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

EMOVIE: A Mandarin Emotion Speech Dataset with a Simple Emotional Text-to-Speech Model

Chenye Cui[1], *Yi Ren*[1], *Jinglin Liu*[1], *Feiyang Chen*[1], *Rongjie Huang*[1], *Ming Lei*[2], *Zhou Zhao*[1]

[1]Zhejiang University, China
[2]Alibaba Group, China

{chenyecui,rayeren,jinglinliu,chenfeiyang,rongjiehuang}@zju.edu.cn,
lm86501@alibaba-inc.com, zhaozhou@zju.edu.cn

Abstract

Recently, there has been an increasing interest in neural speech synthesis. While the deep neural network achieves the state-of-the-art result in text-to-speech (TTS) tasks, how to generate a more emotional and more expressive speech is becoming a new challenge to researchers due to the scarcity of high-quality emotion speech dataset and the lack of advanced emotional TTS model. In this paper, we first briefly introduce and publicly release a Mandarin emotion speech dataset including 9,724 samples with audio files and its emotion human-labeled annotation. After that, we propose a simple but efficient architecture for emotional speech synthesis called EMSpeech. Unlike those models which need additional reference audio as input, our model could predict emotion labels just from the input text and generate more expressive speech conditioned on the emotion embedding. In the experiment phase, we first validate the effectiveness of our dataset by an emotion classification task. Then we train our model on the proposed dataset and conduct a series of subjective evaluations. Finally, by showing a comparable performance in the emotional speech synthesis task, we successfully demonstrate the ability of the proposed model.

Index Terms: emotional speech dataset, speech synthesis, emotional text-to-speech, speech emotion classification

1. Introduction

In the last few years, with the rapid development of text-to-speech (TTS) systems [1, 2, 3, 4, 5], there has been a surge of interest in generating speeches with more emotional information. The specific objective of emotional TTS is to synthesize a more affectionate human-like speech or generate a speech audio sample with expected emotions. And with the ability to generate expressive speeches, the emotional TTS system can be applied to some applications, such as movie dubbing, that could only be done by people with specific skills.

At present, many TTS datasets like VCTK [6], LJSpeech [7] and VCC2018 [8] have been released and have made a significant contribution to speech synthesis tasks. And for Chinese Mandarin tasks, CSMSC[1], AISHELL-3 [9] and DiDiSpeech [10] are published in recent years. Some audio-visual datasets like EmoTV1 [11], ENTERFACE [12], HUMAINE [13], RML [14], VAM [15], CREMA-D [16], RAVDESS [17] and MELD [18] have been established for the multi-model emotional tasks at an earlier stage.

As shown in Table 1, several datasets have been released to conduct emotion-related speech tasks since 2004. AIBO [19] is a cross-linguistic corpus collecting from the interaction between children and robots. PRIORI [20] dataset is recorded from natural conversations during daily smartphone usage and annotates the emotion activation and emotion valence using a 9-point Likert scale. ESD [21] is the first parallel multi-lingual and multi-speaker emotional speech dataset designed for voice conversion tasks and contains five emotional classes in each language. LSSED [22] is a large English dataset designed for SER tasks, which have a nine-classes emotion annotation. EmoV-DB [23] is the first dataset designed for emotional TTS tasks. It is acted by five speakers in two different languages and annotates the audio in five classes of emotions. Among them, no one is collected from the broadcasting television programs or movies, AIBO and PRIORI are collected in a natural environment, and others are acted by specific speakers. Unfortunately, there is no public dataset designed for Mandarin emotional TTS tasks.

Table 1: *Publicly available emotional speech datasets*

Name	Language	Hours	Sampling Rate
AIBO	*Multi*	*9.33*	*44.1 kHz*
PRIORI	*English*	*25.20*	*8 kHz*
ESD	*Multi*	*0.28*	*16 kHz*
LSSED	*English*	*206.42*	*16 kHz*
Emov-DB	*Multi*	*Not Given*	*16 kHz*
EMOVIE	*Mandarin*	*4.18*	*22.05 kHz*

In earlier researches, some deep neural methods like Tacotron [1, 24], FastSpeech [2, 25], and Waveglow [26] have shown their ability to generate fluent speech audio samples as well as expressive singing voice [27, 28] in real-time. Recently, several works have been published to generate emotional speech audio samples and already achieve some signs of progress. The most common methods use emotion embedding, which is extracted with the assistance of a Global Style Token [29] architecture or a speech emotion recognition (SER) model to get a better performance in generating expressive speech. Li's study [30] uses a reference encoder and an auxiliary network to generate the emotional embedding, and calculates a style loss between the predicted spectrograms and the target spectrograms. Wu's work [31] uses the weight of the attention layer in the GST as the probability of each emotion label to calculate the cross-entropy loss and generate the emotion token. Some studies [32] use the speech emotion recognition module as an alternative approach to modeling the emotion feature from existing audio. Several other studies [33, 34] use VAE or DurIAN to model the emotional information and have achieved some progress. The idea most similar to ours is TP-GST [35] and FET [36] which also predict emotion embedding from the text, but they use ex-

[1]https://www.data-baker.com/open%20source.html#/data/index/source

Copyright © 2021 ISCA

http://dx.doi.org/10.21437/Interspeech.2021-1148

tra audio as a reference input when training. In contrast, our model predicts the emotion information from the text context.

However, there are also two remarkable problems in current studies: 1) Many of the emotion speech datasets are designed for speech emotion recognition tasks, which always have ambiguous, unclear, or incomplete speech audio samples and hard to use in speech synthesis. And the high-quality speech datasets for TTS tasks always do not have emotional annotations and cannot be used in emotion tasks. 2) Almost all of the current emotional speech synthesis works need additional reference audio or emotional label input to generate expressive speeches. Thus, the input sequence is often given an inappropriate emotion which always leads to a strange synthesis result.

To solve the first problem, we collect and annotate a new Mandarin emotion speech dataset. The dataset is collected from some movies with natural and expressive speeches and it has an equivalent audio quality to some recording datasets with more natural emotion expression. To our best knowledge, the dataset is the first public Mandarin speech dataset designed for the emotional TTS tasks, and it also can be used for several other emotion-related speech tasks such as speech emotion transfer or SER. The dataset is publicly available at our Github page[2].

To manage the second issue, we propose EMSpeech, an experimental architecture based on the FastSpeech 2 [2], which models a latent mapping from text content to the emotion of the speech, including the duration and the pitch of each phoneme. At the same time, the proposed model could automatically predict emotion through the text input during inference and generates a natural and expressive speech audio sample. Besides, our architecture supports controllable emotional TTS which can receive a manually given emotion label input to generate a speech audio sample with the expected emotion.

In the experiment, we first validate the availability of our dataset by testing the classification accuracy. After that, we make a series of comparative experiments to verify the effectiveness of the proposed method. In order to intuitively see the result of the emotion predicting and emotion controlling, we also analyze the mel-spectrogram and the pitch of the generated speech audio samples.

The main contributions of this paper include: 1) We briefly introduce and publicly release a speech dataset with emotional annotations, which is the first Mandarin movie speech dataset designed for emotional TTS tasks. 2) We establish a simple model for emotional speech synthesis, which achieves a good performance at emotion expression of generated speech.

2. EMOVIE Dataset

In this section, we will introduce EMOVIE, a new Mandarin emotion speech dataset that has high-quality natural emotion speech data and suitable for emotional text-to-speech and some other tasks.

2.1. Data Obtaining and Processing

We get these speech samples from seven Mandarin movies that have a relatively noise-free audio track. To our best knowledge, compared with action movies and war movies, feature movies and comedy movies always have a clearer dialogue environment, so most of the movies are chosen from these kinds of movies. The raw audio with 5.1 channel or 7.1 channel audio tracks is extracted from the movie files using the *ffmpeg* tool. And then we execute the audio track decomposition to get the *Front Center* channel which always has a higher quality dialog audio with less environmental noise.

As we know, embedded subtitles or third-part subtitle files always include the timestamp of each line in the movie. This characteristic can help us cut the full audio track to a single-sentence speech audio sample and correspond the text to its audio fragment.

Moreover, we have further screened the speech audio samples very carefully. At last, we get 9724 samples, 4.18 hours of audio in total.

2.2. Data Annotation

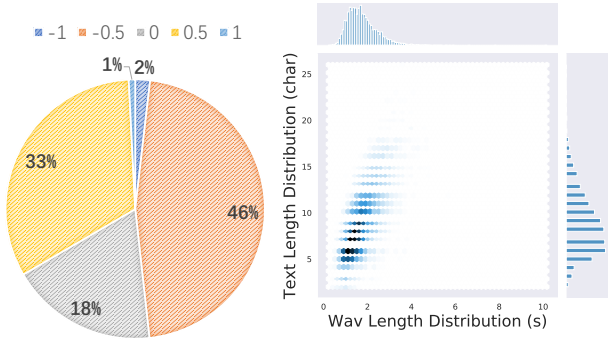
Table 2: *Polarity of the emotion in the dataset*

Polarity	Meaning
-1	The audio has an absolute negative emotion.
-0.5	The audio has an slight negative emotion.
0	The audio has an neutral emotion.
0.5	The audio has an slight positive emotion.
1	The audio has an absolute positive emotion.

We use -1, -0.5, 0, 0.5 and 1 to annotate the emotion polarity of the speech audio sample, and the description of the polarity of the emotion in the annotation is shown in Table 2.

To keep the annotation clearer, the annotators are not allowed to see the text content of the corresponding audio file, which keeps the annotations only obtained by the inherent factors of the audio.

2.3. Data Analysis



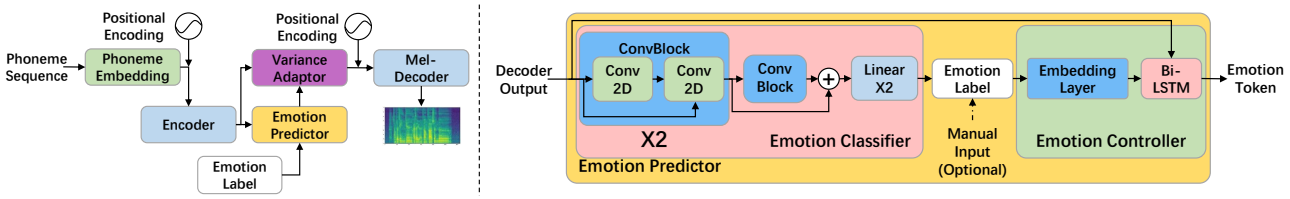
(a) Polarities distribution (b) Length distribution

Figure 1: *Distributions of the dataset*

As shown in Fig. 1(a), we give the emotion polarities distribution of our dataset. Samples with the polarity of '-0.5' and '0.5' are accounted for a larger proportion, which contains 4573 and 3171 samples respectively. The samples of these 2 polarities account for 79% of the total samples. The label second only to them is '0', which has 1783 samples. And there are only 179 and 78 samples for the very obvious polarity '-1' and '1'.

We also plot the joint distribution of the audio samples length and the text length in Fig. 1(b). And the average length of the audio samples and text is 1.78 seconds and 8.93 characters respectively.

[2]https://viem-ccy.github.io/EMOVIE/dataset_release



(a) Pipeline of proposed model

(b) Emotion Predictor architecture

Figure 2: *Model architecture*

3. Model Architecture

Fig. 2(a) shows the framework of our proposed EMSpeech for training and inference. It shares a similar architecture with the FastSpeech 2 [2] which is a fast and high-quality end-to-end TTS system.

In this paper, since the emotional annotations in our dataset do not directly contain any information from the text, we regard the polarity of emotion as an attribute of speech and try to model the implicit mapping between text information and latent speech emotion representations. Finally, we use the emotion information to control the prosody including duration and the pitch, and then emotional speech samples can be generated.

In detail, we use an emotion predictor on the proposed model, which could predict the emotion of the input text and could encode the emotion label to a latent emotion embedding. By adding the emotion embedding to the output of the encoder, the duration predictor and pitch predictor could make the speech sample have a better emotion expression.

The emotion predictor contains two main components: 1) an emotion classifier and 2) an emotion controller.

Also, the emotion controller can receive either the predicted emotion label or the external specific emotion label as input at inference stage, which means the model not only can generate an expressive speech audio sample with an automatically predicted emotion but also can generate a speech audio sample with an expected emotion.

Besides, a speaker embedding is added after the duration predictor, which is adapted for multi-speaker generating.

3.1. FastSpeech 2

The components of the FastSpeech 2 include: a phoneme embedding module, a encoder, a variance adaptor and a mel-spectrogram decoder. The encoder converts the phoneme embedding sequence into the phoneme hidden sequence. The variance adaptor adds different variance information such as duration, pitch and energy into the hidden sequence. The mel-spectrogram decoder converts the adapted hidden sequence into mel-spectrogram sequence in parallel.

3.2. Emotion Classifier

The effectiveness of the emotion predictor mainly depends on the emotion classifier. And we use the ResNet-like network, which is simple enough and has been proven its ability on classification tasks, to predict the emotion label from the input text.

As shown in Fig. 2(b), the framework of the emotion classifier is a 3-layer ConvBlock stack with skip connect operations among the layers. Each ConvBlock has double 2D convolution layers followed by a 2D pooling operation.

3.3. Emotion Controller

The emotion controller aims to map a specific emotion label to an emotion embedding. Considering the emotion embedding is related to a specific phoneme and its position of the input sequence, we choose a network with recurrent structures to encode a sequence-length emotion embedding.

As shown in Fig. 2(b), the Emotion controller uses an Embedding network to model an intermediate feature vector used as the initial hidden state of a Bi-LSTM network. Then we use the Bi-LSTM network to model the phoneme-level emotion embedding for the whole input sequence.

4. Experiment

In this section, we give the details of our experiments on the proposed dataset. Although visualizations for some experiments result is provided below, we recommend readers to listen to the audio samples published on the demo page[3] to have a more intuitive impression.

4.1. Experimental Setup

4.1.1. Dataset Preparation

We first randomly spilt our dataset into 3 sets: 9,524 samples for training, 100 samples for validation and 100 samples for testing. For the audio data, we convert the audio waveform into 80-dimensional mel-spectrograms for the emotional text-to-speech model. The frame size and hop size of the mel-spectrograms are set to 1024 and 256 respectively. Following the approach in previous works, we use an open-source grapheme-to-phoneme tool[4] to convert the text sequence into the phoneme sequence. After that, we use *Montreal Forced Aligner*(MFA) [37] tool to extract the phoneme duration. The emotion data are also processed during the data preparation. We choose the emotion polarity as the emotion information input and convert them to real number labels.

4.1.2. Model Configuration

Our emotional TTS model is based on FastSpeech 2 [2] which consists of 4 feed-forward Transformer (FFT) blocks in the encoder and the mel-spectrogram decoder. Following FastSpeech 2, for each FFT block, both the dimension of phoneme embedding and the hidden size of the self-attention are set to 256. The number of attention heads is set to 2 and the kernel sizes of the 1D-convolution in the 2-layer convolutional network after the self-attention layer are set to 9 and 1. The input/output sizes of the first layer and the second layer in the 2-layer convolutional network are set to 256/1024 and 1024/256 respectively. For the variance predictor, the kernel sizes of the 1D-convolution are

[3]https://viem-ccy.github.io/EMOVIE/

[4]https://github.com/Kyubyong/g2p

set to 3, with input/output sizes of 256/256 for both layers and the dropout rate is set to 0.5.

In the emotion classifier, all the kernel sizes of the 2D-convolution are set to 3x3 and both of the stride and padding sizes are set to 1. And the input/output dimensions of the 3 ConvBlocks are 1/64, 64/128, 128/256 respectively. In the emotion controller, the dimension of the Bi-LSTM layer's input is set to 512 which equal to the size of the encoder layer output plus the size of Embedding layer output. And the size of the hidden state is set to 256.

4.1.3. Training and Inferring

Both the baseline FastSpeech 2 model and our proposed model are trained on a single NVIDIA RTX2080Ti GPU with a batch size of 16 sentences. Adam optimizer is applied to our model and the configuration is set to $\beta_1 = 0.9$,$\beta_2 = 0.98$ and $\epsilon = 10^{-9}$, and the learning rate schedule is set as same as the FastSpeech 2 configuration. Besides, the classifier is optimized with cross-entropy loss, the variance adaptor is optimized with mean square error (MSE) loss and other parts of the model are optimized with mean absolute error (MAE) loss.

To get a more precise emotion modeling, the emotion controller uses a true emotion label as conditional inputs to model the emotion embedding instead of using the predicted label when training. In inference, we use values predicted by the classifier that are jointly trained with the EMSpeech model.

During the inference, the output 80-dimensional mel-spectrograms of our model are transformed into waveform audio by a pre-trained Parallel WaveGAN [38].

4.2. Experiment Result

4.2.1. Emotion Classification on EMOVIE

To verify the reliability of the annotation of the dataset, we conduct an experiment about classification accuracy. We use the output emotion classifier which is jointly trained with the model to calculate the accuracy. The network exhibits a comparable result of emotion polarity classification which achieves an accuracy rate up to 48.2%. Since the accuracy is significantly higher than the random classification, we show that the annotations of the dataset is available.

4.2.2. Evaluation on Audio with Predicted Emotion

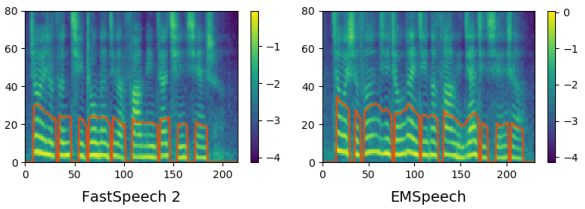


Figure 3: *Generating result of emotion predicting*

To prove that our model can generate more emotional speech audio sample with the assistance of the emotion predictor, we conduct a experiment between the FastSpeech 2 and our EMSpeech. In this experiment, we generate speech audio samples using FastSpeech 2 and our model with the same text input, and next evaluate their emotional expression. We visualize the comparison of spectrograms in Fig. 3. We can see the generating result of EMSpeech has a greater pitch fluctuation than that result from the FastSpeech 2. The scale of the MOS is set between 1 to 5 in emotion predicting results evaluation. MOS from 1 to 5 denotes No, Slight, Ordinary, Good, and Strong expressiveness respectively in the evaluation on audio with predicted emotion. And the MOS in Table 3 illustrates our model can generate speech audio samples containing more emotional expressions. We show that our EMSpeech has the ability to model the mapping between the input text and the speech emotion, which yields more expressive speech than a FastSpeech 2 conditioned on the same data.

Table 3: *MOS of emotion predicting result*

FastSpeech 2	**EMSpeech**
3.45	3.88

4.2.3. Analyses on Emotion Controllability

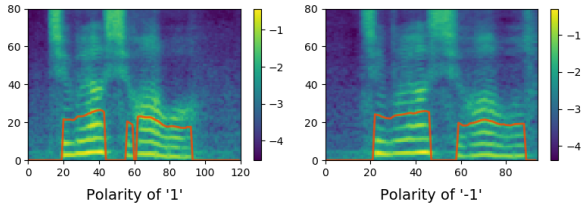


Figure 4: *Generating result of emotion controlling*

We conduct a emotion controllability experiment to further prove that the label we manually input can control the emotion of the generated audio. In the emotion controlling experiment, we use two different emotion labels '1' and '-1' to generate speech audio samples with the same text content and compare their emotional tendencies to verify. Fig. 4 gives a comparison of spectrograms and pitch between the audio sample generated with the two labels. From the figure we can see the audio with the polarity of '1' often has a shorter phoneme duration with a rising pitch trend compared to the audio with the polarity of '-1'. For subject evaluation, we choose the mean opinion score (MOS) to evaluate polarity of the synthesized speeches by our emotional TTS model. In emotion controllability evaluation the scale of MOS is set to between -1 to 1, where -1 denotes strong negative and 1 denotes strong positive polarity. And as shown in Table 4, the MOS shows that the samples generated with the polarity of '-1' often have a more negative emotion than the samples generated with the polarity of '1', indicating that our method can generate controllable speech in excepted polarities.

Table 4: *MOS of emotion controlling result*

Polarity label of '1'	**Polarity label of '-1'**
0.28	−0.40

5. Conclusion

In this paper, we introduced and publicly released a new Mandarin movie speech dataset with emotion polarity annotations. Moreover, an emotional text-to-speech system based on FastSpeech 2 was proposed and achieved a good performance. According to the experimental result, we successfully demonstrated the availability of our EMOVIE dataset and the effectiveness of our proposed EMSpeech in terms of generating speech with more emotion expression and controlling the emotion of generated speech. Even though the system of EMSpeech is a very simple one, the result shows the efficiency of the EMOVIE dataset which is promising for future work. In the future, we will enhance our dataset and improve our model, and seek new breakthroughs in emotion-related tasks.

6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 ICASSP*. IEEE, 2018, pp. 4779–4783.

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

[3] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, T. Qin, and T.-Y. Liu, "Multispeech: Multi-speaker text to speech with transformer," *arXiv preprint arXiv:2006.04664*, 2020.

[4] Y. Ren, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Almost unsupervised text to speech and automatic speech recognition," in *ICML*. PMLR, 2019, pp. 5410–5419.

[5] J. Xu, X. Tan, Y. Ren, T. Qin, J. Li, S. Zhao, and T.-Y. Liu, "Lrspeech: Extremely low-resource speech synthesis and recognition," in *26th SIGKDD*, 2020, pp. 2802–2812.

[6] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "Superseded-cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," 2016.

[7] K. Ito and L. Johnson, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[8] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods," *arXiv preprint arXiv:1804.04262*, 2018.

[9] Y. Shi, H. Bu, X. Xu, S. Zhang, and M. Li, "Aishell-3: A multi-speaker mandarin tts corpus and the baselines," *arXiv preprint arXiv:2010.11567*, 2020.

[10] T. Guo, C. Wen, D. Jiang, N. Luo, R. Zhang, S. Zhao, W. Li, C. Gong, W. Zou, K. Han *et al.*, "Didispeech: A large scale mandarin speech corpus," *arXiv preprint arXiv:2010.09275*, 2020.

[11] S. Abrilian, L. Devillers, S. Buisine, and J.-C. Martin, "Emotv1: Annotation of real-life emotions for the specification of multimodal affective interfaces," in *HCI International*, vol. 401, 2005, pp. 407–408.

[12] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *ICDEW'06*, 2006, pp. 8–8.

[13] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data," in *International conference on affective computing and intelligent interaction*, 2007, pp. 488–500.

[14] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE transactions on multimedia*, vol. 10, no. 5, pp. 936–946, 2008.

[15] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *ICME 2008*, 2008, pp. 865–868.

[16] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[17] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[18] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," *arXiv preprint arXiv:1810.02508*, 2018.

[19] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. J. Russell, and M. Wong, ""' you stupid tin box"-children interacting with the aibo robot: a cross-linguistic emotional speech corpus," 2004.

[20] S. Khorram, M. Jaiswal, J. Gideon, M. McInnis, and E. M. Provost, "The priori emotion dataset: Linking mood to emotion detected in-the-wild," *arXiv preprint arXiv:1806.10658*, 2018.

[21] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," *arXiv preprint arXiv:2010.14794*, 2020.

[22] W. Fan, X. Xu, X. Xing, W. Chen, and D. Huang, "Lssed: a large-scale dataset and benchmark for speech emotion recognition," *arXiv preprint arXiv:2102.01754*, 2021.

[23] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

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

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

[26] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *ICASSP 2019*. IEEE, 2019, pp. 3617–3621.

[27] Y. Ren, X. Tan, T. Qin, J. Luan, Z. Zhao, and T.-Y. Liu, "Deepsinger: Singing voice synthesis with data mined from the web," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1979–1989.

[28] J. Liu, C. Li, Y. Ren, F. Chen, P. Liu, and Z. Zhao, "Diffsinger: Diffusion acoustic model for singing voice synthesis," *arXiv preprint arXiv:2105.02446*, 2021.

[29] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*, 2018, pp. 5180–5189.

[30] T. Li, S. Yang, L. Xue, and L. Xie, "Controllable emotion transfer for end-to-end speech synthesis," in *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2021, pp. 1–5.

[31] P. Wu, Z. Ling, L. Liu, Y. Jiang, H. Wu, and L. Dai, "End-to-end emotional speech synthesis using style tokens and semi-supervised training," in *APSIPA ASC 2019*. IEEE, 2019, pp. 623–627.

[32] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," *arXiv preprint arXiv:2010.13350*, 2020.

[33] V. Aggarwal, M. Cotescu, N. Prateek, J. Lorenzo-Trueba, and R. Barra-Chicote, "Using vaes and normalizing flows for one-shot text-to-speech synthesis of expressive speech," in *ICASSP 2020*. IEEE, 2020, pp. 6179–6183.

[34] Z. Zhang, Q. Tian, H. Lu, L.-H. Chen, and S. Liu, "Adadurian: Few-shot adaptation for neural text-to-speech with durian," *arXiv preprint arXiv:2005.05642*, 2020.

[35] D. Stanton, Y. Wang, and R. Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *SLT 2018 Workshop*. IEEE, 2018, pp. 595–602.

[36] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," *arXiv preprint arXiv:2011.08477*, 2020.

[37] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi." in *Interspeech*, vol. 2017, 2017, pp. 498–502.

[38] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020*. IEEE, 2020, pp. 6199–6203.