Towards Emotionally Consistent Text-Based Speech Editing: Introducing EmoCorrector and The ECD-TSE Dataset

Rui Liu[1], *Pu Gao*[1], *Jiatian Xi*[1], *Berrak Sisman*[2], *Carlos Busso*[3], *Haizhou Li*[4,5]
[1]Inner Mongolia University, Hohhot, China
[2]Center for Language and Speech Processing, Johns Hopkins University, USA
[3]LTI, Carnegie Mellon University, USA
[4]SRIBD, School of Data Science, The Chinese University of Hong Kong, Shenzhen, China
[5]Department of ECE, National University of Singapore, Singapore
liurui_imu@163.com

Abstract

Text-based speech editing (TSE) modifies speech using only text, eliminating re-recording. However, existing TSE methods, mainly focus on the content accuracy and acoustic consistency of synthetic speech segments, and often overlook the emotional shifts or inconsistency issues introduced by text changes. To address this issue, we propose EmoCorrector, a novel post-correction scheme for TSE. EmoCorrector leverages Retrieval-Augmented Generation (RAG) by extracting the edited text's emotional features, retrieving speech samples with matching emotions, and synthesizing speech that aligns with the desired emotion while preserving the speaker's identity and quality. To support the training and evaluation of emotional consistency modeling in TSE, we pioneer the benchmarking Emotion Correction Dataset for TSE (ECD-TSE). The prominent aspect of ECD-TSE is its inclusion of <text, speech> paired data featuring diverse text variations and a range of emotional expressions. Subjective and objective experiments and comprehensive analysis on ECD-TSE confirm that EmoCorrector significantly enhances the expression of intended emotion while addressing emotion inconsistency limitations in current TSE methods. Code and audio examples are available at https://github.com/AI-S2-Lab/EmoCorrector.

Index Terms: Text-based Speech Editing, Emotional Consistency, Retrieval-Augmented Generation (RAG), Emotion Post-Correction

1. Introduction

Text-based Speech Editing (TSE) modifies audio by editing its underlying text rather than the audio signal directly. With the rise of digital media, TSE has become essential for applications like social media content creation , game voiceovers , and film dubbing , as it corrects issues such as mispronunciations, omissions, or stuttering without requiring a full re-recording [1].

Recent advancements in text-to-speech (TTS) have led to the development of neural models for TSE [1, 2, 3, 4, 5]. For example, Despite these advancements, most TSE methods focus on content modification and acoustic quality, often neglecting emotional consistency [6]. As shown in Fig. 1, although there is only one word difference between the edited text and the original text, the emotional expression at the sentence level is very different. Although the traditional TSE model successfully synthesizes the speech segment for the word "bad", it does not express the emotion that the edited text should contain. At this point, the emotional state of the whole sentence needs to be further corrected to achieve emotional consistency.

A natural idea is to use *Emotional Voice Conversion (EVC)* [7, 8] models to transform the emotion of an entire sentence into the same emotional state that the edited text contains without altering the speaker's identity. However, directly using an EVC model presents several issues. 1) Before applying EVC, it is necessary to identify the emotional state embedded in the edited text, a process that may introduce errors. 2) Even if the identified emotion state is accurate, using discrete emotion labels as input for the EVC model may fail to capture rich and nuanced emotional information, thereby limiting the model's performance. 3) The output of the EVC model is inherently uncontrollable, making it difficult to evaluate whether the converted results truly achieve optimal performance. Therefore, directly relying on an EVC model for post-correction does not fully meet our requirements.

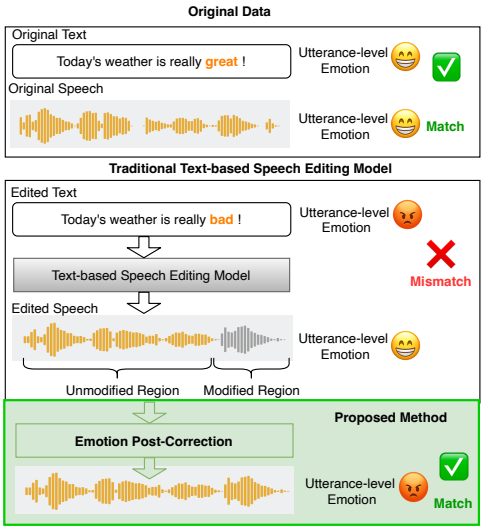

Figure 1: *Our approach lies in correcting the emotional mismatch or inconsistency issue of traditional TSE methods.*

To address the above issues, we propose an end-to-end post-correction approach to rectify emotional inconsistencies for TSE, termed EmoCorrector. Our EmoCorrector is inspired by the concept of Retrieval-Augmented Generation (RAG) [9]. First, emotional embedding are extracted from the edited text, and speech samples with similar emotional embedding are retrieved from the cross-modal retrieval database. Next, the emotional embeddings of these retrieved speech samples are used as a reference, while the speaker features from the original edited speech are retained to preserve speaker identity. Finally, TTS synthesis is performed, enabling the modification of the original edited speech's emotional information and achieving consistent emotional expression. For the feature construction of the cross-modal database, we adopt a contrastive learning [10] strategy focused on emotional features to perform language-audio contrastive learning, aligning the emotional features of

Table 1: *Examples of text data. There are only one or two words different between different texts, but their emotional expressions are completely different.*

Emotion	Text
Neutral	My sister studies in the library.
Sadness	My sister **cries** in the library.
Angry	My sister **yells** in the library.
Happy	My sister studies in the **joyful** library.
Fear	My sister studies in the **haunted** library.

text and speech within a unified space. Additionally, we design a speaker-emotion feature disentanglement strategy [11] based on adversarial learning [12] to ensure the independence of emotional and speaker feature representation.

To support training and evaluation for emotional post-correction, we construct a new Emotion Correction Dataset for TSE (ECD-TSE). This dataset includes original text as well as edited text reflecting different emotional states. We utilized three advanced TTS models to perform emotional speech synthesis for all the text. This emotion-rich speech database, containing text editing information, provides a solid foundation for modeling emotional consistency. Extensive experiments on ECD-TSE demonstrate that EmoCorrector significantly enhances emotional alignment while maintaining natural audio quality. The main contributions of this paper can be summarized as follows: 1) We propose a comprehensive end-to-end emotion post-correction scheme for TSE, named EmoCorrector. 2) Our EmoCorrector draws on the techniques of RAG and designs cross-modal emotion contrastive learning and speaker-emotion disentanglement learning strategies. 3) We introduce the groundbreaking ECD-TSE dataset, which advances emotional consistency modeling and evaluation in the field of TSE.

2. Dataset: ECD-TSE

For the emotionally consistent modeling of TSE, we need a database where sentences conveying similar lexical information with few modifications elicit different emotions. Unfortunately, popular databases such as the ESD dataset [13], MSP-PODCAST [14], and IEMOCAP [15] are not appropriate for this task since they are usually a fixed text corresponding to the speech with various emotions. To this end, we develop the ECD-TSE corpus through two crucial steps: *Text variant generation* and *Emotional speech generation*.

Text Variant Generation: The process of text variant generation is to generate neutral text first, and then generate text variants with different emotions.

First, we instruct ChatGPT-4 to generate texts with a neutral emotion. Our prompt for this step is as follows: "Please help me to build a text dataset, with each sentence consisting of 4-10 words, and in English. Please write 10 sentences, and ensure the text contains words that express neutral emotions." Subsequently, we instruct ChatGPT-4 to modify the neutral texts by altering one or two words, thereby inducing a change in the text's emotional connotation. Our prompt for this modification is as follows (The following example takes "sadness" as an example): "Now, I will give you a sentence. Please modify only one or two words to change the emotion to sadness. Please output only one modified sentence."

As shown in Table 1, in this manner, we obtain texts corresponding to four additional emotional states, with only one or two words differing between the various emotions.

Emotional Speech Generation: For emotional speech generation, we utilize three advanced expressive TTS systems, that are Microsoft Azure [16], CosyVoice2 [17], and F5-TTS [18], as the synthesizer. For Microsoft Azure, when we synthesize each sentence of text, we directly set the emotion label corresponding to the text in the engine, randomly set the speaker identity, and it can synthesize the highly expressive speech of a specific speaker corresponding to the emotion. For Cosyvoice2 and F5-TTS, a reference is required for emotional speech synthesis. Therefore, we randomly select speech samples with matching emotional labels from third-party emotional datasets as references based on the emotional labels of the text to be synthesized. Note that the speaker identity of the synthesized speech is consistent with the reference speech. For the third-party emotional dataset, we use the ESD [13] and MEAD [19] datasets, both of which provide diverse emotional speech samples for diverse speaker identities. Ultimately, the Azure system synthesizes speech for 5 speakers, CosyVoice2 synthesizes speech for another 5 speakers, and F5-TTS synthesizes speech for an additional 2 speakers. The statistic information is shown in Table 2.

Table 2: *Statistics information of the ECD-TSE.*

Attribute	Value
Speech Samples	84,000 = 1,400 * 5 * 12
Emotions	5 (Happy, Sadness, Neutral, Angry, Fear)
Speakers	12 (6 male and 6 female)
Text	7,000 = 1,400 * 5
Total Duration	Approximately 90 hours
Sampling Rate	16,000 Hz

3. Methodology: EmoCorrector

Our proposed EmoCorrector framework ensures emotional consistency in TSE while preserving speaker identity through three stages: 1) Text-Speech Emotion Retrieval Database Construction (Block 1 in Fig.2), 2) Speaker-Emotion Disentanglement Pre-training (Block 2 in Fig.2), and 3) Emotion Post-Correction for TSE (Block 3 in Fig.2). The first two modules require pre-training, and then the third module is trained end-to-end.

3.1. Text-Speech Emotion Retrieve Database Construction

Text-Speech Emotion Contrastive Pre-training: We propose Emotional Contrastive Language-Audio Pretraining (EmoCLAP), inspired by CLAP [20], to learn a shared semantic space for text and speech emotion representations. As shown in Fig. 2 (Block 1.1), the EmoCLAP Text Encoder (based on pre-trained RoBERTa [21]) extracts text emotional features T, while the Speech Encoder (built on emotion2vec [22]) extracts emotional speech features S. Both are projected into a common embedding space via a multi-layer perceptron (MLP): $\mathcal{M}_s = \text{MLP}(S), \quad \mathcal{M}_t = \text{MLP}(T)$.

To enhance text-speech alignment, we employ contrastive learning to maximize the similarity for positive text-speech pairs while minimizing it for negatives. The dot product similarity matrix B is computed as: $B = \mu(\mathcal{M}_s \cdot \mathcal{M}_t^\top)$, where μ is a scaling factor. The contrastive loss is then defined as: $\mathcal{L}_{clap} = -\log \frac{\exp(B)}{\sum_k \exp(B)}$.

Text-Speech Emotion Feature Preparation: After contrastive pre-training, the frozen EmoCLAP Text and Speech Encoders extract fixed emotion embeddings from text and speech (Fig. 2, Block 1.2). These embeddings form the retrieval database, supporting both text-speech cross-modal emotion retrieval and emotion representation during speech synthesis.

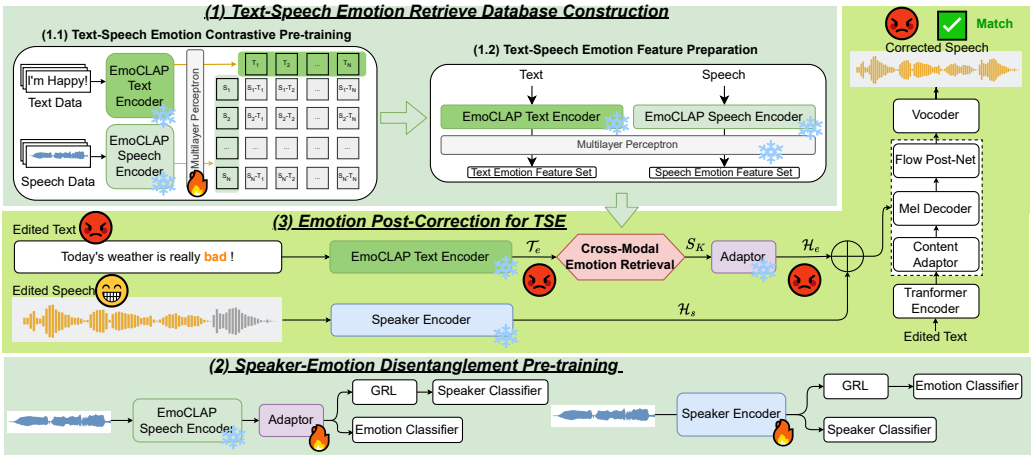

Figure 2: *The overall workflow of EmoCorrector.*

3.2. Speaker-Emotion Disentanglement Pre-training

To disentangle the emotional expression from the speaker's identity and thus serve the subsequent emotion post-correction., we employ adversarial training with a gradient reversal layer (GRL) as the backbone (as shown in Fig. 2 (Block 2)). The EmoCLAP Speech Encoder extracts an emotion representation $\mathcal{H}_e$, which is classified into one of five emotional states using a softmax-based classifier: $P_{\mathcal{H}_e} = C_e(\mathcal{H}_e)$, with the corresponding cross-entropy loss: $\mathcal{L}_e = \mathrm{CE}(P_{\mathcal{H}_e}, \text{emotion\_id})$.

Since $\mathcal{H}_e$ may also encode speaker-specific information, we pass it through a GRL before feeding it into a speaker classifier C_s. This yields: $P^s_{\mathcal{H}_e} = C_s(\mathrm{GRL}(\mathcal{H}_e))$. The adversarial optimization is enforced through an additional cross-entropy loss: $\mathcal{L}^{(e)}_s = \mathrm{CE}(P^s_{\mathcal{H}_e}, \text{speaker\_id})$, ensuring that $\mathcal{H}_e$ retains only emotion-related information while removing speaker-dependent attributes.

In parallel, the Speaker Encoder extracts a speaker representation $\mathcal{H}_s$, which is used for speaker classification with loss $\mathcal{L}_s$. To ensure that $\mathcal{H}_s$ does not capture emotion-related cues, it is also processed through a GRL prior to an emotion classifier, incurring an additional adversarial loss $\mathcal{L}^{(s)}_e$. The overall loss function is the sum of these losses: $\mathcal{L} = \mathcal{L}_s + \mathcal{L}_e + \mathcal{L}^{(e)}_s + \mathcal{L}^{(s)}_e$.

3.3. Emotion Post-Correction for TSE

As illustrated in Fig. 2 (Block 3), the purpose of emotional post-correction is to perform emotional correction on edited speech with inconsistent emotional expressions based on the emotion of the edited text, ensuring that the final speech aligns with the emotional expression of the edited text.

Given an edited text X_t, the EmoCLAP Text Encoder first extracts the text emotion embedding $\mathcal{T}_e$. This embedding is then used as a query vector in the *cross-modal emotion retrieval* module, computed by cosine similarity [11] to retrieve the Top K speech emotion embeddings $S_K = \{E_{i_1}, E_{i_2}, ..., E_{i_K}\}$ most similar to it. Furthermore, S_K are aggregated by averaging them and processed by a pre-trained Adaptor from Block 2 to generate the refined speech emotion embedding $\mathcal{H}_e$.

Simultaneously, the edited speech X_s is processed by the Speaker Encoder to yield the speaker embedding $\mathcal{H}_s$. After that, the $\mathcal{H}_e$ and $\mathcal{H}_s$ are directly added to form a joint embedding, which is then injected to the *Content Adaptor*, *Mel Decoder* and *Flow Post-Net* to integrate emotional characteristics into the synthesized speech. The *Vocoder* synthesizes the final waveform for the given edited text. More details about the TTS synthesizer are referred to [23]. The loss function of Block 3 is also consistent with [23].

4. Experiments and Results

4.1. Experimental Setup

We evaluate EmoCorrector on the ECD-TSE. Precise forced alignment was performed using the Montreal Forced Aligner (MFA) [24]. The dataset was randomly partitioned into training, validation, and test sets in the proportions of 98%, 1%, and 1%, respectively. The EmoCLAP Text Encoder extracts a 768-dimensional emotion embedding, while the Speech Encoder extracts an utterance-level 1024-dimensional emotion embedding. These two embeddings are projected into a multimodal space using two learnable projection matrices, resulting in an output dimension of 1024. The configuration of the Speaker Encoder follows the reference encoder from GST [25], comprising a convolutional stack and a GRU block. Each convolutional layer employs a 3×3 kernel with a stride of 2×2, followed by batch normalization and ReLU activation. The output channels for the six convolutional layers are $32, 32, 64, 64, 128$, and 128, respectively. All classifiers in this architecture consist of a fully connected (linear) layer followed by a softmax layer, and are trained using cross-entropy loss. The pre-trained HiFiGAN [26] vocoder is used to synthesize the speech waveform. An Adam optimizer is used with a learning rate of 1×10^{-7}. We set K to 5. We trained EmoCLAP for 200,000 steps with μ set to 10 steps, followed by training the speaker-emotion disentanglement model for 300,000 steps. All training was conducted on a single A800 GPU.

4.2. Evaluation Metrics

Subjective Metric: Text-Speech Emotion Matching Mean Opinion Score (TSE-MOS) allows listeners to determine whether the emotion of the speech matches the emotion conveyed by the corresponding text. **Objective Metrics:** 1) **Text-Speech Emotion Matching Accuracy (TSEAcc)** is a new LLM-assisted metric. We utilize an advanced Speech Understanding LLM, Qwen2-Audio [27] as the speech emotion recognizer to recognize the emotion in speech and then compare it to the emotion label contained in the text to calculate the accuracy. Due to Qwen2's output characteristics, we calculate accuracy group-wise and then average the results. 2) **Emotional Cosine Similarity (ECS)** adopts emotion2vec [22] to extract the emotional features of the audio before and after correction

Table 3: *Emotional comparison analysis before and after the emotion post-correction of the TSE models.*

Method	TSE-MOS		TSEAcc (%)		ECS	
	Before	After	Before	After	Before	After
Ground Truth	NA	4.67 ± 0.04	NA	52.1%	NA	1.00
Editspeech	3.31 ± 0.03	$\mathbf{4.04 \pm 0.01}$	8.3%	**47.5%**	0.79	**0.97**
A^3T	3.48 ± 0.01	$\mathbf{4.13 \pm 0.03}$	6.8%	**49.3%**	0.73	**0.98**
FluentSpeech	3.17 ± 0.01	$\mathbf{3.93 \pm 0.02}$	7.1%	**46.4%**	0.77	**0.98**
VoiceCraft	3.24 ± 0.02	$\mathbf{4.12 \pm 0.02}$	6.9%	**48.8%**	0.71	**0.97**

Table 4: *A comparative experiment with different values of K.*

Method	TSEAcc (%)		
	K=3	K=5	K=10
Ground Truth	52.1%	52.1%	52.1%
Editspeech	**48.3%**	47.5%	42.5%
A^3T	47.5%	**49.3%**	43.6%
FluentSpeech	46.2%	**46.4%**	41.2%
VoiceCraft	48.7%	**48.8%**	44.3%

Table 5: *Comparison of similarity for overall speech quality before and after correction.*

Method	Energy	MFCC
Editspeech	0.91	0.96
A^3T	0.91	0.94
FluentSpeech	0.89	0.94
VoiceCraft	0.93	0.94

and compute the cosine similarity [11] between these features and the Ground Truth emotional features. The similarity score is in the range of [-1, 1], where a larger value indicates a higher similarity of input samples.

4.3. Baselines

We conduct a comparative study of the speech before and after emotion correction using four advanced TSE models, including: 1) **EditSpeech** [28] generates text based on the surrounding context to maintain naturalness and quality; 2) **A^{3}T** [2] proposes alignment-aware acoustic-text pre-training that takes both phonemes and partially-masked spectrograms as inputs; 3) **FluentSpeech** [1] uses a diffusion model as the backbone and predicts the masked feature with the help of context speech; and 4) **VoiceCraft** [5] introduces a regression Transformer decoder architecture for neural codec token filling. We also include **Ground Truth** speech for comparison.

4.4. Main Results

All the pre-training is trained to the convergent state. In addition, without pre-training, our internal experiments have proved that emotion correction cannot achieve satisfactory results. Due to space constraints, the results of this part of the internal experiments are not displayed. Below we will mainly experiment on the effect of emotion correction and report the results.

In the TSE-MOS evaluation, 25 listeners rated 50 audio samples for the consistency between text and speech emotions. As shown in Table 3, EmoCorrector significantly improved the emotional content of the model, with the TSE-MOS score increasing by an average of 0.73 points. Emotion correction for EditSpeech, A^3T, FluentSpeech, and VoiceCraft resulted in increases of 0.73, 0.65, 0.76, and 0.88 points, respectively. In the TSEAcc evaluation, EmoCorrector also enhanced the accuracy of text and speech emotion matching, with an average improvement from 7.2% to 48%. In the ECS evaluation, EmoCorrector made the emotion-corrected speech more similar to the Ground Truth. The emotion correction for EditSpeech, A^3T, FluentSpeech, and VoiceCraft improved by 0.18, 0.25, 0.21, and 0.26, respectively. These results demonstrate that EmoCorrector effectively addresses the issue of emotional inconsistency between edited speech and text.

4.5. Further Analysis

Analysis of parameter K: In cross-modal emotion retrieval, the number of retrieved Top K samples may have an impact on the final emotion expression effect. We set K to $\{3,5,10\}$ to analyze the speech generation results, since 1 may result in insufficient information and a selection between 5 and 10 might not demonstrate a noticeable variance from 5. All the remaining experimental configurations remain the same as before. As shown in Table 4, the TSEAcc metric achieved the best results with K=5, while performing poorly with 3 and 10. Among these, K=10 yielded the worst outcome, possibly due to the risk of information redundancy when retrieving too many samples.

Analysis of the impact of overall quality: Although our EmoCorrector focuses on emotion correction, it is not intended to compromise the speech quality. To validate this, we randomly select 50 test samples. For each sample, we calculate the Energy and MFCC features before and after correction and then obtain their cosine similarity. The results in Table 5 show that the similarity values of Energy and MFCC remain around 1, indicating that the model maintains the overall speech quality after emotional correction.

Comparison of EVC systems: This section validates whether our EmoCorrector has an advantage over using the EVC model directly. We randomly select 50 test samples and adopt two advanced EVC models, that are Cycle-GAN [29] and VAW-GAN [30], to conduct emotion conversion for the edited speech. We compare the EVC-converted samples with the emotion-corrected samples, and the experimental results in terms of TSEAcc are shown in Table 6 (Due to limited space, we posted the results on the demo website.). It can be seen from the results that our post-processing method can obtain more accurate synthetic speech of emotions than the EVC method, which has an advantage in emotional rendering.

5. Conclusion

This paper presents a novel emotion post-correction scheme for the TSE task, introducing the new benchmarking ECD-TSE dataset and the EmoCorrector that consists of a three-stage pipeline: Text-Speech Emotion Retrieve Database Construction, Speaker-Emotion Disentanglement Pre-training and Emotion Post-Correction for TSE. Experimental results demonstrate that the proposed framework improves emotional consistency in the edited speech. To the best of our knowledge, EmoCorrector and ECD-TSE are the first methods specifically designed for this task. We hope this work provides a foundation for future research in emotion correction for text-based speech editing.

6. Acknowledgement

This work was funded by the Young Scientists Fund (No. 62206136), the General Program (No. 62476146) of the National Natural Science Foundation of China, and the Young Elite Scientists Sponsorship Program by CAST (2024QNRC001). The work by Haizhou Li was supported by the Shenzhen Science and Technology Program (Shenzhen Key Laboratory, Grant No. ZDSYS20230626091302006), the Shenzhen Science and Technology Research Fund (Fundamental Research Key Project, Grant No. JCYJ20220818103001002), and the Program for Guangdong Introducing Innovative and Enterpreneurial Teams, Grant No. 2023ZT10X044.

7. References

[1] Z. Jiang, Q. Yang, J. Zuo, Z. Ye, R. Huang, Y. Ren, and Z. Zhao, "Fluentspeech: Stutter-oriented automatic speech editing with context-aware diffusion models," in *Findings of the Association for Computational Linguistics: ACL 2023*, 2023, pp. 11 655–11 671.

[2] H. Bai, R. Zheng, J. Chen, M. Ma, X. Li, and L. Huang, "A ̈3 t: Alignment-aware acoustic and text pretraining for speech synthesis and editing," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1399–1411.

[3] R. Liu, J. Xi, Z. Jiang, and H. Li, "Fluenteditor: Text-based speech editing by considering acoustic and prosody consistency," *arXiv preprint arXiv:2309.11725*, 2023.

[4] X. Wang, M. Thakker, Z. Chen, N. Kanda, S. E. Eskimez, S. Chen, M. Tang, S. Liu, J. Li, and T. Yoshioka, "Speechx: Neural codec language model as a versatile speech transformer," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[5] P. Peng, P. Huang, S. Li, A. Mohamed, and D. Harwath, "Voicecraft: Zero-shot speech editing and text-to-speech in the wild," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics , ACL 2024, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds. Association for Computational Linguistics, 2024, pp. 12 442–12 462.

[6] Y. Wei, S. Yuan, R. Yang, L. Shen, Z. Li, L. Wang, and M. Chen, "Tackling modality heterogeneity with multi-view calibration network for multimodal sentiment detection," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 5240–5252.

[7] K. Zhou, B. Sisman, R. Liu, and H. Li, "Emotional voice conversion: Theory, databases and esd," *Speech Communication*, vol. 137, pp. 1–18, 2022.

[8] Z. Luo, S. Lin, R. Liu, J. Baba, Y. Yoshikawa, and H. Ishiguro, "Decoupling speaker-independent emotions for voice conversion via source-filter networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 11–24, 2023.

[9] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020.

[10] Y. Meng, X. Li, Z. Wu, T. Li, Z. Sun, X. Xiao, C. Sun, H. Zhan, and H. Meng, "Calm: Contrastive cross-modal speaking style modeling for expressive text-to-speech synthesis," *arXiv preprint arXiv:2308.16021*, 2023.

[11] T. Li, X. Wang, Q. Xie, Z. Wang, and L. Xie, "Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 1448–1460, 2022.

[12] W.-N. Hsu, Y. Zhang, R. J. Weiss, Y.-A. Chung, Y. Wang, Y. Wu, and J. Glass, "Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5901–5905.

[13] K. Zhou, B. Sisman, R. Liu, and H. Li, "Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 920–924.

[14] J. Duret, M. Rouvier, and Y. Estève, "Msp-podcast ser challenge 2024: L'antenne du ventoux multimodal self-supervised learning for speech emotion recognition," *arXiv preprint arXiv:2407.05746*, 2024.

[15] 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, pp. 335–359, 2008.

[16] Microsoft, "Azure speech studio," https://azure.microsoft.com/en-us/services/cognitive-services/speech-services/.

[17] Z. Du, Y. Wang, Q. Chen, X. Shi, X. Lv, T. Zhao, Z. Gao, Y. Yang, C. Gao, H. Wang *et al.*, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[18] Y. Chen, Z. Niu, Z. Ma, K. Deng, C. Wang, J. Zhao, K. Yu, and X. Chen, "F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching," *arXiv preprint arXiv:2410.06885*, 2024.

[19] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, "Mead: A large-scale audio-visual dataset for emotional talking-face generation," in *European Conference on Computer Vision*. Springer, 2020, pp. 700–717.

[20] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] Y. Liu, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, vol. 364, 2019.

[22] Z. Ma, Z. Zheng, J. Ye, J. Li, Z. Gao, S. Zhang, and X. Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation." Association for Computational Linguistics, 2024, pp. 15 747–15 760.

[23] R. Huang, Y. Ren, J. Liu, C. Cui, and Z. Zhao, "Generspeech: Towards style transfer for generalizable out-of-domain text-to-speech," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 970–10 983, 2022.

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

[25] 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.

[26] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[27] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[28] D. Tan, L. Deng, Y. T. Yeung, X. Jiang, X. Chen, and T. Lee, "Editspeech: A text based speech editing system using partial inference and bidirectional fusion," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 626–633.

[29] K. Zhou, B. Sisman, and H. Li, "Transforming spectrum and prosody for emotional voice conversion with non-parallel training data," *arXiv preprint arXiv:2002.00198*, 2020.

[30] K. Zhou, B. Sisman, M. Zhang, and H. Li, "Converting anyone's emotion: Towards speaker-independent emotional voice conversion," in *Interspeech 2020*, 2020, pp. 3416–3420.