

FINE-GRAINED EMOTION STRENGTH TRANSFER, CONTROL AND PREDICTION FOR EMOTIONAL SPEECH SYNTHESIS

Yi Lei, Shan Yang, Lei Xie*

Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University, Xian, China

ABSTRACT

This paper proposes a unified model to conduct emotion transfer, control and prediction for sequence-to-sequence based fine-grained emotional speech synthesis. Conventional emotional speech synthesis often needs manual labels or reference audio to determine the emotional expressions of synthesized speech. Such coarse labels cannot control the details of speech emotion, often resulting in an averaged emotion expression delivery, and it is also hard to choose suitable reference audio during inference. To conduct fine-grained emotion expression generation, we introduce phoneme-level emotion strength representations through a learned ranking function to describe the local emotion details, and the sentence-level emotion category is adopted to render the global emotions of synthesized speech. With the global render and local descriptors of emotions, we can obtain fine-grained emotion expressions from reference audio via its emotion descriptors (for transfer) or directly from phoneme-level manual labels (for control). As for the emotional speech synthesis with arbitrary text inputs, the proposed model can also predict phoneme-level emotion expressions from texts, which does not require any reference audio or manual label.

Index Terms— text-to-speech, expressive speech synthesis, emotion strength, sequence-to-sequence

1. INTRODUCTION

Thanks to the rapid development of deep learning, speech synthesis has been significantly advanced [1, 2, 3]. Recently, with unified acoustic and duration modeling, sequence-to-sequence (seq2seq) based neural speech synthesis has achieved superior performance with extraordinary naturalness compared to the conventional methods [4, 5, 6, 7]. Since natural-sounding can be reasonably produced by current seq2seq-based speech synthesis models learned from a typical corpus with neutral speaking style, there have been increasing interests in how to deliver expressive speeches with these seq2seq models [8, 9, 10, 11]. As we know, human

speech is expressive in nature, with rich style expressions and subtle emotions [9].

To achieve expressive speech synthesis, a common solution is to learn style-related latent representations from reference audio [8, 9, 12, 13]. The goal is to make the synthesized speech to imitate the style of the reference audio, which can be treated as some kind of style transfer. Although it is possible to control the speaking style by analyzing and controlling the learned style representations [14, 9], it is still hard to choose a suitable control vector for arbitrary sentences. Besides, these methods usually encode the reference audio into a fixed-length style representation without explicitly considering the length and contents of the target synthesized speech. When aggregating such reference speech into a fixed-length embedding, essential temporal information may be lost. Thus they can only obtain a global or averaged style [12]. Undoubtedly, human speech contains subtle expressions at various granularities. For instance, phoneme-level prosody variations are important for expressive speech synthesis [15].

This paper proposes a fine-grained control and prediction approach for emotional speech synthesis – a typical case of expressive speech synthesis specifically focusing on emotion rendering. The emotional expressions in human speech are directly affected by their intentions, which leads to different emotion categories such as happy, angry and fear, or even different emotion strengths in each word or phoneme. For example, an angry speaker may particularly put a strengthened focus with strong intensity on ‘hate’ when speaking “I hate this!”. Therefore, in this paper, we treat the emotion category as a *global render* of speech, while the emotion strength in each word or phoneme is defined as a *local descriptor*. Modeling the global render is straight-forward: we directly adopt emotion embeddings to control the emotion category. So for the fine-grained emotional speech synthesis, the key problem is how to model and control the local emotion descriptors.

Since the explicit annotations of emotion strength are unavailable, we adopt a relative ranking method [16, 17] to represent the local emotion descriptors. In detail, we automatically learn the *relative attributes* of emotional speech compared to the neutral speech in the utterance level, where the attributes are proved to be strength-related [18]. In order to achieve fine-grained emotion control, we then extract the

*Corresponding author. This work was supported by the National Key Research and Development Program of China (No.2017YFB1002102).

emotional strength in the phoneme level from the learned ranking function. The phoneme-level local descriptors are then utilized along with text inputs to build the seq2seq acoustic model. Experiments show that it is easy to control the global emotion render as well as the local emotion descriptor of each phoneme from either speech or manual labels.

Besides, even though we could easily control the global render and local descriptor through the above approach, it still needs a method to automatically choose a suitable control vector for arbitrary text inputs during inference [10]. To build an easy-to-use fine-grained emotional speech synthesis model, we further introduce a prediction module in the acoustic model to predict the local descriptors. With such a prediction module, the acoustic model has the ability to produce natural emotional speech according to the contents of the input text. As another advantage, our model can also accept control vectors from reference speech or manual labels to conduct style transfer and control at the same time. The efficacy of the proposed approach is validated from experiments.

2. RELATED WORKS

Recently there are various attempts to model emotions or styles in speech synthesis [8, 9, 12, 14, 15, 19, 20]. The most straightforward way to conduct emotional speech synthesis is using explicit annotations or labels as global render to model expressions [19, 20]. But the global explicit constraints can only provide an “averaged” emotional voice [10], and it is also hard to flexibly control the local emotion variations or even global intensities. To continuously control the global emotion render, the method in [18] introduced a ranking function to represent the emotion strength of speech. In this way, emotional speech can be produced with different strengths. However, it can only control the global render, i.e., global strength, and manual instructions are also necessary to decide the strength of synthesized speech during inference.

To avoid explicit labels, the approach in [8] adopts a reference encoder to extract a latent prosodic representation for style imitation. But the performance of synthesized speech is directly affected by the choice of specific reference audio during inference. Based on the reference encoder method, Global Style Token (GST) is proposed to learn interpretable style embeddings in an unsupervised way [9]. With GST, the proposed model could imitate the style of reference audio and control the style of synthesized speech by choosing specific tokens. As for its application in emotional synthesis, the approach in [14] introduces an inter-to-intra distance ratio algorithm on the learned style tokens, which minimizes the intra-cluster embedding vectors and maximizes the inter-cluster ones. An interpolation technique is further proposed to control emotion intensity. But the above methods still require reference audio or manual labels to guide the generation process and have not explicitly considered controlling local subtle emotion expres-

sions.

Without the need for auxiliary feature during inference, TP-GST is proposed to predict expressive speaking style directly from text [10]. It uses a pre-trained GST model to extract global style tokens of each utterance in the training data, and there is a second task in the encoder to predict style tokens from texts. Again, the TP-GST method only models the global style render of synthesized speech and it is still hard to conduct fine-grained emotional control. In this paper, we focus on fine-grained emotional speech synthesis, which considers the global render and local emotion descriptors at the same time. After extracting granularized phoneme-level emotion strength through the ranking function, we model and predict the local emotion descriptors in a unified model. In this way, the model can generate fine-grained emotional speech directly from text without manual control, and we can also control the generation process through granularized constraints from manual labels or reference audio during inference.

3. PROPOSED MODEL

Fig. 1 shows the proposed fine-grained emotional speech synthesis framework for transfer, control and prediction. It shares similar architecture with Tacotron [4] and Tacotron2 [5], which is composed of a CBHG-based text encoder and an attention-based auto-regressive acoustic decoder to generate mel-spectrogram. As for the emotion expression modeling, the proposed model contains a flexible module to provide emotional information during speech generation, which is learned from text inputs (prediction) or extracted from reference audio (transfer) or manual labels (control). With the emotion expression module, we can conduct emotion transfer, control and prediction in a unified model.

3.1. Local emotion descriptors extraction

For flexible fine-grained emotion speech synthesis, the key aspect is how to represent the local emotion expressions. In our work, we aim to learn a phoneme-level emotion strength representation to conduct fine-grained emotion modeling. Since the phoneme-level emotion descriptors are not readily available and it is hard to annotate manually, we use the concept of relative attributes [17, 18] to learn the emotion strength of phonemes to achieve this goal.

Given two categories of data, the ranking function is to calculate the relative attribute of the data [17]. In this paper, we treat the emotion strength as an attribute of speech. Hence the ranking function aims to learn the relative difference of emotion strength between neutral speech and a kind of emotional speech (such as happy). With the learned ranking function, we can obtain the relative strength for unseen emotional speech, which we treat as local emotion descriptors in this paper.

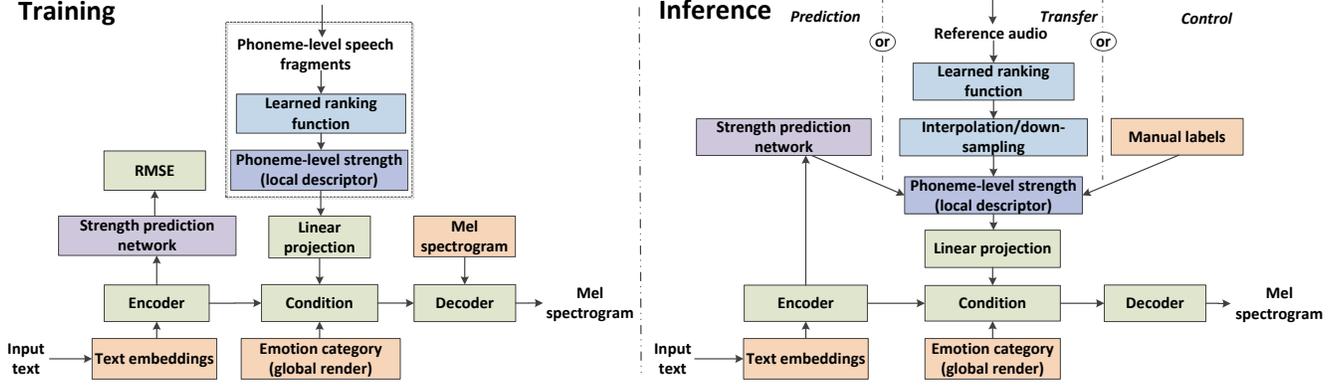


Fig. 1: System overview

Assuming the training set for learning the ranking function is $T = \{t\}$ represented in R^n by acoustic features $\{x_t\}$, and $T = N \cup H$, where N and H are the neutral and happy emotion set, respectively. The goal of relative attributes is to learn ranking function

$$r(x_t) = wx_t \quad (1)$$

satisfying the maximum number of the following constraints:

$$\begin{aligned} \forall (i, j) \in O : wx_i > wx_j \\ \forall (i, j) \in S : wx_i = wx_j \end{aligned} \quad (2)$$

where O and S are the ordered and similar sets respectively. It means that O is composed of sample pairs (i, j) with different categories, such as $i \in H$ and $j \in N$. And S contains sample pairs from the same category. This setting confirms that any sample pair from the ordered set O has the different emotion strength and any sample pair from the S has the similar emotion strength. We believe the emotion strengths of happiness are greater than neutral.

In order to learn the weighting matrix, Parikh *et. al.* [17] proposed to estimate the w by solving the following problem through Newton’s method [21]:

$$\begin{aligned} \text{minimize} \quad & \left(\frac{1}{2} \|w_m^T\|_2^2 + C \left(\sum \xi_{ij}^2 + \sum \gamma_{ij}^2 \right) \right) \\ \text{s.t.} \quad & w_m^T (x_i - x_j) \geq 1 - \xi_{ij}; \forall (i, j) \in O_m \\ & |w_m^T (x_i - x_j)| \leq \gamma_{ij}; \forall (i, j) \in S_m \\ & \xi_{ij} \geq 0; \gamma_{ij} \geq 0 \end{aligned} \quad (3)$$

where C is utilized to control the trade-off between the margin and the size of the slack variables ξ_{ij} and γ_{ij} .

To extract the phoneme-level emotion strength for fine-grained synthesis, we firstly train an HMM-based alignment model to obtain the learned phoneme boundaries of each utterance. With the learned ranking weights w and N speech fragments $\{x_i^1, x_i^2, \dots, x_i^N\}$ according to the boundaries, we could obtain the phoneme-level emotion strength from Eq. (1), which are treated as the local emotion descriptors. We finally normalize the emotion strength into $[0, 1]$ in each emotion category

for easy-to-use in emotion control. Fig. 2 shows an utterance with phoneme-level emotion strength. We can see that different phonemes have different local emotion strengths, although they all have the same utterance-level emotion render – happy.

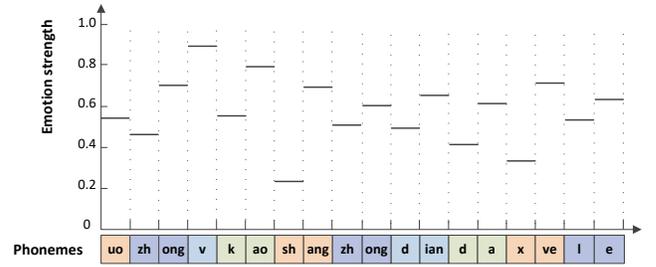


Fig. 2: Phoneme-level emotion strength representation of a “happy” utterance

3.2. Fine-grained emotion transfer and control

In the emotion transfer or control scenario, the generated speech is forced to perform the emotion expression of the reference audio or manual labels. With the above phoneme-level emotion strength representation and global emotion render, we firstly build an attention-based seq2seq model for emotional speech synthesis, as shown in the training part of Fig. 1. In this model, the emotion category embedding is treated as the global emotion render, and the phoneme-level strength represents the local emotion descriptors.

Considering emotion transfer, we can utilize a force-alignment model to obtain the phoneme boundaries of the target reference audio. Then the fragments of the reference audio can be adopted to compute the phoneme-level emotion strength through the learned ranking function. Since the phoneme number of reference audio is different from the input text during inference, we need to conduct linear interpolation or down-sampling to obtain the phoneme-level strength sequence whose length is the same as the text phoneme sequence. In details, assuming there are M phonemes in the

reference audio, we can simply construct a curve through the interpolation of M local emotion descriptors. When the input text contains N phonemes, we evenly split the above curve and treat the boundaries as the target local emotion descriptors.

As for emotion control, the model needs to perform according to manual instructions. Given the input phoneme sequence and the emotion category during inference, we can directly design the local emotion descriptors to satisfy our needs. That is to say, we can assign any value in $[0, 1]$ for each phoneme of the input text to control the generated audio into any trend of emotion expressions as needed. In this way, the proposed model can use the manual designed emotion labels (strength) to obtain fine-grained and flexible control.

3.3. Fine-grained emotion prediction

As discussed above, emotion transfer or control needs reference audio or manual labels to decide the emotional expressions of synthesized speech. But in practice, it is hard to find suitable reference audio or manually-designed emotion labels at the phoneme level. Thus in the proposed fine-grained emotion prediction module, we directly predict phoneme-level local descriptors from phoneme sequences. So the text encoder needs to provide content information for the acoustic decoder and predicted emotion strength information for each phoneme at the same time. As a result, the proposed model can produce natural emotional speech without any reference audio or manual label. As mentioned in section 3.2, we can also conduct emotion transfer and control with this unified model.

In detail, we feed the encoder output to a strength predictor, which has two fully-connected layers followed by the ReLU activation, to predict emotion strength. We minimize the differences between the predicted strength and the ground-truth strength extracted from the relative attributes ranking function. So the final objective of acoustic model is:

$$Loss = Loss_{mel} + \alpha Loss_{strength} \quad (4)$$

where $Loss_{mel}$ means the conventional L1 loss for acoustic modeling, and $Loss_{strength}$ is the L1 loss for emotion strength. α is a tunable weight during training.

During inference, our model will predict the phoneme-level emotion strength directly from text without any reference or label. And the predicted strength will decide the emotional expressions in the generated speech, given the emotion category as a global render. Since the phoneme-level emotion prediction module is relatively independent, we can also use phoneme-level emotion strength from reference audio or manual labels in the same model, which means that the proposed unified model is flexible for emotion transfer, control and prediction.

4. EXPERIMENTS

4.1. Basic setups

In our experiments, we use an internal high-quality emotional speech corpus, which contains about 14-hours of speech from a professional Chinese female speaker. The corpus consists of about 6000 sentences of neutral speech and six categories of emotional speech, including happy, angry, disgust, fear, surprise and sad, where each emotion category has about 600 sentences.

For text representation, we analyze the phone, tone and prosody boundary information through our text analysis module. We extract 80-band mel-scale spectrogram from speech as acoustic features. For both objective and subjective evaluation, we reserve 30 sentences of each emotional category to evaluate the performance of emotional style transfer and control. To reconstruct waveforms from mel-spectrogram, we build a multi-band WaveRNN [22] trained by ground-truth mel-spectrogram for fair comparisons. There are 20 native Chinese speakers taking part in the subjective evaluation. And for the objective evaluation, dynamic time warping (DTW) is adopted to align predicted features and target features.

4.2. Model details

For the fine-grained local emotion descriptors, we firstly extract 384-dimensional emotion-related features from speech using the openSMILE tool [23] and learn the ranking function using the MATLAB codes provided by Parikh *et al.* [17]. We train an HMM-based aligner using the same corpus to obtain the phoneme boundary of each utterance. Finally, we utilize the phoneme-level speech fragments to compute the emotion strength as local emotion descriptors through the learned ranking function. The extracted phoneme-level descriptors are finally normalized into $[0,1]$.

Table 1: MCD of different models for parallel transfer

Method	MCD (dB)
GST	4.89
UET	5.16
proposed-FET	4.91

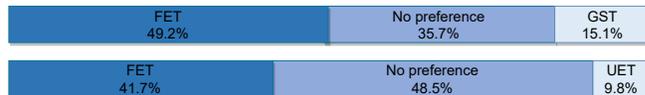


Fig. 3: A/B preference test result for parallel transfer

In the emotional speech synthesis model, we adopt three feed-forward layers as pre-net followed by a CBHG module as text encoder [4]. For the fine-grained emotion control, we

FET 42.2%	No preference 27.8%	GST 30%
FET 39.8%	No preference 32.3%	UET 27.9%

Fig. 4: A/B preference test for non-parallel transfer

concatenate the emotional category embedding with the encoder outputs as global emotion render, while the phoneme-level emotion descriptors are added to the encoder outputs through a linear projection layer with 512 units. As for the emotion prediction, the encoder outputs are fed into the prediction network to predict the phoneme-level emotion strengths. The auto-regressive decoder also contains three feed-forward layers as pre-net and a 2048 units unidirectional LSTM layer. We adopt the robust GMM attention [24] to connect the encoder and the decoder. Finally, the CBHG-based post-net is used to produce mel-spectrogram and waveform is generated through multiband WaveRNN.

4.3. Experimental results

Through our proposed approach, we can conduct fine-grained emotion transfer, control and prediction in a unified model. Hence we will evaluate the performance of the proposed model in the three aspects.

4.3.1. Fine-grained emotion transfer

We first evaluate the ability of emotion transfer for the proposed fine-grained emotion transfer (FET). We train the GST model [9] and the utterance-level emotion transfer model [18] (UET) as baseline systems for comparison. We evaluate the performance of both parallel and non-parallel transfer. For the parallel transfer, the reference audio has the same text content as the target text to be synthesized. In this scenario, the mel-spectrogram of reference audio is fed into the reference encoder of the GST model, and we extract utterance-level and phoneme-level emotion strength of reference audio for the UET model and proposed FET model respectively. As for the non-parallel transfer, where the target text is not necessarily the same as that of the reference audio, the reference audios are randomly selected from the test set to conduct emotion transfer.

Table 1 shows the mel-cepstral distortion (MCD) of different models for parallel transfer. The results indicate that the GST model and the proposed FET model have apparently lower MCD values as compared with the utterance-level emotional strength model. The proposed FET model has a close MCD value with the GST model. Since the goal of emotion transfer is to imitate the emotion of the reference speech, we also conduct A/B preference test to let listeners choose which one is more similar to the reference in emotion expressions, as shown in Fig. 3. Comparing the proposed FET with the GST model, we find that with much more preferred, the

FET model can imitate the local emotional expressions better than the GST model which only can learn an “averaged” emotional embedding from reference. Similarly, the performance of phoneme-level imitation (FET) is much better than the utterance-level transfer (UET).

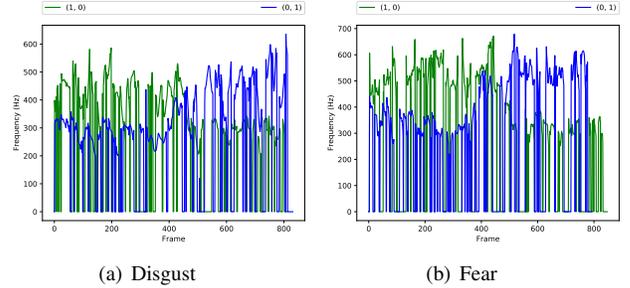
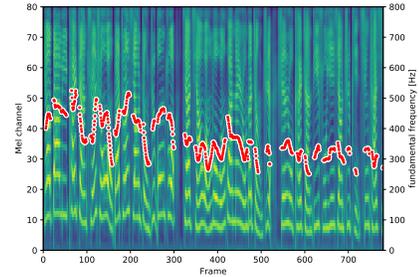
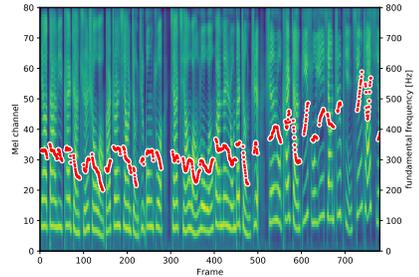


Fig. 5: F0 curves of synthetic samples of different global render and local descriptors. (0,1) means that the local descriptors of the first half phonemes is 0, and the last half is 1.



(a) Gradually decreased strength



(b) Gradually increased strength

Fig. 6: Mel spectrograms and F0 of synthetic samples with gradually changed emotion strength.

FEP 20%	No preference 52.3%	FET 27.7%
------------	------------------------	--------------

Fig. 7: A/B preference for FEP and FET

We also evaluate the performance of non-parallel transfer for the above models. The results of A/B preference test are shown in Fig. 4. As for the non-parallel transfer, we find the proposed FET model also outperforms the GST and UET models.

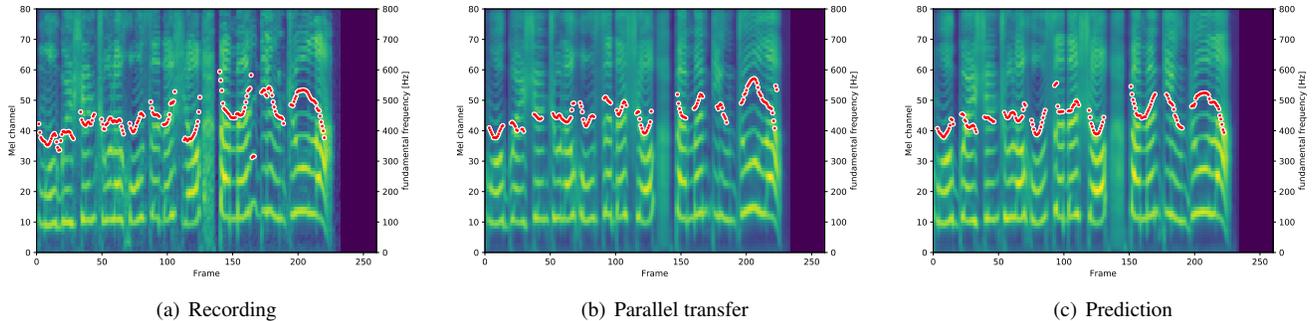


Fig. 8: Mel spectrograms and F0 from recording, parallel emotion transfer and emotion prediction model.

4.3.2. Fine-grained emotion control

Another important aspect of emotional speech synthesis is how to flexibly control the emotional expressions of synthesized speech. For the GST model, we can choose or weigh different tokens to roughly control the speech generation. And the UET model can also control the emotion render with different scales, but only at the whole-utterance level. By contrast, the proposed FET model has the ability of fine controls for both global render and local descriptor. To show such ability, we generate utterances with different global render (emotion category) and local descriptors (phoneme-level emotion strength). Fig. 5 shows two examples of F0 trajectory, with a “disgust” render and a “fear” render. Note that F0 is directly related to emotion strength. In each emotional render, we split the phoneme sequence of the text into two parts, where the emotion strengths of the phonemes in each part are set as either 0 or 1. The F0 trajectories in Fig. 5 show that the utterance part whose emotion strengths are 1 have much higher F0 in the generated speech as compared to the part with strengths of 0. Besides, we also set the strength of the input phoneme sequence gradually decrease from 1 to 0 and gradually increase from 0 to 1 for obtaining gradual strength change. Fig. 6 shows that the F0 trajectory changes gradually as our control. Subjective listening on these samples also indicates that fine-grained emotion strength change can be easily detected. We suggest the readers listen to our online demos¹.

4.3.3. Fine-grained emotion prediction

For the above emotion transfer and control, we still need auxiliary information from reference audio or manual setup to control the emotional expressions. We finally evaluate the proposed fine-grained emotion prediction model (FEP), which directly predicts phoneme-level local emotion strengths from text. Given the emotion category, we generate fine-grained emotional speech with the predicted local emotion strength and calculate the MCD with the target emotional speech, as shown in Table 2. From the results,

¹Samples can be found at https://leiyi420.github.io/pho_ra_strength/

we can find that the generated speech with the ground-truth phoneme strength (FET) has slightly lower MCD than the one using predicted strength (FEP), but the difference is not big. We also conduct A/B test on the FET and FEP systems, as shown in Fig. 7. The subjective result indicates that there is no significant difference between generated speech with predicted strength and ground-truth strength, which shows the ability of the proposed FEP model to accurately predict the local emotional descriptors.

Table 2: MCD of emotion prediction and parallel transfer

Method	MCD (dB)
proposed-FET	4.91
proposed-FEP	5.03

In order to intuitively see the difference between the predicted emotion expression and parallel emotion transfer, we also analyze the mel-spectrogram and F0 of the generated speech from both models. Fig. 8 shows the generated examples from the parallel transfer method, prediction method and real recording. We can find that the synthesized speech using predicted phoneme-level strength can reconstruct satisfied emotion expressions like parallel emotion transfer, even though there are some differences in a few units. The results prove that the proposed method can predict similar emotion expression to both parallel transfer and recordings directly from arbitrary text, which does not need reference audio or manual interventions.

5. CONCLUSIONS

This paper proposed a unified model to conduct emotion transfer, control and prediction for fine-grained emotional speech synthesis. With a sentence-level global render of the emotion category and the phoneme-level local descriptors of the emotion strength learned from a ranking function, the proposed model can transfer details of emotion intensities from reference audio and synthesize speech from manual labels to control emotion expressions. In addition, it can also predict phoneme-level emotional expressions directly from texts. Experimental results show that the proposed method can transfer, control and predict the fine-grained emotion ex-

pression in a unified model, which outperforms the baseline systems with coarse emotional expressions and also improves the flexibility of emotional speech synthesis model.

6. REFERENCES

- [1] Heiga Zen, Andrew Senior, and Mike Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. ICASSP*. IEEE, 2013, pp. 7962–7966.
- [2] Yao Qian, Yuchen Fan, Wenping Hu, and Frank K Soong, “On the training aspects of deep neural network (dnn) for parametric tts synthesis,” in *Proc. ICASSP*. IEEE, 2014, pp. 3829–3833.
- [3] Zhen-Hua Ling, Li Deng, and Dong Yu, “Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis,” *IEEE transactions on audio, speech, and language processing*, vol. 21, no. 10, pp. 2129–2139, 2013.
- [4] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, et al., “Tacotron: Towards end-to-end speech synthesis,” in *Proc. INTERSPEECH*, 2017, pp. 4006–4010.
- [5] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., “Natural tts synthesis by conditioning wavenet on mel spectrogram predictions,” in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.
- [6] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, “Neural speech synthesis with transformer network,” in *Proc. AAAI*, 2019, vol. 33, pp. 6706–6713.
- [7] Shan Yang, Heng Lu, Shiyin Kang, Liumeng Xue, Jinba Xiao, Dan Su, Lei Xie, and Dong Yu, “On the localness modeling for the self-attention based end-to-end speech synthesis,” *Neural Networks*, vol. 125, pp. 121–130, 2020.
- [8] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous, “Towards end-to-end prosody transfer for expressive speech synthesis with tacotron,” *arXiv preprint arXiv:1803.09047*, 2018.
- [9] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” *arXiv preprint arXiv:1803.09017*, 2018.
- [10] Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan, “Predicting expressive speaking style from text in end-to-end speech synthesis,” in *Proc. SLT*. IEEE, 2018, pp. 595–602.
- [11] 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 *Proc. ICASSP*. IEEE, 2019, pp. 6945–6949.
- [12] Yanyao Bian, Changbin Chen, Yongguo Kang, and Zhenglin Pan, “Multi-reference tacotron by intercross training for style disentangling, transfer and control in speech synthesis,” *arXiv preprint arXiv:1904.02373*, 2019.
- [13] Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi, “Deep encoder-decoder models for unsupervised learning of controllable speech synthesis,” *arXiv preprint arXiv:1807.11470*, 2018.
- [14] Se-Yun Um, Sangshin Oh, Kyungguen Byun, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang, “Emotional speech synthesis with rich and granularized control,” in *Proc. ICASSP*. IEEE, 2020, pp. 7254–7258.
- [15] Younggun Lee and Taesu Kim, “Robust and fine-grained prosody control of end-to-end speech synthesis,” in *Proc. ICASSP*. IEEE, 2019, pp. 5911–5915.
- [16] Vittorio Ferrari and Andrew Zisserman, “Learning visual attributes,” in *Proc. NeurIPS*, 2008, pp. 433–440.
- [17] Devi Parikh and Kristen Grauman, “Relative attributes,” in *Proc. ICCV*. IEEE, 2011, pp. 503–510.
- [18] Xiaolian Zhu, Shan Yang, Geng Yang, and Lei Xie, “Controlling emotion strength with relative attribute for end-to-end speech synthesis,” in *Proc. ASRU*. IEEE, 2019, pp. 192–199.
- [19] Jaime Lorenzo-Trueba, Gustav Eje Henter, Shinji Takaki, Junichi Yamagishi, Yosuke Morino, and Yuta Ochiai, “Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis,” *Speech Communication*, vol. 99, pp. 135–143, 2018.
- [20] Younggun Lee, Azam Rabiee, and Soo-Young Lee, “Emotional end-to-end neural speech synthesizer,” *arXiv preprint arXiv:1711.05447*, 2017.
- [21] Olivier Chapelle, “Training a support vector machine in the primal,” *Neural computation*, vol. 19, no. 5, pp. 1155–1178, 2007.
- [22] Chengzhu Yu, Heng Lu, Na Hu, Meng Yu, Chao Weng, Kun Xu, Peng Liu, Deyi Tuo, Shiyin Kang, Guangzhi

Lei, et al., “Durian: Duration informed attention network for multimodal synthesis,” *arXiv preprint arXiv:1909.01700*, 2019.

- [23] Florian Eyben, Martin Wöllmer, and Björn Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proc. ACMMM*, 2010, pp. 1459–1462.
- [24] Eric Battenberg, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby, “Location-relative attention mechanisms for robust long-form speech synthesis,” in *Proc. ICASSP*. IEEE, 2020, pp. 6194–6198.