Interspeech 2022
18-22 September 2022, Incheon, Korea

Improve emotional speech synthesis quality by learning explicit and implicit representations with semi-supervised training

Jiaxu He[1,†], *Cheng Gong*[1,†], *Longbiao Wang*[1,*], *Di Jin*[1],*Xiaobao Wang*[1],*Junhai Xu*[1], *Jianwu Dang*[1,2]

[1]Tianjin Key Laboratory of Cognitive Computing and Application,
College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Japan Advanced Institute of Science and Technology, Ishikawa, Japan
{jiaxu_123,gongchengcheng,longbiao_wang}@tju.edu.cn

Abstract

Due to the lack of high-quality emotional speech synthesis datasets, the naturalness and expressiveness of synthesized speech are still lacking in order to achieve human-like communication. And existing emotional speech synthesis system usually extracts emotional information only from reference audio and ignores sentiment information implicit in the text. Therefore, we propose a novel model to improve emotional speech synthesis quality by learning explicit and implicit representations with semi-supervised learning. In addition to explicit emotional representations from reference audio, we propose an implicit emotion representations learning method based on graph neural network, considering dependency relations of a sentence and text sentiment classification (TSC) task. For the lack of emotion-annotated datasets, we leverage large amounts of expressive datasets to reinforce training the proposed model with semi-supervised learning. Experiments show that the proposed method can improve the naturalness and expressiveness of synthetic speech and is better than the baseline model.

Index Terms: Emotional speech synthesis, BERT,Graph neural networks, text sentiment representations

1. Introduction

In the past few years, neural speech synthesis techniques have experienced significant development. End-to-end text-to-speech (TTS) systems, such as [1, 2, 3] have achieved remarkable results in terms of naturalness and intelligibility of general speech. Benefiting from these techniques, the field of emotional speech synthesis (ESS) [4, 5] has attracted extensive attention from researchers because it is closer to practical applications.

Emotional speech synthesis aims to produce natural and expressive speech using prescribed emotions, usually from one of several predefined emotional categories (happiness, anger, etc.) [6, 7]. To achieve emotional speech synthesis, a common solution is to learn emotion-related latent representations from the reference audio [8, 9]. The goal is to make the synthesized speech imitate the emotion of the reference audio, which can be treated as some kind of style transfer. The methods mentioned above report some promising results in the aspect of emotion expressiveness, but these methods rely on an emotion-annotated dataset that is most likely not available. The lack of emotion-annotated speech datasets is one of the main obstacles that limit the research of emotional speech synthesis.

Therefore, some semi-supervised approaches have been proposed to alleviate the burden of data requirements [10]. Tits et al. [11] investigated how to leverage fine-tuning on a pre-trained Deep Learning-based TTS model to synthesize emotional speech with a small dataset of another speaker. Wu et al. [12] presented an emotional speech synthesis method using style tokens and semi-supervised training. Paper [13] proposed to merge an external SER dataset and a labeled subset of the TTS dataset to train a SER model and label the whole TTS dataset by the trained SER model. These semi-supervised methods can greatly reduce the amount of labeled data required for model training. However, these methods are still not universal enough because a subset of the emotional dataset is usually not high-quality synthesis-specific audios and ignores the implicit sentiment contained in the text.

Speech conveys information not only through audio prosody, but also through its phonetic content [14]. Both implicit linguistic prosody and explicit affective prosody are manifested over a segment of speech beyond the short-time speech frame. Some studies have attempted to extract rhythms from the textual content, Zhou at al.[15] proposed a semantic representation learning method based on graph neural network[16], considering dependency relations of the sentence to enhance expressiveness. And a character-level graph embedding is constructed in [17] to map the input text to graph embedding from time-domain to space-domain with the semantic information embedded. However, all of the above use simple structures designed only from the text and without considering deeper sentiment-related semantics and rich sentiment-related text.

Combining the aforementioned ideas of using semi-supervised approaches on a small emotional dataset and learning from text to enhance expressiveness, we propose a novel ESS model. This model is trained to learn emotional information from both speech and text to generate emotional speech. The contributions of this paper are as follows. I) We propose an implicit sentiment representations learning method based on the graph neural network, considering dependency relations of a sentence and text sentiment classification (TSC) task. II) For the lack of emotion-annotated datasets, we leverage large amounts of expressive datasets to reinforce training the proposed model with semi-supervised learning. III) To the best of our knowledge, this work is the first attempt to improve emotional speech quality using text rather than only explicit emotional representations from reference speech. The objective and subjective evaluation results demonstrate that our ESS system exhibits superior performance compared to the baseline model in terms of speech naturalness and emotion expressiveness.

The remainder of this paper is structured as follows. In Section 2, we describe our proposed method. Section 3 presents the experimental conditions and the results of the subjective and

† equal contribution
* corresponding author

10.21437/Interspeech.2022-11336

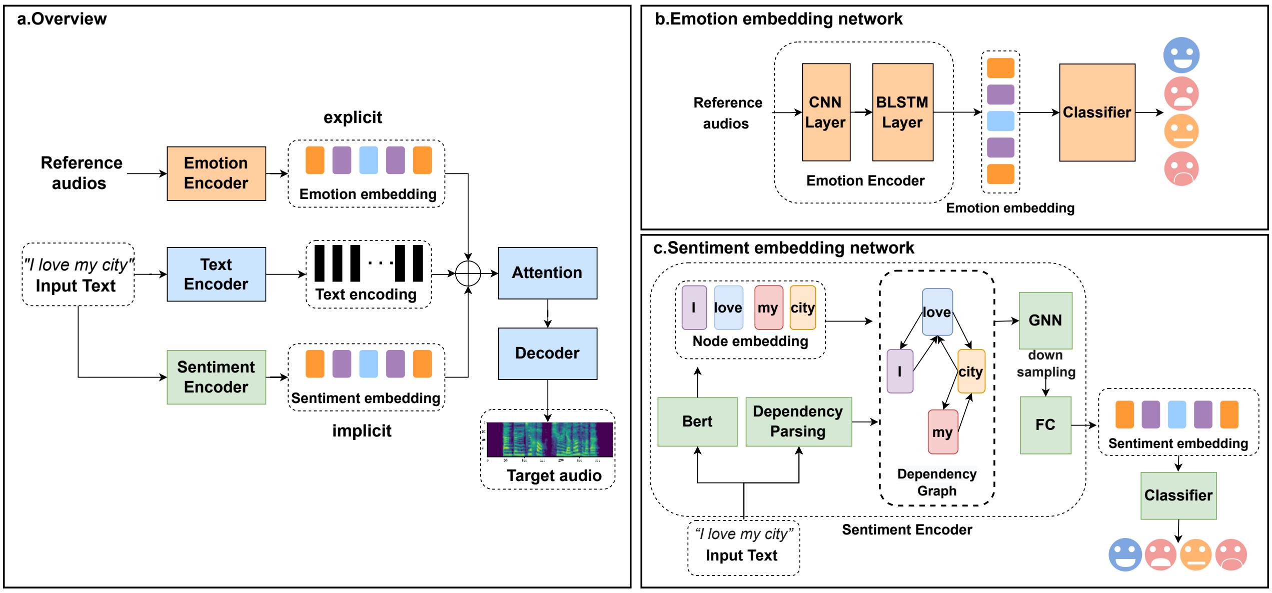


Figure 1: *An overview of the proposed model architecture. The emotion and sentiment classification network are used to extract the corresponding emotion and sentiment representations from the speech and text, respectively.*

objective experiments. Section 4 concludes the paper with our findings and our future work.

2. Proposed method

The proposed method, shown on the left side of Figure 1, is based on the end-to-end TTS architecture Tacotron2 [2]. For the text encoder input x, we used the character sequence of the normalized text for training. Apart from the text encoder, we use the emotion encoder and sentiment encoder to extract explicit emotion representations E_e and implicit sentiment representations S_e from reference Mel spectrum and text respectively. The learned representations (explicit and implicit) and text features T_e jointly dominate the generation of the final Mel spectrum as follows:

$$P(Mel_{out}|E_e, T_e, S_e; \theta) \tag{1}$$

As for representations learning, the proposed model contains two modules to provide emotional information during speech generation, which is learned from text inputs and extracted from reference audios.

2.1. Explicit emotion representations

The explicit emotion representations learning module is shown on the top right of Figure 1, which is comprised of an emotion encoder and an emotion classifier. We use this network to construct an emotion embedding space learned from reference audio samples through the speech emotion recognition task. And we use the explicit emotion representations learned from the emotion encoder to perform emotion transfer during inference.

The emotion encoder is composed of two 2D convolution layers and two Bidirectional Long Short-Term Memory (BLSTM) layers, and the last BLSTM state generates a 100-dimensional emotion representation. The classifier consists of a fully connection layer(FC) and a softmax layer. Finally, the softmax layer outputs the probability of four emotion types, i.e., neutral, happy, angry and sad.

2.2. Implicit sentiment representations

The implicit sentiment representation learning module is shown in the bottom right of Figure 1, which is also divided into two parts: a sentiment encoder and a classifier. And the sentiment encoder mainly consists of a graph construction module and a graph representation module. Like the emotion embedding network, we learn sentiment information from text through the text sentiment classification task. And we only use the sentiment encoder to extract implicit representations from the text during inference.

2.2.1. Graph construction

We use a graph to represent dependency relations of semantic tokens, defined as $G = (V, E)$. Node $V = v_1, v_2, \cdots, v_n$ represents the words with semantic information in a sentence. We use the bidirectional encoder representations from the Bidirectional Encoder Representations from Transformers (BERT) [18] through the text sentiment classification task to extract word-level semantic information from the words $W = w_1, w_2, \cdots, w_n$ in a sentence. Edge E denotes directed edge from node v_i to node v_j with a particular dependency relation. We extract semantic representations and the structure of the dependency tree from W, which can be expressed as:

$$\begin{aligned} [e_1, e_2, \cdots, e_n] &= \text{BERT}(w_1, w_2, \cdots, w_n) \\ E_{dep} &= \text{Dependency}(W) \\ V_{bert} &= [e_1, e_2, \cdots, e_n] \end{aligned} \tag{2}$$

where the e_i is the feature of the i-th word from the Bert, E_{dep} is the edge set of relations from dependency parsing and V_{bert} is set of nodes.

And in graph construction, we use the bidirectional dependencies [17]. As shown in Figure 1 of the sentence ”I love my city”, the node ”city” points to the node ”my” in the original dependency tree. However, we take into account that words have modifying relations with each other. The semantic information of the child node will also affect the parent node. Therefore, we adopt the directions of two information flows to construct the dependency graph.

2.2.2. *Graph representation learning*

To better capture the semantic information of long sentences, we use the Gated Graph Neural Network(GGNN) [19], which maps the dependency graph to semantic representations through information dissemination. The word-level BERT feature e_i is taken as the source state, and the message is passed according to the adjacency matrix to aggregate the message to the target node. Then, we obtain the word-level semantic representation through GGNN.

$$S_w = \text{GGNN}(G) \tag{3}$$

where the S_w is the word-level embeddings.

Because we want the sentiment representation at the sentence level, we need to conduct a graph pooling for downsampling and obtain a 100-dimensional semantic representation after passing through a fully connection layer(FC). The process is as follows:

$$\begin{aligned} S_{sent} &= \text{Pooling}(S_w) \\ S_e &= \text{FC}(S_{sent}) \end{aligned} \tag{4}$$

2.3. Semi-supervised training

The lack of emotion-annotated datasets is one of the main obstacles that limit the research of emotional speech synthesis. Therefore, we designed a semi-supervised training to leverage both annotated emotional speech data and other un-annotated emotional speech or text data. And the various types of data used in our method are listed as follows, where x means the text and y means speech:

- $D_A = \{x_a, y_a\}$ means small annotated emotional speech data (with text transcript).
- $D_T = \{x_t\}$ means annotated sentiment text data (without audio).
- $D_S = \{y_s\}$ means annotated emotional speech data (without text transcript).
- $D_G = \{x_g, y_g\}$ means general speech data (without emotional annotates).
- $D_N = \{x_n, y_n\}$ means expressive speech data (without emotional annotates).

The general speech data D_G is used to learn alignment, and single modality text data D_T and speech data D_S are used to pre-train the sentiment classification and emotion recognition models for text and speech, respectively. This semi-supervised training process is shown in Algorithm 1.

Note that we adopt two semi-supervised strategies: S-first and G-first. S-first means to first train the emotion encoder with D_A data and then train the sentiment encoder with D_N data, while G-first means the reverse training order. The above two different strategies are mainly used to distinguish the effectiveness of unlabeled and labeled data in semi-supervised training.

3. Experiments and results

3.1. Experimental Setup

Datasets. We conducted experiments on five different corpora. We use IEMOCAP [20] as (D_S) to pre-train SER model and use selected GoEmotions as (D_T) to pre-train TSC model. Then,we train our model with LJSpeech [21] as D_G, Blizzard Challenge 2013 (BC) [22] as D_N, and Emotional Speech Datasets (ESD) [23] as D_A. The details of the corpora are listed in Table 1.

Algorithm 1 Semi-training algorithm.

```
Input: Datasets: $D_A, D_T, D_S, D_G, D_N$,
    $\{SER, TSC\}$ ← initialization with random weights.
Output: The trained proposed ESS model
  Pre-training:
  Pre-train SER model with $\{D_S\}$
  Pre-train TSC model with $\{D_T\}$
  Training: Train ESS model
  $\{ESS\}$ ← initialization with pre-train sentiment and emotion encoder's parameters of SER and TSC.
  Train ESS model with $\{D_G\}$, while fix the parameters of sentiment and emotion encoder.
  if S-first then
      Train ESS model with $\{D_A\}$, while only fix the parameters of sentiment encoder.
      Train ESS model with $\{D_N\}$, while only fix the parameters of emotion encoder.
  else if G-first then
      Train ESS model with $\{D_N\}$, while only fix the parameters of emotion encoder.
      Train ESS model with $\{D_A\}$, while only fix the parameters of sentiment encoder.
  end if
```

Table 1: *Details of the corpus mainly used in our experiments. The emotional categories are respectively are neutral(Ne),sad(Sa),happy(Ha),angry(an) and expressive(Ex).*

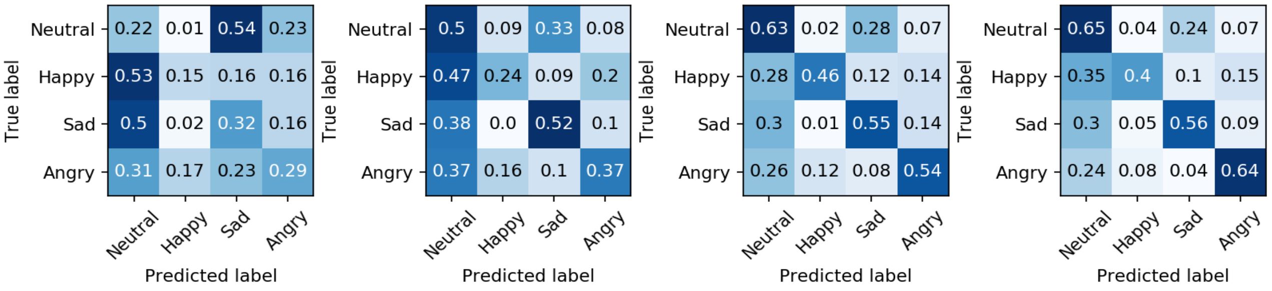
	Dur(h)	**Sent**	**Emo**
LJSpeech	23.4	13100	Ne
BC	16.8	8248	Ex
ESD	8.5	14,000	Ne,Sa,Ha,An
IEMOCAP	12	36600	Ne,Sa,Ha,An
GoEmotions	-	5525	Ne,Sa,Ha,An

Method. We trained the following three models:

- **T-GST:** We use GST-Tacotron [8] as the baseline.
- **T-SER:** Only using the SER model to learn emotion information like the SER part of [14]. We also use this model as another baseline.
- **T-G-S:** Our proposed model that learns sentiment and emotion information from both text and audio. (G-first)
- **T-S-G:** Our proposed model that learns emotion and sentiment information from both audio and text. (S-first)

Evaluation metrics. In terms of subjective evaluation for naturalness, the mean opinion score (MOS) was calculated on a scale from 1 to 5 with 0.5-point increments. To subjectively evaluate the emotion expressiveness performance of our proposed method, the emotion classification test was conducted, and the synthetic utterances of all emotions were played in random order. Each subject was asked to choose the emotion they perceived for each utterance.

We also conducted an ABX test. The rating criterion was determined by answering the question “Which one's speaking emotion is closer to the target audio emotion?” with one of three choices: (1) the first is better, (2) the second is better, and (3) neutral. In all tests, 10 native listeners were asked to rate the performance of 40 randomly selected synthesized utterances from the test set.



(a) T-GST (b) T-SER (c) T-G-S (d) T-S-G

Figure 2: *Confusion matrices of synthesized speech from T-GST,T-SER,T-G-S and T-S-G. The X-axis and Y-axis of subfigures represent predicted and truth emotion label, respectively.*

Table 2: *MOS results with 95% confidence intervals computed from the t-distribution.*

Emotion	T-GST	T-SER	T-G-S	T-S-G
Neutral	3.14±0.32	3.54±0.24	3.55±0.27	3.80±0.23
Sad	3.40±0.27	3.65±0.20	3.70±0.20	3.73±0.18
Happy	3.40±0.32	3.43±0.27	3.48±0.25	3.62±0.25
Angry	2.72±0.25	3.18±0.30	3.65±0.30	3.98±0.20

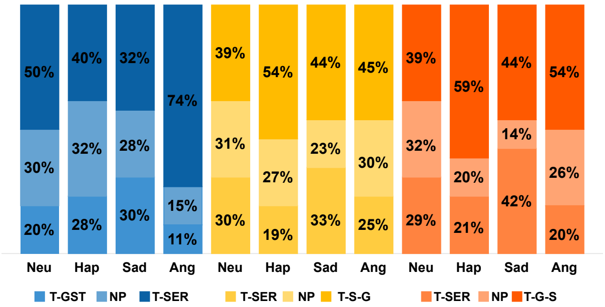


Figure 3: *The preference test results between T-GST vs. T-SER and T-SER vs. proposed model.*

3.2. Result and analysis

Subjective evaluation of speech naturalness. The results of the MOS test are presented in Table 2. The proposed model T-S-G and T-G-S demonstrated a significantly better naturalness than the baseline models T-SER and T-GST. This result shows the advantage of learning emotion representations from both text and audio rather than only audio. Besides, the speech quality of T-S-G is slightly better than T-G-S. This result mainly stems from two reasons: on the one hand, the audio quality of expressive data BC is much better than ESD; on the other hand, we can learn implicit sentiment from the text of BC, which improves the expressiveness of synthesis speech. In a word, refining the model with BC instead of ESD at the final step could avoid the negative effects of the low audio quality and low text coverage of the small annotated emotional speech data.

Subjective sentiment classification test for emotional expressiveness. Figure 2 shows the confusion matrices of the subjective emotion prediction results. From Figure 2, we can see that there were plenty of recognition errors in the baseline models. Of course, the T-GST baseline gets the worst performance in this test. This result means that it is difficult for the model to emotion control and transfer without the constraints of emotion labels. When learning emotion information from both text and audio as the proposed model, the confusion matrix appeared in a clear diagonal form. What's more, listeners are often confused between the synthesized speech of happy and neutral. This may be a problem in the training corpus, and it is difficult to build a corpus with a high degree of discrimination for each emotion.

Subjective evaluation of ABX test for emotion expressiveness. Figure 3 shows the results of the ABX test. A gap between our proposed model (T-G-S and T-S-G) and the baseline models (T-GST and T-SER) is visible. This shows that the proposed model can produce better latent emotion representations, which results in better emotion expressiveness. The results of the ABX test for the baseline model show that T-SER was much better than T-GST, and there is a significant difference. This result proves that using the sentiment classification task in the pre-train model is necessary. Comparing T-G-S and T-S-G, refining the model with BC instead of ESD at the final step could improve the emotional synthesis model performance.

In summary, using explicit and implicit information learned from audio and text can more accurately model emotional representations. In addition, the semi-supervised training algorithm can reduce the problem of lankness of annotated datasets. Evaluations showed that our proposed model could synthesize emotional speech with more natural and expressive. We present synthetic samples at https://xuyouning.github.io/ESS/demo.html

4. Conclusions

Emotional speech synthesis is challenging due to labeled data sparsity. In addition, there is much more sentiment text and expressive speech data than annotated emotional speech dataset. In this paper, we propose a novel model to improve emotional speech synthesis quality by learning explicit and implicit representations with semi-supervised learning. In addition to explicit emotional representations, we propose an implicit sentiment representations learning method based on the graph neural network, considering dependency relations of a sentence and text sentiment classification (TSC) task. For the lack of emotion-annotated datasets, we leverage large amounts of expressive datasets to reinforce training the proposed model with semi-supervised learning. Experiments show that the proposed method can improve the naturalness and expressiveness of synthetic speech and is better than the baseline model.

5. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grant 62176182.

6. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, Z. Y. Jaitly, Y. Xiao, Z. Chen, S. Bengio, Q. Le *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017.

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

[3] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706–6713.

[4] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "An investigation to transplant emotional expressions in dnn-based tts synthesis," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1253–1258, 2017.

[5] H. Choi, S.Park, J.Park, and M. Hahn, "Multi-speaker emotional acoustic modeling for cnn-based speech synthesis," *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6950–6954, 2019.

[6] S.-Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7254–7258.

[7] M. Schröder, "Emotional speech synthesis: A review," in *Seventh European Conference on Speech Communication and Technology*. Citeseer, 2001.

[8] 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*. PMLR, 2018, pp. 5180–5189.

[9] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Processing Letters*, vol. 26, no. 9, pp. 1383–1387, 2019.

[10] 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," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 5734–5738.

[11] N. Tits, K. El Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional tts," in *Intelligent Systems and Applications*, Y. Bi, R. Bhatia, and S. Kapoor, Eds. Cham: Springer International Publishing, 2020, pp. 52–60.

[12] 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 *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2019, pp. 623–627.

[13] Y. Gao, W. Zheng, Z. Yang, T. Kohler, C. Fuegen, and Q. He, "Interactive text-to-speech via semi-supervised style transfer learning," *onikle*, 2020.

[14] R. Liu, B. Sisman, G. Gao, and H. Li, "Expressive tts training with frame and style reconstruction loss," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1806–1818, 2021.

[15] Y. Zhou, C. Song, J. Li, Z. Wu, and H. Meng, "Dependency parsing based semantic representation learning with graph neural network for enhancing expressiveness of text-to-speech," *arXiv preprint arXiv:2104.06835*, 2021.

[16] H. Choi, S.Park, J.Park, and M. Hahn, "J.zhou and g.cui and z.zhang and c.yang and z.liu and l.wang and c. li and m. sun," *arXiv preprint arXiv:1812.08434*, 2018.

[17] A. Sun, J. Wang, N. Cheng, H. Peng, Z. Zeng, and J. Xiao, "Graphtts: graph-to-sequence modelling in neural text-to-speech," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6719–6723.

[18] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *Universal Language Model Fine-tuning for Text Classification*, p. 278, 2018.

[19] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow, "Gated graph sequence neural networks," in *Proceedings of ICLR'16*, 2016.

[20] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[21] K. Ito, "The lj speech dataset," *https://keithito.com/LJ-Speech-Dataset/*, 2017.

[22] S. King and V. Karaiskos, "The blizzard challenge 2013," *SynSIG*, 2013.

[23] K. Zhou, B. Sismah, 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.