Low-Resource VITS-Based Emotion Speech Synthesis Using KNN Algorithm

Zedong Xing[1[0000−1111−2222−3333]], Bicheng Xiong[1[0009−0005−2814−1504]], and Weiping Wen[2(✉)]

[1] Xiangtan University, Hunan Province, 411105,CN
202221632998@smail.xtu.edu.cn
[2] Peking University, Beijing, 100871, CN
weipingwen@pku.edu.cn

Abstract. Recently, the application scenarios of Text-to-Speech (TTS) have been expanding. Along with the simple replication of voice tones, there's a growing demand for multi-emotional speech. However, traditional emotional speech synthesis typically relies on a large amount of annotated data, which might be scarce in certain applications or domains. In this paper, we aim to synthesize multi-emotional speech with minimal neutral samples using the K-Nearest Neighbors (KNN) algorithm and an enhanced VITS model for speech synthesis. Experimental results indicate that our approach improves both emotional expressiveness and speech clarity compared to traditional methods. Additionally, it achieves low-cost multi-emotional speech synthesis, making it suitable for resource-constrained applications.

Keywords: Multi-emotional speech synthesis · Enhanced VITS model · KNN algorithm.

1 Introduction

In recent years, the field of deep learning in speech has witnessed rapid development, with significant advancements in speech synthesis[19]. A series of Text-to-Speech (TTS) models have emerged capable of faithfully reproducing the voice characteristics of target speakers at high quality and have found widespread application in various scenarios such as voice cloning[11, 17, 21]. Traditional speech models were limited to generating speech with a fixed tone. However, in real-life situations, people's speech carries different emotional nuances, with tones varying along with the content. Therefore, the practical effectiveness of traditional speech models appears rigid and mechanical. With the continuous complexity of usage scenarios, we aim for speech synthesis to achieve expressive multi-emotional expressions akin to human speech.

Currently, there are numerous studies focusing on expressive speech synthesis, aiming to convey various speaking styles and emotions [2], which can be categorized into supervised and unsupervised approaches. Supervised methods involve models trained on datasets annotated with emotion labels [6, 23], where

these labels guide the model training to learn accurate weights. The most direct way to represent annotated emotional labels from the dataset as inputs for TTS models is by using one-hot vectors, with common emotions (e.g., neutral, happy, sad, angry) as the values of the vectors. Such traditional multi-emotional speech synthesis still requires emotional data of the target speaker's voice as a reference. However, acquiring multi-emotional speech datasets for target voices is costly, making it challenging to obtain sufficient emotional data for specific emotions in most scenarios. Even with ample data, the effectiveness of utilizing pre-trained emotion classification models on old voice data to distinguish the emotions in new voice data without added emotional labels is not ideal. Hence, the challenge lies in how to acquire labeled emotional datasets with minimal cost. Due to the high cost of labeled datasets, many studies have proposed unsupervised methods, where models are trained to extract speaking styles or emotions from expressive speech data through unsupervised means [4] Unsupervised models typically utilize reference speech as input for TTS models. The TTS model extracts style or prosodic embeddings, which are then used to synthesize speech resembling the style of the input reference. However, the style representations of the hidden variables extracted from the reference speech are uninterpretable. In practical applications, it is challenging to accurately and conveniently select suitable references for arbitrary text content.

Regarding the above issue, we focus on the task of synthesizing multi-emotional speech with a low-cost approach, given access to a small amount of neutral speech samples from the target speaker. This entails achieving multi-emotional speech synthesis with the target speaker's voice at minimal expense, while automatically adapting emotions based on the text content. Previous research [1] has demonstrated that the complexity of language model processing is not linearly correlated with its effectiveness in classic tasks such as speech transformation. Simply utilizing the self-supervised features of pre-trained models for straightforward feature matching transformation can yield results comparable to those of more complex models. Inspired by this, we opt to integrate feature matching algorithms into the feedforward processing of text-to-speech.

In our study, we propose a **L**ow-**R**esource **V**ITS-Based **E**motion **S**peech synthesis called **LRVESpeech** , which improves upon the baseline TTS model by incorporating the k-nearest neighbor algorithm [9]. Specifically, we conducted two main training stages.

In the first stage, we implemented the emotion conversion module. Initially, we extracted feature sequences from both the source utterance and the reference utterance using a self-supervised speech representation model. Subsequently, we performed classification analysis using the k-nearest neighbor algorithm, replacing each frame of the source representation with the closest neighbor from the reference utterance, thus achieving the transformation to the target speaker. Finally, we used a neural vocoder to acoustically encode the transformed features to generate the target speaker's speech. This process was repeated, transforming the voice from the existing labeled emotion dataset into the target speaker's voice.

In the second training stage, we trained the VITS [11] model using the generated speech dataset. In this stage, we modified the input of the VITS model to include speaker ID, emotion, and text, further enhancing the model's performance and capabilities.

We summarize the contributions of our method as follows:

- We introduce a multi-emotion speech TTS based on the k-nearest neighbor algorithm, enabling simple and controllable synthesis of emotional speech.
- The proposed method allows for the synthesis of high-quality multi-emotion speech datasets by utilizing a small number of neutral speech samples from the target speaker.
- During the prediction process, due to the characteristics of VITS, the generated speech is diverse, resulting in high-quality speech without a robotic feel.

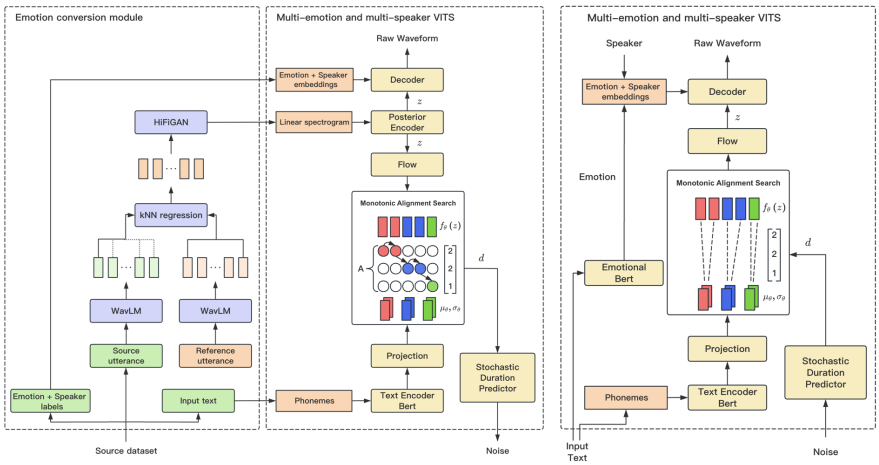
2 LRVESpeech

2.1 Model Overview

Our overall framework, as illustrated in Fig. 1, consists of a speech synthesis framework and an emotion conversion module. We choose the VITS model [11] as our speech synthesis framework and make several enhancements. In the decoder module, we incorporate speaker and emotion feature embeddings, which assist us in better controlling the timbre and emotion of the generated speech during the inference stage. Additionally, based on the findings of research [13], which suggest that incorporating pre-training information from NLP models can aid TTS in learning prosody for more natural-sounding speech, we introduce the BERT language model [8] to extract semantic features from the text, enhancing the model's prosodic expression capabilities. Since VITS is an end-to-end model that does not require complex multi-training processes, it simplifies our training process while ensuring the robustness of the system. The emotion conversion module utilizes the k-nearest neighbor algorithm to generate multi-emotion speech datasets from single speech samples.

Based on our proposed model framework, we need to follow these steps to achieve the synthesis of multi-emotion speech from a small number of neutral samples:

1. Pre-train the emotion conversion module by fine-tuning the VITS vocoder with neutral samples from the target speaker to improve the VITS's ability to represent target samples.
2. Set emotion conversion parameters to transform the source emotion dataset into a dataset with the target speaker's voice using k-nearest neighbor.
3. Train the improved VITS model using the generated emotion-labeled dataset to achieve the goal of synthesizing multi-emotion speech.



(a) Training process (b) Inference process

Fig. 1. Architecture of our method. As shown in (a), during the training process, we first utilize the neutral samples of the target speaker. Through the retrieval and matching process of the k-nearest neighbors (kNN) algorithm, we convert them into a multi-emotion dataset with emotion labels. Subsequently, we train the improved VITS model using this dataset. In the inference process depicted in (b), we directly label the input text with emotion and speaker tags, allowing the model to generate speech samples with emotions for the target speaker.

2.2 Emotion Conversion Module

The design of this module adopts an encoder-converter-vocoder framework similar to that used in [1]. Firstly, we utilize a self-supervised language model as the encoder component. We extract self-supervised vector representations from the source emotion dataset and aggregate them. Then, we extract self-supervised vector representations from the speech of the target speaker and form a lexicon. This allows for the extraction of both audio and text features simultaneously. In the converter component, we use the k-nearest neighbor algorithm to map the original emotion data frames to the nearest neighbors in the lexicon. Finally, the transformed vector features are decoded into audio linear spectrograms using a pre-trained vocoder. Below is a detailed description of each component.

Encoder. In the encoder stage, our objective is to extract relevant semantic information from the target speaker and the source emotion dataset using a self-supervised model, thereby establishing the source dataset SSL feature library and the target SSL feature library. Existing research suggests that self-supervised models perform well in various downstream tasks, including speaker identification and prosody extraction. To fully retain emotion information, we adopt the WavLM [3] self-supervised model, which has state-of-the-art performance in emotion-

related tasks. WavLM is a self-supervised model that jointly learns masked speech prediction and denoising, enabling it to handle end-to-end downstream speech tasks such as speaker verification, speech separation, and speech recognition. Previous studies [18] have found that compared to frames with different phonemes, the features from the 6th layer of WavLM can effectively map frames with the same phoneme closer together in feature space while preserving speaker timbre information. This implies that we can utilize the feature vectors from the 6th layer for speech conversion, preserving text information while changing speaker timbre. Therefore, we choose to replace frames at the 6th layer.

Similar to [1], throughout our entire experimental process, we directly use the pre-trained WavLM model without fine-tuning.

Retrieval Matching Process. In this stage, we replace each vector in the source dataset feature library with content from the target feature library. To make the entire timbre transformation process simpler and faster, our matching strategy employs the k-nearest neighbors (kNN) algorithm. The advantage of the kNN algorithm over the commonly used Gaussian clustering algorithm like EM is that it can be used directly without the need for training. Specifically, for the target speaker, we use reference audio from the WavLM model of the target speaker to extract WavLM features and construct a reference matching pool. Similarly, for the original emotion dataset, we also use the WavLM model to extract features and use the kNN algorithm to search for the K nearest SSL features in the reference matching pool. We then average these K features for replacement, thereby transforming the speaker's timbre while retaining prosody and content. This approach fully utilizes features from the target speaker and can potentially eliminate timbre leakage. Since kNN is a lazy learning algorithm, no explicit training is required for the converter during the training process.

Vocoder. The objective of the reconstructor is to take the retrieved and matched audio features and text features from the source dataset as input and transform them into audio linear spectrograms through a vocoder. Traditional vocoders lack explicit pitch modeling, whereas pitch, especially the fundamental frequency (F0), is crucial for expressing emotions [5]. Vocoder models that do not consider pitch features may result in poor quality of emotion dataset after transformation, failing to achieve satisfactory speech synthesis effects. Hence, we employ Hifi-gan [12] as the underlying vocoder. This vocoder takes both the fundamental frequency and acoustic features as input, which significantly enhances the quality of emotional expression during inference.

To improve the reconstruction speed, we fine-tune the target feature library as the training set for the vocoder during the training process. Here, we do not need to worry about overfitting issues. On the contrary, this operation can enhance the reconstructed sound quality.

2.3 Base TTS Framework

Our TTS framework is based on the VITS architecture with enhancements. After inference through the emotion conversion module, as we use the multi-emotion speech data of the target speaker along with corresponding texts as the training set for VITS, to better generate emotional speech for specific speakers, we add emotion and speaker ID embeddings to the decoder of VITS. During training, the emotional labels for the target speaker are specified by the source dataset. In the inference phase, we use a pretrained BERT model for sentiment classification of the text content, specifying the model to choose from the emotional labels in the training dataset we use, thus automating the control of the emotional categories generated by the VITS model. Additionally, in the text encoder part, we also use BERT to enhance the training of semantic features.

3 Experiment

3.1 Datasets

We have chosen the Emotional Speech Database (ESD) [23] as the source emotional dataset for our emotion conversion. All speech data in this dataset is recorded in professional indoor recording studios, with a signal-to-noise ratio of 20 dB or higher, and a sampling frequency of 16 kHz. The ESD database comprises recordings from 10 native English speakers and 10 native Chinese speakers. Each person was assigned 350 sentences to express five emotion categories: neutral, happy, angry, sad, and surprised. Here, we mainly utilize the data from speakers whose native language is English.

For the emotional data of the target speaker, we randomly selected 4 speakers from the VCTK dataset [22]. Each speaker provided 20 sentences of speech as the test samples for the target speaker's emotions. All our speech data has been downsampled to 16 kHz.

3.2 Model Settings

Emotion Conversion Module. For the encoder, we utilize pre-trained WavLM to extract SSL features from both the source emotional dataset and the emotional data of the target speakers. In the k-nearest neighbors (kNN) process, we set k=4 to align with our goal of generating emotional speech with a limited number of target speakers. Additionally, we employ cosine distance to compare features.

Regarding the neural vocoder, we employ Hifigan as the foundational framework and train it using the VCTK dataset with a sampling rate of 16 kHz. During training, we first transform the training data into 1024-dimensional vectors using WavLM. Subsequently, we modify Hifigan to accept these transformed vectors as input. For the training process, we opt for the Adam optimizer with a learning rate of 10^{-4}, conduct training for 200,000 steps, and set the batch size to 16.

TTS Framework. The VITS model is employed as the base, with training parameters consistent with those provided by the official VITS. Pre-trained BERT models include DeBERTa-v3-large[10]for English training and Chinese-RoBERTa-WWM-Ext-Large [7] for Chinese training. The emotion label generation BERT is pretrained by ourself.

3.3 Emotional Dataset Conversion Results

Emotion Preservation Results. t-SNE is a technique for nonlinear dimensionality reduction and visualization of high-dimensional data. It preserves the structure of the data while better reflecting the clustering structure and relationships between categories within the data. In this experimental section, we conducted separate validations on four speakers selected from the VCTK dataset. The specific procedure involves extracting feature vectors from emotion datasets transformed by four different speakers. Subsequently, these feature vectors are passed through a pre-trained emotion classifier to obtain emotion-specific feature vectors. Finally, t-SNE visualization is performed on these feature vectors.

As depicted in Fig. 2, each point represents a speech embedding, with points of the same color indicating the same emotion. The distance between two points represents their similarity; the closer the points, the more similar they are. We observed that in the emotional speech datasets of the four speakers post-conversion, points corresponding to the same emotion tended to cluster together, while those representing different emotions were notably separated. This clearly demonstrates the robustness of our emotion preservation during the voice conversion process. We successfully maintained the emotional content of the original dataset even after voice conversion, highlighting the stability of emotion preservation throughout the process. At the same time, we also observed that "sad" and "normal" emotions are relatively close in distribution. We speculate this could be due to the smaller fluctuation in sad emotion expressions, leading to minimal differences from normal emotions. Alternatively, it might indicate that our training data inadequately captures the nuances of expressing sadness.

Timbre Conversion Results. In this section, we subjectively evaluate the similarity of voice timbre before and after voice conversion. We acknowledge that expressions under different emotions can result in variations in voice perception. Prior to conversion, our speech is predominantly neutral, while after conversion, it incorporates emotionally-inflected speech. Consequently, we decided to provide participants in the evaluation with a three-tiered rating scale to assess the timbre similarity between the two audio samples: "very similar," "not sure," and "totally different." For each speaker, we randomly prepared two audio samples, one neutral and the other randomly generated from various emotions such as happiness, sadness, anger, and surprise. The samples were randomly grouped and played to mitigate the influence of order effects. Participants were asked to assess the timbre similarity of each audio sample group within a specified time frame. The results, as depicted in the Table 1, indicate that the majority of participants

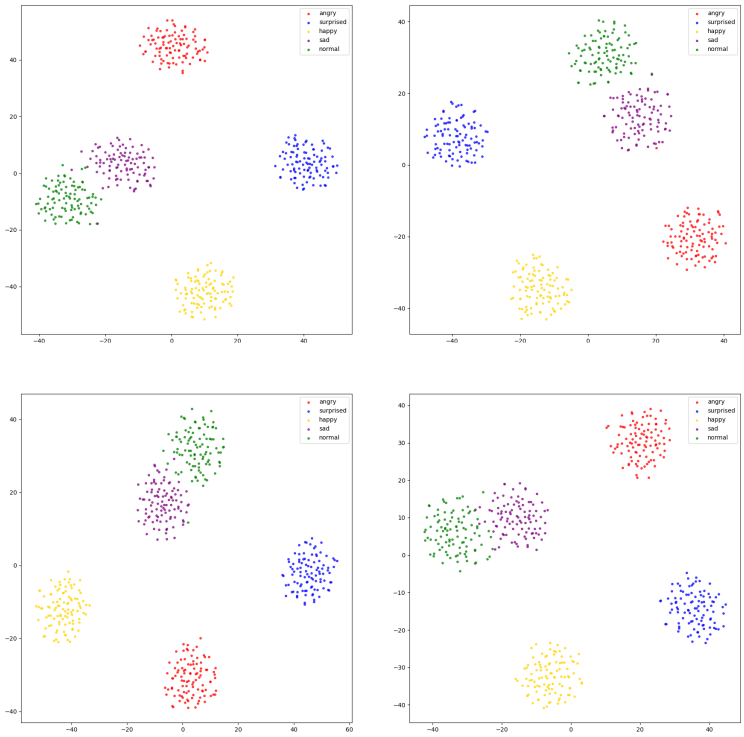


Fig. 2. Visualization of speaker embedding.

consider the audio samples before and after conversion for all four speakers to be relatively similar in timbre. Based on these results, we can reasonably conclude that our voice conversion process is successful and universally effective.

Table 1. Similarity test.

speaker id	speaker 1	speaker 2	speaker 3	speaker 4
totally different	4	2	3	4
not sure	5	3	4	3
verysimilar	91	95	93	93

3.4 The Performance of Emotional Speech Generation

Subjective Evaluation. In this section, we conducted Mean Opinion Score (MOS) testing to evaluate the naturalness of generated speech and the clarity of speaker emotion expression. We employed different methods for generating emotional speech, categorizing them into five emotions, with five speech samples

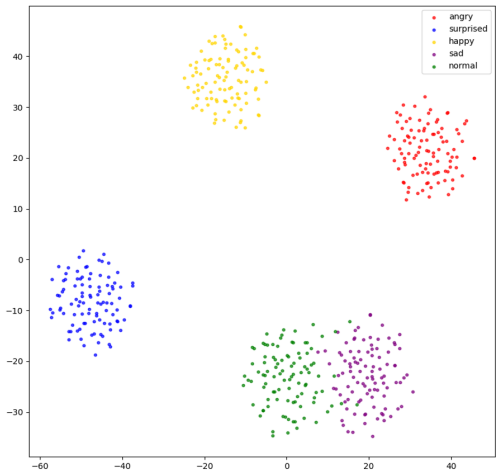


Fig. 3. Emotional discrimination test.

generated for each emotion. Specifically, we utilized speech samples generated by OpenVoice[15], Cross-Speaker Emotion Transfer TTS(CSET TTS)[20], and PromptStyle[14] for comparative analysis. In the scoring phase, we recruited a total of 15 native Mandarin Chinese speakers to participate in the rating experiment. As per the requirements, they were tasked with scoring the samples generated by different methods based on a scale from 1 to 5 for both the naturalness of generated speech and the clarity of speaker emotion expression within the same time frame. The results, as shown in Table 2, indicate that our emotion expression performance is generally on par with other methods.

Objective Evaluation. Firstly, we conducted accuracy testing on the generated speech to evaluate whether the quality of emotion datasets generated by the emotion conversion module would affect the accuracy of speech synthesis itself. Here, we utilized the VTCK original dataset and, under the condition of the same speaker, selected an equal amount of speech as the training set to train the original VITS model. As a control group, each of the two generating models produced 100 audio sentences. Using Whisper [16] provided by OpenAI, we

Table 2. Naturalness and Emo-expression MOS results

Models	Naturalness	Emo-expression
OpenVoice	4.23±0.08	4.15±0.09
CSET TTS	3.88±0.12	3.73±0.10
PromptStyle	4.14±0.06	4.09±0.09
Our method	4.21±0.11	4.17±0.08

Table 3. VITS and LRVESpeech WER results

Models		**WER(%)**
VITS-vctk		8.32
LRVESpeech	normal	7.35
	happy	9.61
	angry	11.16
	sad	8.75
	surprised	9.29

conducted ASR testing to calculate the Word Error Rate (WER). The results, as shown in Table 3, indicate that the Character Error Rate (CER) of LRVESpeech did not increase significantly compared to the control group. Generally, low-quality datasets can affect the performance of VITS. This strongly suggests that our model, after improvements such as Vocoder pre-training, achieves the goal of generating multi-emotion speech while maintaining a trustworthy and acceptable speech synthesis quality. Additionally, the WER of the Normal group was even lower than that of the baseline model, which may be attributed to the incorporation of BERT semantic feature synthesis.

Secondly, we conducted t-SNE testing on the generated speech to evaluate whether the samples could effectively express and differentiate between different emotions. Due to the tedious process of generating speech with varied emotions and content, we chose to test t-SNE using samples from a single speaker for simplicity. As shown in Fig. 3, we can clearly observe distinct separation of samples from different emotion categories in the t-SNE space, indicating that our generated speech samples possess good discriminative ability in emotional expression.

4 Conclusion

In this work, we propose LRVESpeech, a multi-emotional speech synthesis model that synthesizes four different emotional types of speech from a limited number of neutral samples under low-resource conditions. Additionally, we validated the possibility of acquiring a substantial multi-emotional dataset through voice conversion using neutral samples, offering a novel approach to the problem of generating multi-emotional speech when there is a shortage of emotional samples. The conclusions drawn from our experiments indicate that the emotional speech generated by LRVESpeech is clear and effective, achieving the synthesis of multi-emotional speech with a limited number of samples.

References

1. Baas, M., van Niekerk, B., Kamper, H.: Voice conversion with just nearest neighbors. In: Harte, N., Carson-Berndsen, J., Jones, G. (eds.)

24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023. pp. 2053–2057. ISCA (2023). https://doi.org/10.21437/INTERSPEECH.2023-419, https://doi.org/10.21437/Interspeech.2023-419
2. Barakat, H., Türk, O., Demiroglu, C.: Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources. EURASIP J. Audio Speech Music. Process. **2024**(1), 11 (2024). https://doi.org/10.1186/S13636-024-00329-7, https://doi.org/10.1186/s13636-024-00329-7
3. Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., Wei, F.: Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE J. Sel. Top. Signal Process. **16**(6), 1505–1518 (2022). https://doi.org/10.1109/JSTSP.2022.3188113, https://doi.org/10.1109/JSTSP.2022.3188113
4. Choi, H., Park, S., Park, J., Hahn, M.: Multi-speaker emotional acoustic modeling for cnn-based speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019. pp. 6950–6954. IEEE (2019). https://doi.org/10.1109/ICASSP.2019.8683682, https://doi.org/10.1109/ICASSP.2019.8683682
5. Chou, J., Lee, H.: One-shot voice conversion by separating speaker and content representations with instance normalization. In: Kubin, G., Kacic, Z. (eds.) 20th Annual Conference of the International Speech Communication Association, Interspeech 2019, Graz, Austria, September 15-19, 2019. pp. 664–668. ISCA (2019). https://doi.org/10.21437/INTERSPEECH.2019-2663, https://doi.org/10.21437/Interspeech.2019-2663
6. Cornille, T., Wang, F., Bekker, J.: Interactive multi-level prosody control for expressive speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022. pp. 8312–8316. IEEE (2022). https://doi.org/10.1109/ICASSP43922.2022.9746654, https://doi.org/10.1109/ICASSP43922.2022.9746654
7. Cui, Y., Che, W., Liu, T., Qin, B., Yang, Z.: Pre-training with whole word masking for chinese BERT. IEEE ACM Trans. Audio Speech Lang. Process. **29**, 3504–3514 (2021). https://doi.org/10.1109/TASLP.2021.3124365, https://doi.org/10.1109/TASLP.2021.3124365
8. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/V1/N19-1423, https://doi.org/10.18653/v1/n19-1423
9. Fix, E.: Discriminatory analysis: nonparametric discrimination, consistency properties, vol. 1. USAF school of Aviation Medicine (1985)
10. He, P., Liu, X., Gao, J., Chen, W.: Deberta: decoding-enhanced bert with disentangled attention. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=XPZIaotutsD
11. Kim, J., Kong, J., Son, J.: Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24

July 2021, Virtual Event. Proceedings of Machine Learning Research, vol. 139, pp. 5530–5540. PMLR (2021), http://proceedings.mlr.press/v139/kim21f.html
12. Kong, J., Kim, J., Bae, J.: Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html
13. Laurer, M., van Atteveldt, W., Casas, A., Welbers, K.: Building efficient universal classifiers with natural language inference. CoRR **abs/2312.17543** (2023). https://doi.org/10.48550/ARXIV.2312.17543, https://doi.org/10.48550/arXiv.2312.17543
14. Liu, G., Zhang, Y., Lei, Y., Chen, Y., Wang, R., Xie, L., Li, Z.: Promptstyle: Controllable style transfer for text-to-speech with natural language descriptions. In: Harte, N., Carson-Berndsen, J., Jones, G. (eds.) 24th Annual Conference of the International Speech Communication Association, Interspeech 2023, Dublin, Ireland, August 20-24, 2023. pp. 4888–4892. ISCA (2023). https://doi.org/10.21437/INTERSPEECH.2023-1779, https://doi.org/10.21437/Interspeech.2023-1779
15. Qin, Z., Zhao, W., Yu, X., Sun, X.: Openvoice: Versatile instant voice cloning. CoRR **abs/2312.01479** (2023). https://doi.org/10.48550/ARXIV.2312.01479, https://doi.org/10.48550/arXiv.2312.01479
16. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA. Proceedings of Machine Learning Research, vol. 202, pp. 28492–28518. PMLR (2023), https://proceedings.mlr.press/v202/radford23a.html
17. Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., Liu, T.: Fastspeech 2: Fast and high-quality end-to-end text to speech. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=piLPYqxtWuA
18. Sha, B., Li, X., Wu, Z., Shan, Y., Meng, H.: Neural concatenative singing voice conversion: Rethinking concatenation-based approach for one-shot singing voice conversion. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024. pp. 12577–12581. IEEE (2024). https://doi.org/10.1109/ICASSP48485.2024.10446066, https://doi.org/10.1109/ICASSP48485.2024.10446066
19. Tan, X., Qin, T., Soong, F.K., Liu, T.: A survey on neural speech synthesis. CoRR **abs/2106.15561** (2021), https://arxiv.org/abs/2106.15561
20. Terashima, R., Yamamoto, R., Song, E., Shirahata, Y., Yoon, H., Kim, J., Tachibana, K.: Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation. In: Ko, H., Hansen, J.H.L. (eds.) 23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022. pp. 3018–3022. ISCA (2022). https://doi.org/10.21437/INTERSPEECH.2022-11278, https://doi.org/10.21437/Interspeech.2022-11278
21. Wang, Y., Skerry-Ryan, R.J., Stanton, D., Wu, Y., Weiss, R.J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q.V., Agiomyrgiannakis, Y., Clark,

R., Saurous, R.A.: Tacotron: Towards end-to-end speech synthesis. In: Lacerda, F. (ed.) 18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017. pp. 4006–4010. ISCA (2017). https://doi.org/10.21437/INTERSPEECH.2017-1452, https://doi.org/10.21437/Interspeech.2017-1452

22. Yamagishi, J., Veaux, C., MacDonald, K.: Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). https://doi.org/10.7488/ds/2645 (2019). https://doi.org/10.7488/ds/2645, [sound]
23. Zhou, K., Sisman, B., Liu, R., Li, H.: Emotional voice conversion: Theory, databases and ESD. Speech Commun. **137**, 1–18 (2022). https://doi.org/10.1016/J.SPECOM.2021.11.006, https://doi.org/10.1016/j.specom.2021.11.006