End-to-End Emotional Speech Synthesis Using Style Tokens and Semi-Supervised Training

Pengfei Wu*, Zhenhua Ling*, Lijuan Liu†, Yuan Jiang†, Hongchuan Wu† and Lirong Dai*

* University of Science and Technology of China, Hefei, China

E-mail:wpf0610@mail.ustc.edu.cn, {zhling, lrdai}@ustc.edu.cn

† iFLYTEK Research, iFLYTEK Co., Ltd., Hefei, P.R.China

E-mail:{ljliu, yuanjiang, hcwu4}@iflytek.com

Abstract—**This paper proposes an end-to-end emotional speech synthesis (ESS) method which adopts global style tokens (GSTs) for semi-supervised training. This model is built based on the GST-Tacotron framework. The style tokens are defined to present emotion categories. A cross entropy loss function between token weights and emotion labels is designed to obtain the interpretability of style tokens utilizing the small portion of training data with emotion labels. Emotion recognition experiments confirm that this method can achieve one-to-one correspondence between style tokens and emotion categories effectively. Objective and subjective evaluation results show that our model outperforms the conventional Tacotron model for ESS when only 5% of training data has emotion labels. Its subjective performance is close to the Tacotron model trained using all emotion labels.**

Keywords: emotional speech synthesis, end-to-end, Tacotron, global style tokens, semi-supervised training

I. INTRODUCTION

The text-to-speech (TTS) methods based on deep learning techniques have developed rapidly in the past few years. They can improve the naturalness of synthetic speech significantly compared with the traditional approaches. Zen et.al. [1] made the first attempt of building acoustic models using deep neural networks (DNNs), which can model the complex dependencies between linguistic information and acoustic features. With similar number of parameters, it achieved better naturalness than hidden Markov model (HMM)-based one. Since it was difficult to incorporate long time span contextual effects into DNNs [2], recurrent neural network (RNN)-based methods [3] were then proposed to better capture temporal information. In recent years, the audio generative models that can generate high quality natural-sounding speech [4] and the end-to-end acoustic models which predicted acoustic features directly from phoneme or grapheme sequences [5], [6] have also been proposed and attracted the attention of many researchers.

How to generate speech with expected emotions is also an important topic in TTS [7]. At the early stage of emotional speech synthesis (ESS) research, the rule-based methods were studied which controlled the speech parameters for emotion expression through manually-defined rules [8], [9]. Although this approach can control a large number of parameters related with both voice source and vocal tract [10], the resulting speech sounded unnatural and the emotion discriminability was not satisfactory. The unit selection-based methods for ESS have also been studied. However, obtaining a large-scale emotional speech corpus for unit selection is usually difficult and costly. Compared with unit selection, statistical parameter speech synthesis (SPSS) requires much smaller speech corpus and is far more flexible. A lot of studies have been made to build acoustic models for ESS using hidden Markove models [11], [12] and neural networks [13], [14], [15]. In [13], emotion-dependent and emotion-independent acoustic models based on RNNs with long short-term memory (LSTM) units were investigated. The end-to-end Tacotron model has also been applied to ESS in [14].

All of the ESS methods mentioned above rely on supervised training, i.e., all training or adaptation utterances have explicit emotion labels. This condition is easy to satisfy when the datasets are built based on the performed emotions of actors. However, for spontaneous expressive speech, to annotate emotion labels for all utterances is difficult and costly. Therefore, some unsupervised approaches [16], [17], [18] of expressive speech synthesis using neural networks have been proposed recently. Although these methods can learn speech variations without explicit annotations for expressiveness, the learned representations or latent embeddings were usually not fully controllable and had poor interpretability. As mentioned in [17], one token may capture multiple attributes in speech. Besides, the attributes that the learned tokens correspond to can't be determined at the training stage.

In this paper, we propose a semi-supervised training method using global style tokens (GSTs) [17] for ESS aiming at the condition that only a small portion of training data has emotion labels. This model is built based on the GST-Tacotron framework. A bank of tokens with the same number as emotion categories are defined and a cross entropy loss is introduced between the token weights and the emotion labels in order to establish a one-to-one correspondence between tokens and emotions. The model parameters are estimated by multi-task learning when only the emotion labels of a few training samples are available. In our experiments, we found that this method can achieve the interpretability of tokens and the accurate recognition of emotion labels. When only 5% of training data had emotion labels, our model achieved similar subjective performance with the Tacotron-based ESS model with fully supervised training.
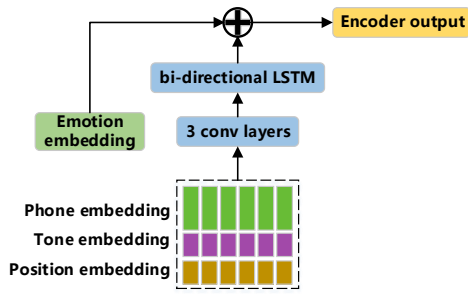
Fig. 1. The architecture of the encoder using in our baseline Tacotron model.

Fig. 2. The model architecture of GST-Tacotron with semi-supervised training.

II. RELATED WORK

A. End-to-end speech synthesis with style tokens

Recently, end-to-end acoustic models with style tokens for speech synthesis have been proposed [16], [17], [18]. A Tacotron-based end-to-end speech synthesis architecture that learnt a latent embedding space of prosody was proposed in [16]. In this model, a reference encoder was defined to encode the prosody of reference speech into a fixed-length vector which contained the information not provided by the text and speaker identity. The experimental results demonstrated that this encoder can transfer prosody between utterances in an almost speaker-independent fashion. An unsupervised style modeling, control and transfer method was proposed in [17]. It extended Tacotron by adding a style token layer consisting of a bank of embeddings named global style tokens in it. The embeddings were trained without explicit labels, and were learnt to model the expressiveness-related acoustic variations in speech.

B. Neural network-based ESS

LSTM-RNN models for ESS were studied in [13]. Two modeling approaches, emotion-dependent modeling and unified modeling with emotion codes, were designed. Experimental results showed that both approaches achieved significant better naturalness of synthetic speech than HMM-based emotion-dependent modeling. Emotion control can be achieved by interpolating or extrapolating the emotion codes in the unified model. An end-to-end ESS model was introduced in [14], which focused on modifying Tacotron to get better alignment. This paper instead studies how to build an end-to-end ESS model with limited emotion labels using GSTs.

III. METHODS

A. Tacotron-based acoustic model for ESS

This paper builds end-to-end ESS models based on Tacotron, and refers to an open source implementation[1] of Tacotron 2 [6]. Since our emotional speech corpus is recorded in Chinese, some modifications are made to the encoder module of Tacotron, as shown in Fig. 1. Instead of using Chinese character sequences directly, phoneme and tone sequences given by text analysis are adopted as input. Here, phonemes
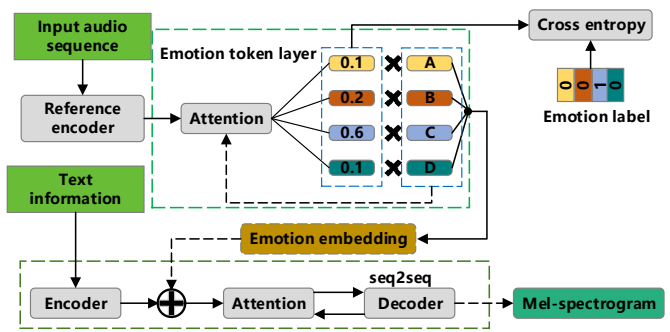
---

[1]https://github.com/Rayhane-mamah/Tacotron-2.

are defined as the initials and finals of Chinese, and the tone of each phoneme is defined as the tone of the syllable in which current phoneme is located. Besides, in order to get better break and rhythm performance, a feature which indicates the position of each phoneme in current sentence is added. Therefore, for each phoneme, its representation is made up by concatenating a phone embedding, a tone embedding and a position embedding.

In order to build our Tacotron-based baseline models for ESS, we implement two approaches of emotional modeling, i.e., emotion-dependent (ED) modeling which trains a model for each emotion separately and emotion-independent (EI) modeling that trains a unified model for all emotions. In the EI model, an emotion embedding vector is added to the output of text encoder as shown in Fig. 1.

B. GST-Tacotron with semi-supervised training for ESS

This paper proposes a semi-supervised training method for Tacotron-based ESS. This method adopts almost the same model structure as GST-Tacotron, including an emotion token layer and a Chinese Tacotron sequence-to-sequence module, as shown in Fig. 2. The emotion token layer converts log-mel spectrogram into a fixed-length emotion embedding by two steps. First, a reference encoder is utilized to compress the input log-mel spectrogram into a fixed-length vector. Then, the vector is passed to the emotion token layer as the query vector for attention and to calculate a set of weights which measure its similarity to each token. The weighted sum of tokens forms the emotion embedding which is added to the encoder output for every time step as shown in Fig. 1.

As introduced above, conventional GST-Tacotron models lack interpretability of the learnt tokens. In order to guarantee that each learned token corresponds to one emotion in our method, several modifications are made to the conventional GST-Tacotron model. First, single-head attention is used instead of multi-head attention in the emotion token layer to achieve a single-vector representation of token weights. Second, the number of tokens is set to be consistent with the number of emotion categories in the corpus. Finally, one-hot emotion labels are introduced into the emotion token layer as the target of token weights. A cross entropy loss between the emotion labels and the token weights is added into the loss

TABLE I
DETAILS OF THE CORPUS USED IN OUR EXPERIMENTS.

emotion	train		test	
	duration (min.)	# of utt.	duration (min.)	# of utt.
neutral	84.68	759	3.20	30
happy	62.86	926	2.02	30
sad	87.33	882	3.02	30
angry	23.71	370	2.07	30
total	258.57	2937	10.30	120

TABLE II
OBJECTIVE EVALUATION RESULTS OF ED AND EI MODELS.

Model	Emotion	MCD(dB)	F_0RMSE(Hz)	V/UV(%)	FFE(%)
ED	neutral	2.75	57.54	7.94	20.00
	happy	2.65	73.87	9.06	25.14
	sad	2.61	53.99	6.95	17.24
	angry	2.69	77.88	10.90	28.68
EI	neutral	2.66	58.19	7.20	20.19
	happy	2.63	74.59	8.53	26.30
	sad	2.54	51.47	6.82	15.48
	angry	2.61	76.94	10.34	27.31

function of the GST-Tacotron model as follows,

$$loss = \sum_s L(\mathbf{c}, \hat{\mathbf{c}}) + \sum_{s'} CE(\mathbf{e}, \hat{\mathbf{e}}), \qquad (1)$$

where $\mathbf{c}$ and $\hat{\mathbf{c}}$ denote the targets and predictions of an utterance in the training set, $\mathbf{e}$ and $\hat{\mathbf{e}}$ represent the emotion label vector and the token weight vector of a training utterance respectively, $L(\cdot)$ and $CE(\cdot)$ are Tacotron loss and cross entropy loss functions, s denotes the complete training set and s' denotes the part of training utterances with emotion labels.

At the synthesis stage, the emotion embedding is first determined by selecting the emotion token corresponding to the expected emotion. Then, the embedding vector is added to the output of text encoder at each time step for following attention-based decoding. For both the baseline method describe in Section III-A and our proposed method, a WaveNet vocoder [19] conditioned on mel-spectrogram is trained to reconstruct waveform samples from the output of Tacotron.

IV. EXPERIMENTS

A. Experimental conditions

A high-quality emotional speech corpus recorded in a professional studio was used in our experiments. The 3057 Chinese sentences of four emotions (neutral, happy, sad and angry) were uttered by a female speaker. The recordings were about 4.48 hours with 16kHz sampling rate and 16bits quantization. The details of the corpus are shown in Table I.

Based on the speech corpus above, four types of models were constructed for comparison.

- **ED** Four emotion-dependent Tacotron models using all emotion labels as introduced in Section III-A.
- **EI** An emotion-independent Tacotron model using emotion embeddings and all emotion labels as introduced in Section III-A. This model served as the upper-bound model for evaluating our proposed semi-supervised training method.
- **Semi-EI** Same as **EI**, but only using the emotion labels of a part of training utterances. For the utterances without emotion labels, a zero emotion embedding vector is used. This model served as the baseline model for evaluating our proposed semi-supervised training method.
- **Semi-GST** A GST-Tacotron model using the semi-supervised training method introduced in Section III-B.

The output acoustic features of all models were 80-dimensional mel-spectrogram. The dimensions of phone embedding, tone embedding and position embedding were 384,

64 and 64 respectively. The settings of Tacotron models and the emotion token layer in Semi-GST followed previous studies [6], [17]. The WaveNet neural vocoder had 24 dilated casual convolution layers which were divided into 4 convolution blocks. Each block contained 6 layers and their dilation coefficients were {1,2,4,8,16,32}. The output distribution for each waveform sample was single Gaussian. When building the ED models, it was difficult to get Tacotron models which can generate normal alignment between phoneme and acoustic feature sequences since the training data of a single emotion was limited. Therefore, a pre-training and adaptation strategy was utilized. First, a unified Tacotron model was estimated with 1000 epochs using all training data without emotion labels. Then, the data of each emotion was utilized to fine-tune the model with 100 epochs to get corresponding ED model. All the EI, Semi-EI and Semi-GST models were trained with 1000 epochs.

B. Comparison between ED and EI models

Objective evaluations were conducted to compare the performance of ED and EI models, which both followed the supervised training framework. The evaluation metrics were the average distortions of acoustic feature prediction, including mel-cepstrum distortions (MCD), root mean square error of F_0 (F_0RMSE), voiced/unvoiced decision error (V/UV) and F_0 frame error (FFE) [20], on the test set of each emotion. The FASTDTW [21] algorithm was adopted to align the predicted acoustic feature sequences towards the natural ones. Then, the distortions between them were calculated frame-by-frame.

The results are shown in Table II. From the table, we can see that in both ED and EI models, "neutral" and "sad" emotions obtained better results than "happy" and "angry" emotions on almost all metrics. It can be explained from two aspects. First, the later two emotions had higher and more variational F0 contours which were more difficult for modeling. Second, the "angry" emotion had much less training data as shown in Table I. Comparing ED and EI models, it can be observed that the EI model achieved slightly better results than ED for all emotions. This can be attributed to the advantage of unified modeling and data sharing.

C. Objective evaluation on our proposed method

First, we conducted an emotion recognition experiment to verify the interpretability of the tokens learned by the Semi-

TABLE III
CONFUSION MATRICES OF EMOTION RECOGNITION USING THE SEMI-GST MODELS TRAINED WITH DIFFERENT PROPORTIONS OF EMOTION LABELS. HERE, $\bar{w}$ STANDS FOR THE AVERAGE TOKEN WEIGHT OF THE TRUE EMOTION.

(a) 2% emotion labels

true label	recognized label				
	neutral	happy	sad	angry	$\bar{w}$
neutral	**26**	0	4	0	0.8554
happy	2	**21**	7	0	0.7387
sad	4	5	**21**	0	0.6553
angry	2	8	**20**	0	0.0336

(b) 5% emotion labels

true label	recognized label				
	neutral	happy	sad	angry	$\bar{w}$
neutral	**29**	1	0	0	0.9598
happy	0	**30**	0	0	0.9927
sad	0	1	**29**	0	0.9507
angry	0	0	0	**30**	1.0000

(c) 10% emotion labels

true label	recognized label				
	neutral	happy	sad	angry	$\bar{w}$
neutral	**30**	0	0	0	0.9890
happy	0	**30**	0	0	0.9999
sad	0	1	**29**	0	0.9667
angry	0	0	0	**30**	1.0000

(d) 20% emotion labels

true label	recognized label				
	neutral	happy	sad	angry	$\bar{w}$
neutral	**30**	0	0	0	1.0000
happy	0	**30**	0	0	1.0000
sad	0	0	**30**	0	0.9993
angry	0	0	0	**30**	1.0000

TABLE IV
OBJECTIVE EVALUATION RESULTS OF EI, SEMI-EI AND SEMI-GST MODELS.

Model	MCD(dB)	F_0RMSE(Hz)	V/UV(%)	FFE(%)
EI	2.61	63.77	7.98	21.66
Semi-EI	2.71	68.95	8.97	24.93
Semi-GST	2.64	64.04	8.26	21.81

TABLE V
AVERAGE PREFERENCE SCORES(%) AMONG EI, SEMI-EI AND SEMI-GST MODELS ON NATURALNESS OF SYNTHETIC SPEECH, WHERE N/P STANDS FOR "NO PREFERENCE" AND p DENOTES THE p-VALUE OF A t-TEST BETWEEN TWO MODELS.

Semi-GST	EI	Semi-EI	N/P	p
66.25	-	18.33	15.42	2.30×10^{-34}
35.00	38.96	-	26.04	0.31

and Semi-GST models. The metrics were the same as the ones used in Table II and the results averaged across all emotions are shown in Table IV. We can see that Semi-GST performed better than Semi-EI on all metrics. The performance of the Semi-GST model was very close to that of the EI model although it only utilized 5% emotion labels.

D. Subjective evaluation on our proposed method

To subjectively evaluate the performance of our proposed method on the naturalness and emotion expressiveness of synthetic speech[2], two groups of ABX preference tests and three emotion classification tests were conducted by 12 subjects. In both subjective tests, ten sentences were selected randomly for each emotion from the synthesized test set.

In ABX preference tests, the subjects were asked to give one of the three choices, (1) A is more natural, (2) no preference, (3) B is more natural, for each pair of synthetic speech played in random order. In addition to the average preference scores, the p-value of t-test was used to measure the significance of the difference between two models in each test. The results are shown in Table V. We can see that Semi-GST significantly outperformed Semi-EI ($p < 0.01$) and there was no significant difference between Semi-GST and EI ($p = 0.31$).

In emotion classification tests, the synthetic utterances of all emotions were played in random order. Each subject was asked to choose the emotion they perceived for each utterance. They were also allowed to chose the "other" option if they didn't think the utterance belonged to any of the the four emotions. The results are shown in Table VI. First, we can see that all these three models achieved the best accuracy for the "sad" emotion and the worst accuracy for the "angry" emotion. This is consistent with the objective evaluation results shown in Table II. Further, both EI and Semi-GST models performed better than the Semi-EI model. The average classification accuracies of the three models were 91.46% (EI), 86.25% (Semi-EI), and 91.88%(Semi-GST) respectively, which shows

GST model. Four Semi-GST models were trained which utilized 2%, 5%, 10% and 20% emotion labels in the training set respectively. In this experiment, we fed the 120 test utterances (30 per emotion) in the test set into the emotion token layer and got 120 token weight vectors. For each vector, the index of its maximum value was treated as the emotion recognition result of this utterance. Table III shows the confusion matrices of emotion recognition. The $\bar{w}$ value in Table III denotes the weight of the token corresponding to the true emotion and averaged across the 30 test utterances. We expect that all $\bar{w}$ values are close to 1 which indicates a one-to-one correspondence between the learnt tokens and the emotion categories.

From Table III, we can see that there were plenty of recognition errors when only 2% emotion labels were utilized. When increasing the proportion of available emotion labels to 5%, the confusion matrix appeared in a clear diagonal form and the $\bar{w}$ values became very close to 1. Further, there were no classification errors when using 20% emotion labels. These results reflect that the Semi-GST model can achieve the emotion-related interpretability of tokens with only 5% emotion labels.

Then, an experiment was conducted to compare the performance of EI, Semi-EI and Semi-GST models. In this experiment, 5% emotion labels were used for both Semi-EI

---

[2]The audio samples can be found at http://home.ustc.edu.cn/~wpf0610/APSIPA2019.html

TABLE VI
THE PERCENTAGES OF PERCEIVED EMOTIONS FOR SPEECH SYNTHESIZED
USING DIFFERENT MODELS.

(a) EI

true label	perceived label				
	neutral	happy	sad	angry	other
neutral	**91.67**	5.00	0	0.83	2.50
happy	12.50	**85.00**	0	0	2.50
sad	0	0	**100.00**	0	0
angry	5.83	2.50	0	**89.17**	2.50

(b) Semi-EI

true label	perceived label				
	neutral	happy	sad	angry	other
neutral	**87.50**	9.17	0.83	0.83	1.67
happy	5.83	**91.67**	0	0	2.50
sad	1.67	0	**96.67**	0.83	0.83
angry	9.17	15.00	0.83	**69.17**	5.83

(c) Semi-GST

true label	perceived label				
	neutral	happy	sad	angry	other
neutral	**92.50**	2.50	1.67	0.83	2.50
happy	4.17	**94.17**	0	0	1.67
sad	0	0	**100.00**	0	0
angry	8.33	5.83	0.83	**80.83**	4.17

that the Semi-GST model can express emotions as effectively as the EI model.

V. CONCLUSIONS

This paper have presented an end-to-end emotional speech synthesis method using style tokens and semi-supervised training. The model is implemented based on the GST-Tacotron framework. A cross-entropy loss function was introduced to achieve the emotion-related interpretability of style tokens and the semi-supervised training using a small portion of emotion labels in the training data. Experimental results confirmed the effectiveness of our proposed method. Objective and subjective evaluation results demonstrated that our model outperformed the conventional Tacotron model for emotional speech synthesis when only 5% of training data has emotion labels. The naturalness and emotion expressiveness of our proposed method using 5% emotion labels were close to the Tacotron model using all emotion labels. To apply our proposed method to spontaneous emotional and expressive speech will be a task of our future work.

ACKNOWLEDGMENT

REFERENCES

[1] H. Zen, A. W. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 7962–7966.

[2] H. Zen, "Statistical parametric speech synthesis: from hmm to lstm-rnn," 2015, lecture given at RTTH Summer School on Speech Technology, Barcelona, Spain.

[3] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 1964–1968.

[4] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, and O. V. et al., "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[5] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, and R. J. W. et al., "Tacotron: Towards end-to-end speech synthesis," in *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017*, 2017, pp. 4006–4010.

[6] J. Shen, R. Pang, R. J. Weiss, M. Schuster, and N. J. et al., "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*, 2018, pp. 4779–4783.

[7] G. Hofer, K. Richmond, and R. A. J. Clark, "Informed blending of databases for emotional speech synthesis," in *INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 2005, pp. 501–504.

[8] J. Cahn, "The generation of affect in synthesized speech," *Journal of the American Voice I/O Society*, vol. 8, pp. 1–19, 1990.

[9] I. R. Murray and J. L. Arnott, "Implementation and testing of a system for producing emotion-by-rule in synthetic speech," *Speech Communication*, vol. 16, no. 4, pp. 369–390, 1995.

[10] M. Schröder, "Emotional speech synthesis: a review," in *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3-7, 2001*, 2001, pp. 561–564.

[11] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis," *IEICE Transactions*, vol. 88-D, no. 3, pp. 502–509, 2005.

[12] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi, "A style control technique for hmm-based expressive speech synthesis," *IEICE Transactions*, vol. 90-D, no. 9, pp. 1406–1413, 2007.

[13] S. An, Z. Ling, and L. Dai, "Emotional statistical parametric speech synthesis using lstm-rnns," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2017, Kuala Lumpur, Malaysia, December 12-15, 2017*, 2017, pp. 1613–1616.

[14] Y. Lee, A. Rabiee, and S. Lee, "Emotional end-to-end neural speech synthesizer," *CoRR*, vol. abs/1711.05447, 2017.

[15] J. Lorenzo-Trueba, G. E. Henter, S. Takaki, J. Yamagishi, and Y. M. et al., "Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis," *Speech Communication*, vol. 99, pp. 135–143, 2018.

[16] R. J. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, and D. S. et al., "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 4700–4709.

[17] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, and E. B. et al., "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, 2018, pp. 5167–5176.

[18] K. Akuzawa, Y. Iwasawa, and Y. Matsuo, "Expressive speech synthesis via modeling expressions with variational autoencoder," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 3067–3071.

[19] W. Ping, K. Peng, and J. Chen, "Clarinet: Parallel wave generation in end-to-end text-to-speech," *CoRR*, vol. abs/1807.07281, 2018. [Online]. Available: http://arxiv.org/abs/1807.07281

[20] Wei Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 3969–3972.

[21] S. Salvador and P. Chan, "Fastdtw: Toward accurate dynamic time warping in linear time and space," in *KDD workshop on mining temporal and sequential data.* Citeseer, 2004.