INTERACTIVE TEXT-TO-SPEECH VIA SEMI-SUPERVISED STYLE TRANSFER LEARNING

Yang Gao[1,⋆] *, Weiyi Zheng*[2]*, Zhaojun Yang*[2]*, Thilo Köhler*[2]*, Christian Fuegen*[2]*, Qing He*[2,†]

[1]Carnegie Mellon University [2]Facebook AI

ABSTRACT

With increasing interests in interactive speech systems, speech emotion recognition and multi-style text-to-speech (TTS) synthesis are becoming increasingly important research areas. In this paper, we combine both. We present a method to extract speech style embeddings from input speech queries and apply this embedding as conditional input to a TTS voice so that the TTS response matches the speaking style of the input query. Specifically, we first train a multi-modal style classification model using acoustic and textual features of speech utterances. Due to a limited amount of labeled data, we combined the emotional recognition dataset: the interactive emotional dyadic motion capture database (IEMOCAP) with a small labeled subset of our internal TTS dataset for style model training. We take the softmax layer from the style classifier as style embedding and then apply this style embedding extraction model to generate soft style labels for our unlabelled internal TTS dataset. With this semi-supervised approach, reliable style embeddings are extracted to train a multi-style TTS system. As a result, we developed a controllable multi-style TTS system whose response matches the given target styles embedding, which could be extracted from the input query or manually assigned.

Index Terms— Text-to-speech synthesis, emotion, style, semi-supervised

1. INTRODUCTION

Speech is a crucial part of human-computer interactions and high-quality TTS synthesis plays an important role in mimicking natural human communications. With recent technology advancements in speech synthesis [1–3], TTS systems can achieve near human quality [4]. One popular topic in the recent research of TTS is expressive TTS, which is aiming at achieving controllable style synthesis in TTS [5–7]. To avoid the difficulty of hand-labeling prosody and speaking style, style embedding could be extracted using a style encoder and concatenated with the textual feature, to guide the attention module of the synthesizer during training/inference [5, 6]. The latent styles could also be factorized using a token table [7].

These prosody transfer methods could be used to create an interactive TTS system mimicking the style of the input audio query. However, to learn specific styles, there are limitations with unsupervised style factorization learning [7]. Since the disentanglement of different styles is heavily influenced by randomness and the choice of hyper-parameters [8], the learning of target styles is not controllable. To have interpretable representation learning, some (at least weak) supervision is in need [8]. Under supervision with explicit prosody labels, the styles could be learned with clear guidance [9, 10]. But supervised learning requires a large amount of labeled data, giving difficulties in the development of expressive TTS research and applications.

In [11], an external dataset IEMOCAP [12] is used to train an emotion classifier and the trained model is then employed to label previously unlabeled TTS training data: Blizzard 2017 [13]. It proposed a way to enable automatic emotion labeling of the unlabeled dataset and to produce a controllable TTS. But, the external dataset IEMOCAP and the synthesis dataset Blizzard 2017 [13] have significant differences in background noise, recording environment, speech quality, etc. With the differences between these two datasets, the classifier trained using an external dataset may not be well-adapted to extract representations from the synthesis data. The final emotion synthesis accuracy is 41% on four emotions [11] evaluated by listeners, while the human judgment accuracy on emotions in real speech is around 50% reported in [11, 14]. Since there are no emotion labels on the synthesis dataset Blizzard 2017, the prediction accuracy of the TTS data is not directly evaluated.

In this paper, we propose a semi-supervised approach to learn reliable style representation on the synthesis dataset. Apart from making usages of the external IEMOCAP dataset to help the learning on 4 emotion classes as in [11], we further improve the transfer learning effect using the labeled subset of our TTS dataset. To increase the quality of the learned style representation and have more reliable style embedding in the expressive TTS training, we combined the IEMOCAP dataset with the labeled subset of our internal dataset to create a joint dataset and train a multi-modal style classification model using this joint dataset. Taking the softmax layer of the style classifier as style embedding, the classifier serves as a style embedding extraction model, which lets us then generate style embedding for our unlabelled internal TTS dataset. By using the style embedding as additional auxiliary features for the TTS system, we could train a controllable multi-style TTS system that learns to respect given target styles. During speech synthesis, style embedding could be extracted from the input speech query and fed into our TTS system, which will then produce its response in matching styles as the input query. As a result, we developed a novel interactive multi-style TTS system. The multi-style TTS system is evaluated using comprehensive subjective experiments.

2. RELATED WORK

2.1. Expressive TTS

Expressive TTS has been studied for years from the HMM-based synthesis using style modeling with control vector [15–17] to the state-of-the-art prosody transfer expressive TTS work [5–7]. In [7], a token table is learned through the "style token layer" to represent a variety of prosodic dimensions. Furthermore, controllable TTS is discussed in [10] for the emotion nuances style learning. In [11], an external dataset is used to help the learning of control dimensions in the TTS dataset. In their work, they used statistical parametric

⋆ Work performed during internship at Facebook AI

† Corresponding author

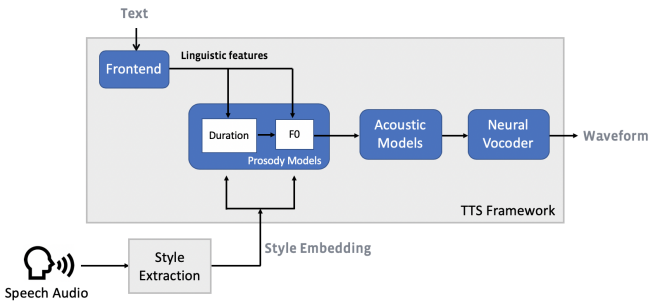

Fig. 1. Style embedded TTS framework

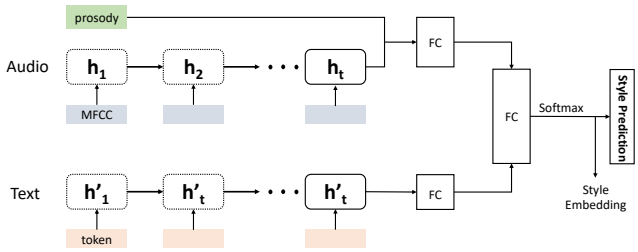

Fig. 2. Multimodal style classifier

speech synthesis (SPSS) with the control vector as an auxiliary input feature to drive their expressive TTS system.

2.2. Emotion recognition

Early approaches on emotion recognition mostly have been inspired by psychology studies [18, 19]. Recently, deep neural networks (DNNs) have first been used to learn high-level representations for utterance-level emotion recognition [20]. Trigeorgis *et al.* further applied convolutional neural networks (CNNs) to model context-aware emotion-relevant features, which are then combined with long term-short memory (LSTM) networks aiming towards end-to-end emotion modeling [21]. Fundamentally, the expression of emotions is usually conveyed through multi-modal behavior channels, including speech, language, body gestures, or facial expressions. Thus, emotion recognition is often formulated as a classification problem of utterances using these multi-modal signals. [22] proposed a multi-modal dual recurrent encoder to simultaneously model the dynamics of both text and audio signals within an utterance to predict emotion classes. This architecture has achieved state-of-the-art performance on IEMOCAP [12] dataset which is a multi-modal emotion dataset and has been widely used in the affective computing community. In this work, our speech style recognizer is built based on this architecture.

3. MODELS AND FRAMEWORKS

Figure 1 shows the architecture of the expressive TTS system. It consists of a standalone style embedding extraction component that generates the style embedding from audio input, and a TTS framework which takes the style embedding as input and synthesizes the response in matching style.

3.1. Semi-supervised style learning

The multimodal dual recurrent encoder (MDRE) model we used for speech style classification is adapted from the state-of-the-art emotion recognition model introduced in [22]. As shown in Figure 2, the model is composed of two separate recurrent encoders for audio and text modeling, respectively. The audio model uses Mel-frequency Cepstral Coefficients (MFCC) features and utterance level prosody feature as inputs and the text model uses token representations as described in [22]. The audio encoder output is concatenated with the text encoder output, then fed into a fully-connected layer to produce the final classification. We changed the loss function from sigmoid cross-entropy to softmax cross-entropy as it produced significantly better results for our training task.

Since our internal dataset is only partially labeled, we train the semi-supervised style classifier using the labeled portion and then infer the rest of the dataset's style embedding using this classifier. The style embedding we chose to compare are 1) the bottleneck layer feature before the final prediction 2) the final softmax features. During the experiments, we found the softmax layer output works better. This feature also has the added advantage that it can be interpreted as a probability representation of different speaking styles. The softmax feature as embedding is shown in Figure 2.

3.2. Style embedded TTS system

As shown in Figure 1, our TTS pipeline is a multi-model framework that consists of a text-processing frontend, a set of prosody models, acoustic models and a neural vocoder. This framework is ideal for controllable expressive speech synthesis because it separates speech style modeling from speech audio modeling. Specifically, the input text is first converted to linguistic features. Then, the linguistic features along with any conditional features such as style embedding, speaker ID are used to produce the prosody features such as duration and F_0. Linguistic features combined with prosody features are used to generate spectral acoustic features and, at the last stage, a conditional neural vocoder takes in the spectral features to synthesize the audio waveform. The speaking style of the synthesized speech is controlled by the conditional style embedding feature. During inference, we would use the style classifier to extract the style embedding from a short speech segment as the input query or manually assigned. In an interactive speech system, this speech segment can be the input utterance from a user.

4. EXPERIMENTS

4.1. Datasets

4.1.1. Internal dataset

The internal TTS dataset was recorded in voice production studio by multiple professional voice talents. It has balanced phonemic and textual information and the sampling rate is 48kHz. Only a small subset of the internal dataset has style labels. In addition to the IEMOCAP styles, the internal data has two more styles: fast and soft. Details of the data are summarized in Table1. These utterances are used to train a multi-speaker style classifier. During the synthesis, only one speaker's voice is used which has 40,244 utterances with only around 3000 utterances labeled. The style representations are extracted on the rest of this single speaker dataset and used in the TTS training.

Table 1. Total dataset style labels statistics

Dataset	Split	Fast	Soft	Neutral	Happy	Angry	Sad
Internal	Train	1145	1814	4481	885	140	35
	Dev	105	161	439	79	13	3
	Test	124	220	506	93	17	2
	All	1374	2195	5426	1057	170	40
IEMOCAP	Train	–	–	1390	1307	865	883
	Dev	–	–	100	90	61	62
	Test	–	–	218	239	177	139
	All	–	–	1708	1636	1103	1084

4.1.2. IEMOCAP dataset

To compensate for the limited amount of labeled data in our internal dataset, we chose an open-source dataset, IEMOCAP [12], which is widely used for emotion recognition, to complement our training data. In this dataset, both video and audio were recorded from ten actors in dyadic sessions under both scripted and spontaneous communication scenarios. The dataset contains 12.5 hours of recordings with a sampling rate of 22kHz. It has utterance-wise emotion labels such as neutral, happy, sad, anger, surprise, etc. To be consistent with former research [11, 22] and also be suitable for our own interaction goal, we select the following emotions in our study: neutral, happy, sad and angry. Similar to the approach in [22], we merge utterances with excited emotion with those of happy emotion.

4.2. Implementation details

The style classification model is adapted from [22] and is shown in Figure 2. Specifically, we set the batch normalization layer with 0.9 momentum to help cross-domain adaptation. To compensate for the imbalance among style labels, we weighted the by-class loss function with an inverse of the style label prior while capping the neutral label prior as 0.25. In addition, AdaBN [23] is implemented in this model to boost domain adaptation performance between the internal and external datasets.

The multi-style TTS system is trained using the internal dataset with style embedding features as conditional input features. The style embedding labels were generated by passing each utterance through the style classification model as described in Section 3.1. In the synthesis phase, the style embedding features could be automatically extracted from input query or manually assigned as a combination of different styles.

5. RESULTS

5.1. Style classification

In the style classification task, we first tested the style classifier's model performance on the IEMOCAP train/test split. It achieves an overall accuracy of 72.7% which is similar to the reported state-of-the-art [22]. The confusion matrix is shown in Figure 3. To improve the embedding quality on the internal dataset, the IEMOCAP dataset and the labeled subset of the internal dataset were combined during training. The results show that the style classifier achieves 91.4% overall accuracy and 71.5% weighted accuracy on the internal labeled dataset. Figure 4 shows the confusion matrix on the internal dataset validation set. The number of utterances among different classes are unbalanced, with the neutral class having the most number of utterances, as shown in Table 1. With a lack of labelled data in anger and sadness in the internal dataset, the prediction accuracy on these two classes are not high. The style classification accuracy decreased slightly on the IEMOCAP dataset after joint training, likely due to mismatch between the internal and IEMOCAP datasets.

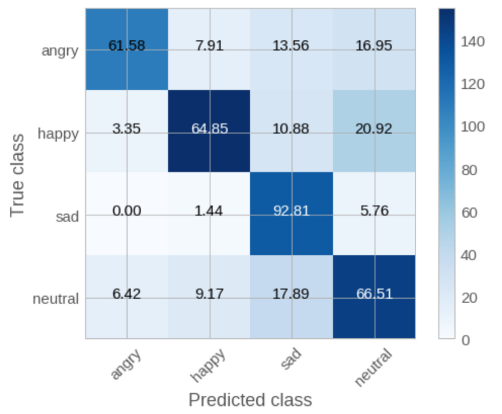

Fig. 3. Confusion matrix on IEMOCAP data

To choose the best input features, we have also done feature normalization experiments. The normalization is done corpus-wise to compensate for the domain difference between our internal dataset and the IEMOCAP dataset. Table 3 shows that normalizing both MFCC and prosody provides the best classification accuracy on the internal dataset's validation set. So in the final model, we normalized both MFCC features and prosody features. The final classification accuracy for the internal dataset is in Table 2.

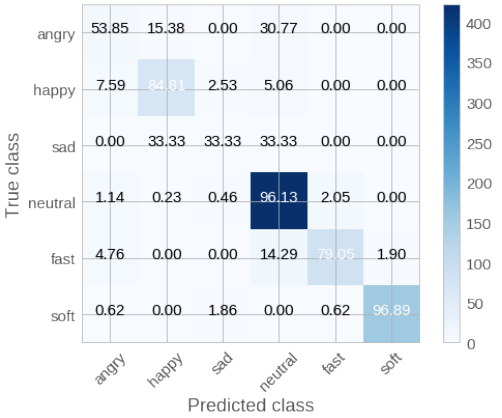

Fig. 4. Confusion matrix on internal labeled data

5.2. Multi-style TTS with conditional style embedding

To evaluate our expressive TTS's performance, we collected subjective evaluation responses from 22 subjects. As reported in [11, 14, 24], the human perception on the emotions of natural speech is only around 50%, showing the ambiguity of emotion perception. Hence, instead of evaluating the subjective style accuracy on the multi-style synthesis results, we are doing the ABX test and preference test. The demo page for our system's synthesis results is at [25].

Table 2. Style classification on internal data

Dataset	Trick	Neutral	Fast	Soft	Happy	Angry	Sad	**Accuracy** Weighted	Unweighted
Train	BN	0.984	0.871	0.964	0.892	0.176	0.0	0.779	0.973
	AdaBN	0.953	0.847	0.918	0.903	0.353	0.0	0.915	0.957
Dev	BN	0.979	0.819	0.994	0.81	0.385	0.0	0.686	0.931
	AdaBN	0.927	0.8	0.963	0.873	0.538	0.333	0.766	0.904
Test	BN	0.984	0.871	0.964	0.892	0.176	0.0	0.683	0.940
	AdaBN	0.953	0.847	0.918	0.903	0.353	0.0	0.715	0.914

Table 3. Feature selection

Features	**Accuracy** Weighted	Unweighted
Unnormalized	0.726	0.875
Normalized MFCC	0.673	0.840
Normalized prosody	0.494	0.62
Normalized both	0.766	0.904

Table 4. Subjective Preferences

	Baseline	Neutral Style	Other Styles
Preference (%)	28.0	54.2	17.8

5.2.1. ABX test and preference score

We designed the ABX test as follows. With two different styles, we randomly chose an example from each style. Let's call these two examples A and B. We also choose a different sample X from one of these two styles and call it the references style. The listener will decide which of A, B has the same style as the reference X. We prepared 15 such tests. The aggregated results gave an 82.42% accuracy, showing the styles can be distinguished between classes.

For preference score test, we asked the listeners to choose between three styles synthesized with the same text: baseline TTS model (i.e., TTS without style embedding), neutral style and other hand-chosen styles. As in the conclusion in [11], the listeners prefer appropriate variation over random variation. So we manually assign soft style labels (not neutral style) in the inference to accompany the text information in that utterance. The evaluation results are shown in Table 4. The neutral style created by our system is the best accepted by the listeners, showing training the TTS with style labels can improve the quality of synthesis results. However, the other hand-picked styles are not best-accepted by listeners who refer to the neutral style as more peaceful and relaxing, which might be influenced by the F0 preference. This may be further improved by disentangling F0 with prosody variation in future studies.

Table 5. TTS data F0 statistics

Style	F0	Count
Angry	195.5±30.8	6817
Happy	214.8±37.3	3576
Sad	197.3±30.8	5137
Neutral	183.7±10.3	16431
Fast	181.9±12.8	5579
Soft	180.5±14.7	2704

5.2.2. Multi-style response to real life input query

We conducted experiments to evaluate the generalization capacity of the close-loop style extraction and multi-style TTS system. We recorded speech queries from multiple speakers by letting them read the queries freely in a conference room. We then generated TTS responses for each query by conditioning on its style embedding.

We randomly selected 10 query/responses pairs. We let the listeners compare the multi-style TTS responses with the baseline TTS responses, and select which response's style matches the input query better. Our results show that over 40% of test pairs have more than 60% matching rate; 10% of test pairs have less than 40% matching rate; and 50% has around 50% matching rate. When the speaking style of the input query is strong, the TTS response can match the input style to a certain extent. Samples are at [25].

6. DISCUSSIONS

Due to the lack of anger and sadness samples in the internal dataset, the learned styles in these classes are transferred from the class representation in IEMOCAP dataset which may not represent the ideal sad or angry speaking style for TTS. Also, we noticed F0 differs significantly in the synthesis results for different classes, which may be because the mean and variance of F0 are different among predicted classes in TTS training data, as shown in Table 5. We also notice that since the style embedding is a weighted representation of different styles, decreasing the weight of a certain style weakens that style's effects on the synthesis outputs, shown in [25]. In the future, the performance of the multi-style expressive TTS system can be further improved with a training dataset that contains more balanced style labels and more significant emotion and prosody variations.

7. CONCLUSIONS

In conclusion, we developed an interactive TTS system that has the potentials to synthesize matching speaking styles as the input query. It is composed of a multi-modal style classifier and a neural-network-based TTS system. The style classifier is jointly trained using our labeled internal data and the IEMOCAP open source dataset. With a limited amount of style labeled TTS data, we used a semi-supervised approach to train the TTS system such that it can generate controllable multi-style TTS responses in a reliable manner.

8. REFERENCES

[1] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio," *arXiv preprint arXiv:1609.0349*, 2016.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *3rd International Conference on Learning Representations (ICLR)*, 2015.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in neural information processing systems (NeurIPS)*, 2017, pp. 5998–6008.

[4] 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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[5] Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A Saurous, "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[6] 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," *International Conference on Machine Learning (ICML)*, 2018.

[7] 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," *International Conference on Machine Learning (ICML)*, 2018.

[8] Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem, "Challenging common assumptions in the unsupervised learning of disentangled representations," *International Conference on Machine Learning (ICML)*, 2018.

[9] Noé Tits, Fengna Wang, Kevin El Haddad, Vincent Pagel, and Thierry Dutoit, "Visualization and interpretation of latent spaces for controlling expressive speech synthesis through audio analysis," *INTERSPEECH*, 2019.

[10] Gustav Eje Henter, Jaime Lorenzo-Trueba, Xin Wang, and Junichi Yamagishi, "Principles for learning controllable TTS from annotated and latent variation.," in *INTERSPEECH*, 2017, pp. 3956–3960.

[11] Zack Hodari, Oliver Watts, Srikanth Ronanki, and Simon King, "Learning interpretable control dimensions for speech synthesis by using external data.," in *INTERSPEECH*, 2018, pp. 32–36.

[12] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[13] Simon King and Vasilis Karaiskos, "The Blizzard challenge 2017," in *Proc. Blizzard Challenge*, 2017, vol. 2017, pp. 1–10.

[14] Rainer Banse and Klaus R Scherer, "Acoustic profiles in vocal emotion expression.," *Journal of personality and social psychology*, vol. 70, no. 3, pp. 614, 1996.

[15] Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi, "Modeling of various speaking styles and emotions for HMM-based speech synthesis," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[16] Makoto Tachibana, Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi, "HMM-based speech synthesis with various speaking styles using model interpolation," in *International Conference on Speech Prosody*, 2004.

[17] Junichi Yamagishi, Makoto Tachibana, Takashi Masuko, and Takao Kobayashi, "Speaking style adaptation using context clustering decision tree for HMM-based speech synthesis," in *2004 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2004, vol. 1, pp. I–5.

[18] Chi-Chun Lee, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[19] Emily Mower, Maja J Mataric, and Shrikanth Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 19, no. 5, pp. 1057–1070, 2010.

[20] Kun Han, Dong Yu, and Ivan Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH*, 2014.

[21] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[22] Seunghyun Yoon, Seokhyun Byun, and Kyomin Jung, "Multimodal speech emotion recognition using audio and text," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 112–118.

[23] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu, "Adaptive batch normalization for practical domain adaptation," *Pattern Recognition*, vol. 80, pp. 109–117, 2018.

[24] Klaus R Scherer, Rainer Banse, Harald G Wallbott, and Thomas Goldbeck, "Vocal cues in emotion encoding and decoding," *Motivation and emotion*, vol. 15, no. 2, pp. 123–148, 1991.

[25] Yang Gao, "Demo for 'interactive text-to-speech via semi-supervised style transfer learning'," https://github.com/Yolanda-Gao/Interactive-Style-TTS, 2019, Accessed: 2019-10-21.