Interspeech 2022
18-22 September 2022, Incheon, Korea

Text aware Emotional Text-to-speech with BERT

Arijit Mukherjee[†], *Shubham Bansal*[†], *Sandeepkumar Satpal, Rupeshkumar Mehta*

Microsoft STC India

{armuk, shbansa, ssatpal, rupeshme}@microsoft.com

Abstract

Emotional text to speech is the idea of synthesizing emotional audio via a text-to-speech model. With neural text-to-speech, sentence-level naturalness has improved a lot and is almost at par with human speech, but the current approach to emotional text-to-speech models heavily relies on the user to input the expected emotion along with the text to synthesize the desired speech. In this work, we propose a novel text-aware emotional text-to-speech system that leverages a pre-trained BERT model to get a deep representation of the emotional context from the text both during training and inference. We show that our proposed method synthesizes emotional audio with emotion depending on the emotional context of the input text. We also show that our method outperforms baseline systems in varying the emotional intensity depending on the text.

Index Terms: text-to-speech , emotional TTS, text embeddings, BERT

1. Introduction and Related work

Recent advancements in the area of neural text-to-speech (TTS) [1, 2, 3] have led to almost human parity in terms of the naturalness of speech. Additionally, there has been huge interest in building an emotional and highly expressive TTS. The objective of text-aware emotional TTS in our work is to intelligently synthesize the emotional speech depending on the emotion of the corresponding text. The majority of the work in emotional TTS proposes different techniques to learn a representative embedding for each emotion and use these learned embeddings to control the emotion of the synthesized speech. [4] proposes to use the one-hot representation of the emotion label as the emotion embedding during the training. Further, global style tokens (GST) [5] based framework has also been used to model emotion embedding. [6] proposes using the centroid of GST embeddings for each emotion and [7] shows that the GST embeddings for each emotion could be highly dispersed and proposes using an inter-to-intra distance ratio algorithm that considers the distribution of each emotion.

It is to be observed that the above-discussed approaches rely on the user to input an appropriate emotion for each text and hence the scalability and automation of emotional TTS becomes challenging when applied to a wide range of scenarios such as call center, audio-book reading, etc. In one of the earliest works, TP-GST [8] learns to predict stylistic renderings from text alone, neither needs explicit labels during training but focuses majorly on improving the overall expressiveness of speech. In one of the recent works, [9] proposes jointly training a ResNet-like text classification network on the given dataset to predict the emotion label from the input text. During inference, the predicted emotion label is automatically mapped to the latent representation as per the embedding table learned during the training. Additionally, [10] proposes training the text-based local strength predictor and utterance variance predictor to model the fine-grained emotion on the basis of the text.

In the majority of the emotional speech datasets, audio emotion may not always agree with the emotion of the corresponding text as also observed by [9, 10]. We henceforth refer to this problem as text-audio emotion disagreement. Such disagreements may lead to inaccurate emotion prediction from the text during inference when a text emotion classifier is learned directly from the emotional speech dataset as described in [9]. Similarly, local strength predictor and utterance variance predictor learning in [10] may also not be optimal. In our work, we propose techniques to handle the audio-text emotion disagreement in the emotional speech datasets.

Bidirectional encoder representations for transformers (BERT) model [11] is a pre-trained encoder-transformer model and it produces powerful context representations and provides the state-of-art results in several NLP tasks including text emotion classification [12, 13, 14]. BERT has also been successfully used to improve the prosody of the synthesized speech in [15]. To explicitly model the emotion of the text, several high-quality datasets [16, 12, 17] have been proposed. Training a text emotion classification model and using the predicted label's one-hot or fixed embedding representation as an input to the emotional TTS model [6, 9] often results in an averaged expressiveness for each emotion. In our work, we train the BERT-based text classification model on GoEmotions dataset [12] and propose to directly use the embedding from its penultimate layer rather than the emotion label both during the training and inference. Such a technique offers three-fold advantages: 1. BERT embedding may capture the fine-grained as well as global representation of the emotion of given text with a single embedding unlike [10, 18] . 2. Emotion label present in the emotional speech dataset does not necessarily need its corresponding label in the text classification model. 3. If we assume that there is minimal audio-text emotion disagreement, there is no need for labeled data as model training and inference does not rely on the emotion labels. Although our text classifier is pre-trained on an external text emotion dataset [12], we still need to handle the audio-text emotion disagreement in the proposed system because we use the BERT embeddings directly during the training.

We show that our model is preferred in subjective evaluations in varying the emotion intensity according to the text over the baselines.

2. Neural Text to Speech

Neural text-to-speech uses deep learning based methods to synthesize audios from a given text. It involves mainly two components: **(1) Frontend :** which minimally includes a text normalizer to normalize the text followed by a lexicon and grapheme-to-phoneme model. **(2) Backend :** mainly consists of an acoustic model that takes sequence of characters or phonemes and

†: Both authors have contributed equally.

10.21437/Interspeech.2022-11293

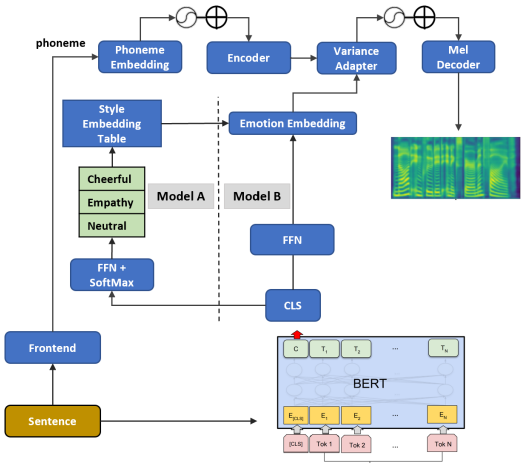

Figure 1: *Text Aware Emotional TTS Model A represents Text Classifier + Multi Emotion model 3.2.2 and Model B represents Text Aware Emotional TTS with BERT Embedding 3.3*

converts them to a mel-spectrogram followed by a wave generation model also called as vocoder which converts the mel-spectrogram to its corresponding waveform. In all our experiments, we use Fastspeech 2 [3] as our acoustic model which is a non auto regressive transformer based model [19] and MelGAN [20] as our vocoder.

3. System Overview

In this section, we first describe our approach to text emotion modeling. We then discuss our baseline systems followed by our proposed text-aware emotional TTS with BERT embeddings. All the described methods synthesize the emotional speech directly from the text without any manual emotion label during the inference.

3.1. BERT-based Text Emotion Modelling

In recent years, several model architectures[12, 13, 14, 21, 22] have been proposed for text emotion modeling. Irrespective of the chosen model, a large amount of highly accurate emotion-labeled text is needed to train them. The majority of the emotional speech datasets have the following limitations: 1. Emotion labels are indicative of audio emotion and may not be directly applicable to its corresponding text. 2. Collecting a large emotional speech dataset is expensive as well as laborious and hence there might not be enough transcripts available in the form of textual data to train a high-quality text emotion classifier.

In our text modeling approach, we leverage a pre-trained BERT model to build our text emotion classification model. We use the final hidden states h of the first token $[CLS]$ as the representation of the whole text. We henceforth refer to this representation as BERT embedding. We then add a standard feed-forward layer and softmax on top of the BERT embedding to predict the probability of an emotion label c. Finally, we finetune the entire network on GoEmotions[12] dataset containing approximately 58K emotion labeled text by minimizing the cross-entropy loss. We use the sentiment mapping suggested in[12] and map positive, neutral, negative to cheerful, neutral, and empathy emotions respectively and we ignore the ambiguous labeled data in our training. We get an accuracy of roughly 74% and an F1 score of 0.64. We freeze the model weights and use this same model for all our experiments.

3.2. Baselines

3.2.1. Implicit data pooled model

Recent neural TTS architectures like Fastspeech2 [3] extract duration, pitch, and energy from the speech waveform and learns to model them from the encoded phoneme embeddings during the training. Since these properties are highly representative of emotion, we hypothesize that the neural TTS network might be able to implicitly learn to generate emotive speech from the text if trained with the emotional speech dataset. In this approach, we use the neural TTS architecture as defined in Section 2 which serves as the simplest baseline.

3.2.2. Text classifier + Multi-emotion model

In this baseline, we follow the approach similar to [9] but we use an external BERT-based text emotion classification model rather than training the text emotion classifier afresh on the emotional speech data due to the reasons stated in Section 3.1. We present the model architecture in Figure 1. During training, the audio emotion label from the emotional speech dataset is used as a conditional input to the emotion embedding table instead of using the label predicted by the text emotion classification model for better emotion modeling and to avoid the text-audio emotion disagreement. During inference, we use this learned embedding table to map the predicted text emotion to a fixed size emotion embedding. Since phonemes and their corresponding position information are already encoded in the encoder output, we concatenate the encoder output for each phoneme with the emotion embedding and use a simple feed-forward network to model the phoneme-dependent emotion embedding for the given sentence.

We also observe that in this approach, it is necessary that all the emotions present in the emotional speech dataset must have its corresponding emotion label in the text emotion classification model and vice-versa. We consider this requirement a major bottleneck in scaling the text-aware emotional TTS across different emotional speech datasets and scenarios. We address this with our proposed approach in the following Section 3.3.

3.3. Text-aware Emotional TTS with BERT Embeddings

BERT encodes the bidirectional representation of text and $[CLS]$ token embedding captures both the semantic and syntactic meaning of a given text. We hypothesize that fine-tuning the BERT model on a downstream text emotion classification task results in diverse BERT embeddings capable of representing the variations in the emotion of the text. In this approach, we propose to concatenate the BERT embedding extracted from the text with the encoder output of each phone after applying a feed-forward layer as shown in Figure 1. Similar to the multi-emotion model, we use another feed-forward network to model the phoneme-dependent emotion embeddings for the given sentence.

Another advantage is that, unlike the multi-emotion model, this model relies on using the BERT embeddings directly, and hence it is not necessary that corresponding to all the emotions present in the emotional speech dataset, there must be a corresponding emotion label in the text emotion classification model and vice-versa. Thus, a universal text classifier model could be built to be used across all emotional speech datasets.

This approach also simplifies the training and inference by relying solely on the BERT embedding but suffers from a major

limitation: Text-audio emotion disagreement. In the majority of the emotional speech datasets, there is no guarantee that the text is spoken in the most appropriate emotion and hence resulting in the jumbled-up BERT embeddings. We visualize these BERT embeddings in 2-D TSNE space with respect to their audio emotion labels for 2 different emotional speech datasets in Figure 2. We propose ways to handle the text-audio emotion disagreement in the subsequent subsection.

3.3.1. Handling text-audio emotion disagreement

Algorithm 1 Emotion Speech Data Filtering

Input: speech data $D \leftarrow \forall < text, audio, emotion >$
Output: Filtered training data $O \leftarrow []$
Init M: $Textclassifier Model()$
if All speech emotion class exists in M **then**
  for all $d \in D$ **do**
    $emo_{text} =$ **M**(d_{text})
    if $emo_{text} == d_{emotion}$ **then**
      $O.add(d)$
    end if
  end for
else
  Init I: $I \leftarrow []$
  for all $d \in D$ **do**
    $emb_{text} =$ **ExtractBERTEmbedding**(d_{text}, M)
    $I.add(< emb_{text}, d_{emotion} >)$
  end for
  Train highly regularized classifier F_c with I
  for all $d \in D$ **do**
    $emb_{text} =$ **ExtractBERTEmbedding**(d_{text}, M)
    $Pred_{emo} =$ **PredictEmotion**(emb_{text}, F_c)
    if $Pred_{emo} == d_{emotion}$ **then**
      $O.add(d)$
    end if
  end for
end if

In the above algorithm 1, we propose two ways of handling audio-text emotion disagreement: If corresponding to all the emotions present in the emotional speech dataset, there is a mapped emotion label from the text emotion classifier, we follow the first approach. If there are certain emotional speech labels that cannot be mapped to any of the output labels of the text classifier, we train a highly regularized classifier where we learn to predict the audio emotion label from the BERT embedding of a given text. We observe that this highly regularized classifier can act as a noise filter to remove the disagreeing samples. Both the approaches resulted in well-separated clusters of BERT embeddings. We visualize these embeddings with respect to the audio emotion label as shown in Figure 3(b).

4. Experiments & Results

In this section, we will compare our proposed approach for text-aware emotional TTS with the baseline systems presented in Section 3.2. The text emotion classifier is described in Section 3.1. The frontend and vocoder remains the same across all our experiments.

4.1. Training Data

In our experiments, we use two internal single-speaker emotional speech datasets in en-US accent as described below:
SpeakerF1 : This set consists of emotional speech recordings in "Cheerful" and "Empathy" emotions with 200 utterances each. Since, there is no "Neutral" emotion audios and "Cheerful" , "Empathy" emotion are opposite in polarity, we hope that there is less text-audio emotion disagreement in this dataset.
SpeakerF2 : This set consists of emotional speech recordings in three emotions: "Cheerful", "Neutral" and "Empathy" with 3900, 4900 and 4300 utterances in each of them respectively.

4.2. Implicit Data-pooled Model vs Text-Aware Emotional TTS with BERT Embeddings

The aim of this experiment is to validate that if an additional text embedding performs better than the implicit modeling capabilities of FastSpeech 2 [3]. We thus compare the approaches described in Section 3.2.1 and section 3.3. For "SpeakerF1", since, 400 utterances are not sufficient to train a Neural TTS model, we fine-tune on top of a base model trained on multiple speakers. For "SpeakerF2", given the sufficient number of utterances, we train the Neural TTS model without any initialization. We discuss the results in the subsequent section.

4.2.1. Results & Analysis

To evaluate, we perform a style CMOS with a set of 15 native en-US judges to compare both the systems and score in the range of [+3,-3] based on the following question: **"Which audio has more appropriate emotion depending on its corresponding text?"**. Positive score suggest that 3.3 is better than 3.2.1. Results are shown in Table 1.

Experiment Data	Implicit vs Text-aware Style CMOS
SpeakerF1	+0.491
SpeakerF2	-0.059
SpeakerF2 (After filtering 3.3.1)	+0.408

Table 1: *Text-aware style CMOS for two speakers, higher number means our model is better than the baseline*

From table 1, with text embedding, we observe significant improvement for SpeakerF1 but marginal regression for SpeakerF2. We hypothesize that this is because of the text-audio emotion disagreement. To validate our hypothesis, we have analyzed the 2D TSNE [23] projection of BERT embeddings for all the texts in both the datasets with respect to their audio emotion label. Figure 2 shows that for SpeakerF1, BERT embedding clusters are well separable with respect to their audio emotion label, which is not observed in the case of SpeakerF2. In the next section, we discuss the results after handling these text-audio emotion disagreements.

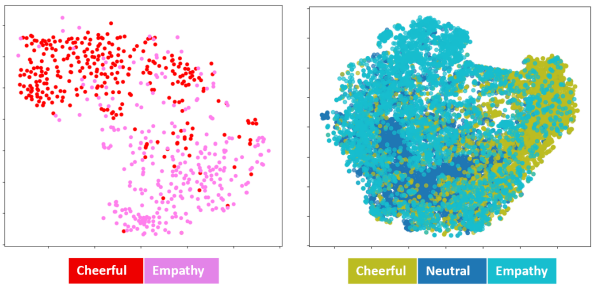

Figure 2: *TSNE Visualization of BERT embedding (a) Left: SpeakerF1 (b) Right: Speaker F2. Each point represents a training data point, the color of the point signifies the audio emotion label in the training data*

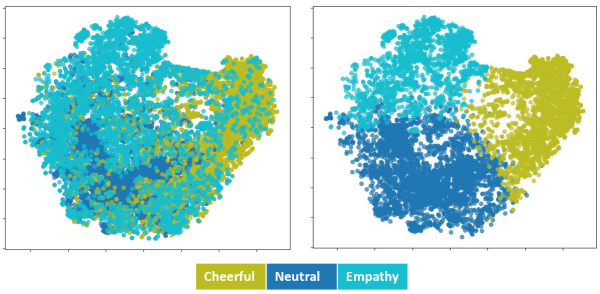

Figure 3: *TSNE Visualization of text embedding for data of SpeakerF1 (a) Left: Before and (b) Right: After handling text-audio emotion disagreement*

4.3. Handling Text-Audio Emotion disagreement

For SpeakerF2, where we observed a significant amount of text-audio emotion disagreement, we identified and filtered out 36% of the training data with the algorithm proposed in Section 3.3.1. We re-plot the TSNE visualization of selected data for SpeakerF2 training in Figure 3(b). We observe that the BERT embedding clusters are now well separable with respect to their audio emotion labels. To validate the effect of improved clustering, we train another model with our proposed approach in Section 3.3 and compare it against the Implicit data-pooled model described in Section 3.2.1. We observe that the style CMOS gets significantly improved to **+0.408** from **-0.059** justifying the need to handle text-audio emotion disagreement.

4.4. Text Classifier + Multi-emotion model vs Text-Aware Emotional TTS with BERT Embeddings

In this experiment, we compare our proposed approach of text-aware with BERT embeddings (we refer to it as TABE) described in Section 3.3 against the text-classifier+multi-emotion approach (we refer to it as TCME) described in Section 3.2.2 for speeakerF2. For TABE, we use the model trained on data obtained after handling the text-audio emotion disagreement as described in Section 4.3 and TCME is trained on the entire data because we use audio emotion label during training and hence TCME does not suffer from the disagreements.

4.4.1. Results & Analysis

Our test set comprises 100 sentences split into five categories based on its text emotion: (1) Cheerful, (2) Neutral Cheerful, (3) Neutral, (4) Neutral Empathy, and (5) Empathy. Such a test-set helps us to judge the model's ability to produce more fine-grained and diverse emotion in the final audio as per the emotion of the text. This test set has been created by consensus voting among several language experts. Following are the results of the tests performed:

(1) End-to-end emotion accuracy : In this test, we evaluate with the set of 5 native judges. Judges are provided with the following task: **"Assign a label among either of the 5 buckets on the basis of how the audio emotion is being perceived."**.

In Table 2, we observe that in the case of "Neutral-Cheerful" and "Neutral-Empathy" text-set, TABE is able to outperform TCME by a significant margin in emotion accuracy, whereas for "Neutral" text-set, we observe slight regression with TABE compared to TCME, as some of the neutral sentences are synthesized with "Neutral Cheerful" and "Neutral Empathy" emotion. We hypothesize that such behavior in TABE is due to the fact that TABE produces a diverse set of emotions whereas TCME synthesizes consistent emotion for the "Neutral" text.

Text Aware with BERT Embedding (TABE)					
Emotion	Cheerful	Empathy	Neutral	Neutral Cheerful	Neutral Empathy
Cheerful	**95.00**	0.00	2.50	1.25	1.25
Empathy	2.50	**90.00**	1.25	2.50	3.75
Neutral	0.00	1.25	**87.50**	3.75	7.50
Neutral Cheerful	12.50	3.75	53.75	**26.25**	3.75
Neutral Empathy	0.00	47.50	21.25	0.00	**31.25**
Text classifier + Multi-emotion (TCME)					
Emotion	Cheerful	Empathy	Neutral	Neutral Cheerful	Neutral Empathy
Cheerful	**95.00**	0.00	1.25	2.50	1.25
Empathy	2.50	**92.50**	2.50	2.50	0.00
Neutral	0.00	0.00	**97.50**	0.00	2.50
Neutral Cheerful	28.75	5.00	58.75	7.50	0.00
Neutral Empathy	0.00	65.00	20.00	0.00	15.00

Table 2: *Percentage of assignment of audio emotion. Each row and column corresponds to text and audio emotion respectively*

(2) Overall emotion preference In this test, we evaluate with a set of fifteen judges. Judges are asked the following question: **"Which audio emotion is preferred for the given text?"**. Judges are also allowed to mark both the candidates as "Equal". In Table 3, we observe that for "Neutral" and "Empathy" text-set, both candidates are preferred a similar number of times, whereas, in the case of "Cheerful", "Neutral-Empathy", and "Neutral-Cheerful" text-set, we see a sizeable preference for TABE.

Text Emotion	**TCME Win**	**TABE Win**	**Equal**
Cheerful	15.0%	35.0%	**50.0%**
Empathy	30.0%	30.0 %	**40.0%**
Neutral	**40.0%**	**40.0%**	20.0%
Neutral Cheerful	25.0%	**45.0%**	30.0%
Neutral Empathy	30.0%	**45.0%**	25.0%
Total	28.0%	**39.0%**	33.0%

Table 3: *Average percentage preference in each category*

(3) MOS with Neutral voice: We also perform a MOS test to validate if our proposed text-aware emotional TTS with BERT embeddings is comparable to the neutral TTS voice in terms of overall voice quality. From Table 4, we observe that all the 3 models are almost at par with each other in terms of overall quality.

TCME	**TABE**	**Neutral TTS**
4.36	4.33	4.39

Table 4: *MOS between TCME, TABE and Neutral TTS*

5. Conclusion

We propose a text-aware emotional TTS architecture capable of identifying the emotion variation from the input text with the help of a pre-trained BERT model and synthesize the audio with the correct emotion. We also discuss the challenges of text-audio emotion disagreement while training such models along with a filtering-based approach to solve such challenge. We further showed that our proposed method can achieve better emotion accuracy while not regressing on the overall MOS of the synthesized audio.

6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. J. Skerry-Ryan,

R. A. Saurous, Y. Agiomyrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," *CoRR*, vol. abs/1712.05884, 2017. [Online]. Available: http://arxiv.org/abs/1712.05884

[2] N. Li, S. Liu, Y. Liu, S. Zhao, M. Liu, and M. Zhou, "Close to human quality TTS with transformer," *CoRR*, vol. abs/1809.08895, 2018. [Online]. Available: http://arxiv.org/abs/1809.08895

[3] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," 2020. [Online]. Available: https://arxiv.org/abs/2006.04558

[4] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *arXiv preprint arXiv:1711.05447*, 2017.

[5] 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.

[6] 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.

[7] 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.

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

[9] C. Cui, Y. Ren, J. Liu, F. Chen, R. Huang, M. Lei, and Z. Zhao, "Emovie: A mandarin emotion speech dataset with a simple emotional text-to-speech model," *arXiv preprint arXiv:2106.09317*, 2021.

[10] Y. Lei, S. Yang, X. Wang, and L. Xie, "Msemotts: Multi-scale emotion transfer, prediction, and control for emotional speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 853–864, 2022.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[12] D. Demszky, Dorottya andA Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "Goemotions: A dataset of fine-grained emotions," 2020. [Online]. Available: https://arxiv.org/abs/2005.00547

[13] M. Munikar, S. Shakya, and A. Shrestha, "Fine-grained sentiment classification using BERT," *CoRR*, vol. abs/1910.03474, 2019. [Online]. Available: http://arxiv.org/abs/1910.03474

[14] L. Luo and Y. Wang, "Emotionx-hsu: Adopting pre-trained BERT for emotion classification," *CoRR*, vol. abs/1907.09669, 2019. [Online]. Available: http://arxiv.org/abs/1907.09669

[15] L. Chen, Y. Deng, X. Wang, F. K. Soong, and L. He, "Speech bert embedding for improving prosody in neural tts," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6563–6567.

[16] F. Barbieri, J. Camacho-Collados, L. Neves, and L. E. Anke, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," *CoRR*, vol. abs/2010.12421, 2020. [Online]. Available: https://arxiv.org/abs/2010.12421

[17] S. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko, "SemEval-2018 task 1: Affect in tweets," in *Proceedings of The 12th International Workshop on Semantic Evaluation*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 1–17. [Online]. Available: https://aclanthology.org/S18-1001

[18] Y. Lei, S. Yang, and L. Xie, "Fine-grained emotion strength transfer, control and prediction for emotional speech synthesis," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 423–430.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[20] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brebisson, Y. Bengio, and A. Courville, "Melgan: Generative adversarial networks for conditional waveform synthesis," 2019. [Online]. Available: https://arxiv.org/abs/1910.06711

[21] L. Luo and Y. Wang, "Emotionx-hsu: Adopting pre-trained bert for emotion classification," *arXiv preprint arXiv:1907.09669*, 2019.

[22] F. A. Acheampong, H. Nunoo-Mensah, and W. Chen, "Transformer models for text-based emotion detection: a review of bert-based approaches," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 5789–5829, 2021.

[23] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html