the target speaker cluster, and the angry cluster is closer to the source speaker cluster. It means that the baseline performs better in neutral and worse in angry, which highly matched the subjective results in Table 1. The objective results again prove that the proposed method performs better than the baseline in the timbre similarity. Moreover, it observes that there are several relatively separated clusters in the recordings of the source speaker, proving that even for the same speaker, his/her tim-
Table 1. Subjective timbre similarity evaluation results of different model, with confidence intervals of 95%. The higher value means the better timbre similarity and the bold indicates the best performance in all the models.

| emotions       | baseline   | M1        | M2        | proposed   |
|----------------|------------|-----------|-----------|------------|
| neutral        | 4.05 ± 0.14| 3.19 ± 0.13| 3.91 ± 0.15| 3.50 ± 0.13|
| happy          | 3.37 ± 0.13| 4.03 ± 0.05| 3.77 ± 0.13| 4.00 ± 0.07|
| sad            | 3.00 ± 0.13| 2.89 ± 0.09| **3.96 ± 0.16**| 3.36 ± 0.09|
| angry          | 2.96 ± 0.06| 3.21 ± 0.09| 3.33 ± 0.11| **3.67 ± 0.10**|
| surprise       | 3.16 ± 0.19| 3.91 ± 0.09| 3.85 ± 0.09| 3.76 ± 0.15|
| scare          | 3.11 ± 0.11| 3.35 ± 0.07| **3.52 ± 0.09**| 3.48 ± 0.12|
| hate           | 3.33 ± 0.22| 4.02 ± 0.11| 4.01 ± 0.12| 3.85 ± 0.13|
| average        | 3.28 ± 0.10| 3.51 ± 0.11| **3.76 ± 0.07**| 3.66 ± 0.07|

Table 2. Stability comparison feedback error rates of different models (%), because one utterance may contain multiple stability problems, the total may is not equal to the sum of the first 4 rows. The lower value means the better stability performance and the bold indicates the best performance in all the models.

| stability problems | baseline | M2 | proposed | p  |
|--------------------|----------|----|----------|----|
| missing words      | 0.67     | 0.33| 0.67     |    |
| speaking rate      | 6.67     | 6.00| **4.67** |    |
| blurred speech     | 25.67    | 20.67| **10.00**|    |
| pronunciation defect| 5.67   | **2.67**| 6.00     |    |
| total              | 32.67    | 27.30| **18.33**|    |

Table 3. Average preference scores (%) of the emotion perceive evaluations, where N/P stands for “no preference”, and p denotes the p-value of a t-test between two models. The higher value means stronger preference.

| baseline | M2       | proposed | N/P | p     |
|----------|----------|----------|-----|-------|
| 40.57    | -        | **53.24**|       | 6.19  | <0.01 |
| -        | 46.10    | **46.57**| 7.33 | 0.87  |
5. REFERENCES

[1] Yuchen Fan, Yao Qian, Feng-Long Xie, and Frank K. Soong, “TTS synthesis with bidirectional LSTM based recurrent neural networks,” in INTERSPEECH 2014, pp. 1964–1968.

[2] Jose Sotelo, Soroush Mehri, Kundan Kumar, João Felipe Santos, and Kyle Kastner et al., “Char2wav: End-to-end speech synthesis,” in ICLR 2017.

[3] Yuxuan Wang, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, and Ron J. Weiss et al., “Tacotron: Towards end-to-end speech synthesis,” in Interspeech 2017, 2017, vol. 70, pp. 195–204.

[4] Sercan Ömer Arik, Mike Chrzanowski, Adam Coates, Gregory Frederick Diamos, and Andrew Gibiansky et al., “Deep voice: Real-time neural text-to-speech,” in ICML 2017, 2017, vol. 70, pp. 2415–2424.

[5] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, and Navdeep Jaitly et al., “Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions,” in ICASSP 2018, pp. 4779–4783.

[6] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, and Oriol Vinyals et al., “Wavenet: A generative model for raw audio,” in The 9th ISCA Speech Synthesis Workshop. 2016, p. 125, ISCA.

[7] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, and Shubham Jain et al., “Samplernn: An unconditional end-to-end neural audio generation model,” in ICLR 2017.

[8] Nal Kalchbrenner, Erich Elsen, and Karen Simonyan et al., “Efficient neural audio synthesis,” in ICML 2018, vol. 80, pp. 2415–2424.

[9] Shumin An, Zhenhua Ling, and Lirong Dai, “Emotional statistical parametric speech synthesis using Istm-rnns,” in APSIPA ASC 2017, pp. 1613–1616, IEEE.

[10] Younggun Lee, Azam Rabiee, and Soo-Young Lee, “Emotional end-to-end neural speech synthesizer,” CoRR, vol. abs/1711.05447, 2017.

[11] R. J. Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, and Daisy Stanton et al., “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” in ICML 2018, pp. 4700–4709.

[12] Yuxuan Wang, Daisy Stanton, Yu Zhang, R. J. Skerry-Ryan, and Eric Battenberg et al., “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in ICML 2018, pp. 5167–5176.

[13] Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo, “Expressive speech synthesis via modeling expressions with variational autoencoder,” in Interspeech 2018, pp. 3067–3071, ISCA.

[14] Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in ICASSP 2019, pp. 6945–6949, IEEE.

[15] Yanyao Bian, Changbin Chen, Yongguo Kang, and Zhenglin Pan, “Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis,” CoRR, vol. abs/1904.02373, 2019.

[16] Peng-Fei Wu, Zhen-Hua Ling, Li-Juan Liu, Yuan Jiang, Hong-Chuan Wu, and Lirong Dai, “End-to-end emotional speech synthesis using style tokens and semi-supervised training,” in APSIPA ASC 2019, pp. 623–627, IEEE.

[17] Matt Whitehill, Shuang Ma, Daniel J. McDuff, and Yale Song, “Multi-reference neural TTS stylization with adversarial cycle consistency,” in Interspeech 2020, 2020, pp. 4442–4446.

[18] Raza Habib, Soroosh Mariooryad, Matt Shannon, Eric Battenberg, and R. J. Skerry-Ryan et al., “Semi-supervised generative modeling for controllable speech synthesis,” in ICLR 2020.

[19] Tao Li, Xinsheng Wang, Qicong Xie, Zhichao Wang, and Lei Xie, “Controllable cross-speaker emotion transfer for end-to-end speech synthesis,” CoRR, vol. abs/2109.06733, 2021.

[20] Isaac Elias, Heiga Zen, Jonathan Shen, Yu Zhang, Ye Jia, Ron J. Weiss, and Yonghui Wu, “Parallel tacotron: Non-autoregressive and controllable TTS,” in ICASSP 2021, pp. 5709–5713.

[21] Mingjian Chen, Xu Tan, Bohan Li, Yanqing Liu, Tao Qin, Sheng Zhao, and Tie-Yan Liu, “Adaspeech: Adaptive text to speech for custom voice,” in ICLR 2021, 2021.

[22] Laurens Van der Maaten and Geoffrey Hinton, “Visualizing data using t-sne,” Journal of machine learning research, vol. 9, no. 11, 2008.