FS-CLAP: FUSION-SEPARATION CLAP FOR EMOTIONAL SPEAKING STYLE SPEECH RETRIEVAL

Name of author

Address - Line 1
Address - Line 2
Address - Line 3

ABSTRACT

General contrastive cross-modal pre-trained models perform well on coarse-grained downstream tasks. However, they do not generalize effectively to complex, fine-grained tasks. One example is the emotional speaking style description task. Therefore, fine-tuning on datasets for this task is essential. Simply aligning the coarse-grained features from the CLAP model is not enough. This often leads to insufficient alignment or representation collapse. To address this, we propose FS-CLAP. It uses a fuse-separate strategy to align fine-grained features from both modalities. Meanwhile, we also utilize a loss function based on hard negative samples to enhance the alignment effect of hard negative samples. Experiments on multiple natural language emotion datasets show our model's effectiveness. FS-CLAP surpasses the baseline CLAP model. It also demonstrates strong competitiveness against current state-of-the-art methods.

Index Terms— contrastive language-audio pretraining, emotion, speaking style description, retrieval

1. INTRODUCTION

Contrastive learning paradigms, exemplified by CLIP [1] in vision and CLAP [2] in speech, have achieved remarkable success in learning general-purpose representations with strong zero-shot capabilities. However, their reliance on coarse-grained global features leads to poor generalization on fine-grained downstream speech tasks, as they fail to model detailed acoustic information like emotional fluctuations and timbral qualities.

This limitation is particularly acute in Speech Emotion Recognition (SER), a field historically constrained by labeled data scarcity and rigid, predefined emotion categories. While self-supervised learning (SSL) models like Wav2Vec 2.0 [3] mitigate data dependency, they still demand extensive supervised fine-tuning for SER, thus not fully resolving the reliance on large-scale labeled data. As the research focus evolves from simple classification to more descriptive tasks like Speech Emotion Captioning (SEC) [4] and Speaking Style Captioning (StyleCap) [5], the need for fine-grained modeling has become paramount. Yet, existing CLAP fine-tuning methods like EmotionRankCLAP [6] continue to operate on coarse-grained feature alignment.

To address the pressing user need for retrieving speech via natural language, the Emotional Speaking Style Retrieval (ESSR) task has emerged, aiming to model complex details like dynamic emotional expressions and unique speaker characteristics. The recent RA-CLAP [7] has validated the feasibility of applying a contrastive framework to ESSR.

We find that pre-trained CLAP models exhibit significant performance bottlenecks on the Emotional Speaking Style Retrieval (ESSR) task, both in zero-shot settings and after fine-tuning. We attribute this to their coarse-grained pre-training, which fails to disentangle fine-grained emotional and stylistic traits, leading to two critical issues: insufficient cross-modal feature alignment and representation collapse during the fine-tuning process. To address these problems, we propose FS-CLAP, a fine-grained cross-modal based on contrastive learning method. Its Fusion-Separation Block simultaneously achieves detailed feature alignment while enhancing the discriminative power of the model's representations, thereby effectively mitigating collapse and significantly boosting the CLAP framework's performance on ESSR tasks. The core contributions of this paper are as follows:

- We propose a Fusion-Separation Block to achieve fine-grained feature fusion and alignment while preserving the CLAP contrastive learning architecture.

- Experiments on four public datasets show that FS-CLAP exhibits superior performance on the ESSR task, validating the effectiveness and competitiveness of the method.

- We propose a standardized pipeline for audio-text data collection and cleaning to facilitate the rapid expansion of high-quality cross-modal databases for research.

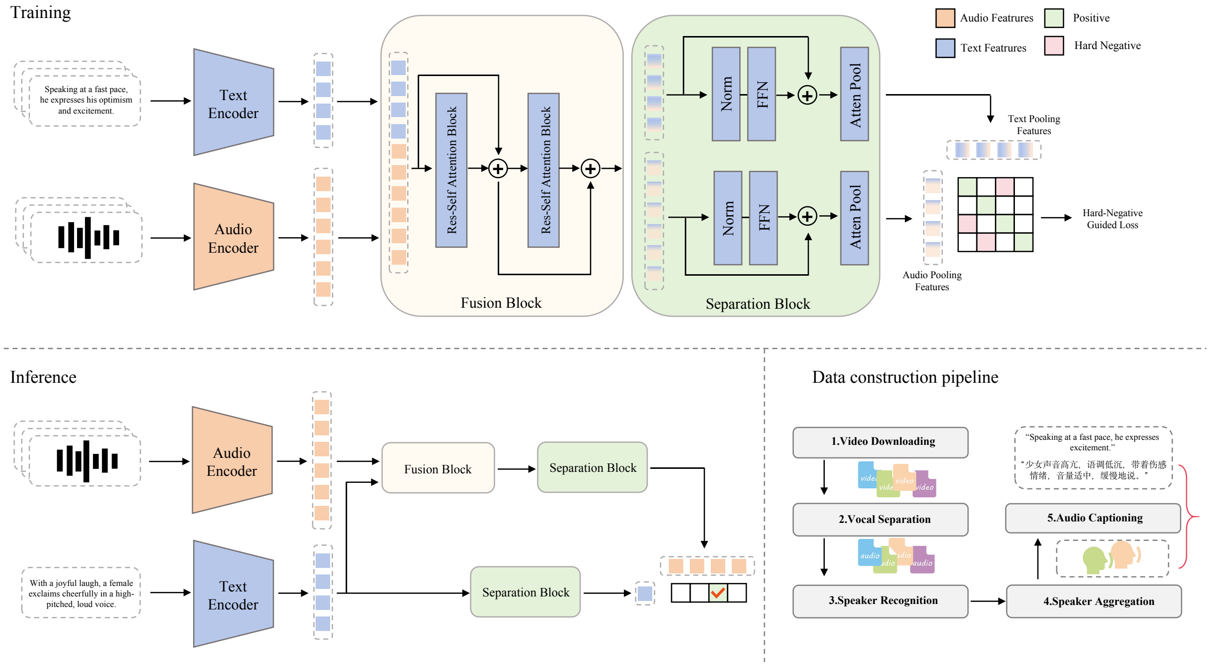

Fig. 1: An overview of our proposed model: unimodal features are processed by a Fusion module for deep interaction, followed by a Separation module for individual refinement and pooling. The resulting global features are then aligned using a HN loss.

2. RELATED WORK

2.1. Contrastive Language-Audio Pre-training

Contrastive cross-modal pre-training, originating in the vision-language domain with the pioneering CLIP model [1], learns concepts from raw text to enable strong zero-shot transfer. Inspired by this, the CLAP framework [8, **?**] established a foundational paradigm in the audio domain by constructing a unified embedding space for audio-language retrieval. Subsequent research has expanded upon this foundation: AudioCLIP [9] extended the framework to a tri-modal (image, text, audio) representation to set a new baseline for zero-shot classification, while MGA-CLAP [10] leveraged a modality-shared codebook to enhance fine-grained correspondence, boosting performance on multi-grained tasks.

2.2. Emotional Speaking Style CLAP for Speech Retrieval

The focus of contrastive cross-modal audio pre-training has shifted from general-purpose representations to specialized, fine-grained applications. This includes using contrastive learning for style modeling in controllable speech generation (e.g., Calm [11]) and for deeper emotional speech understanding, such as modeling emotion intensity gradients (e.g., EmotionRankCLAP [6]). This evolution has culminated in the new task of Emotional Speaking Style Retrieval (ESSR). RA-CLAP [7] is the first dedicated solution designed for this task.

3. METHODOLOGY

As shown in Fig. 1, the FS-CLAP framework extracts fine-grained features using a CLAP encoder, which are then processed by a Fusion-Separation module. The entire model is optimized with a Hard Negative Guided contrastive loss.

3.1. Fusion-Separation Module

Multi-modal Representations in CLAP. CLAP's bi-encoder architecture is designed to align audio and text representations. It uses a dedicated audio encoder f to map a waveform x_i features p_i and a text encoder g to map text y_i to features Q_i.

These discrete sequence features have dimensional differences and cannot be directly compared. Therefore, an aggregator, h (like mean pooling or attention pooling), is usually needed to compress these sequences into single, fixed-dimension fine-grained features, $\hat{p}_i$ and $\hat{q}_i$. Finally, a symmetric contrastive loss function is used. In a common semantic space, it pulls matched audio-text pairs $(\hat{p}_i, \hat{q}_i)$ closer together while pushing mismatched pairs $(\hat{p}_i, \hat{q}_j)$ further apart. This ultimately creates an aligned semantic space.

The architecture suffers from two critical flaws. The first is representation collapse, where the dominance of one modality causes its representations to become overly clustered, thereby losing both discriminability and diversity. The second is the loss of fine-grained correspondence between the features of the two modalities, caused by premature feature compression. These issues prevent the model from learning

Table 1: Performance comparison on the GigaSpeech, TextrolSpeech and Zhvoice datasets. The best results are marked in **bold**, and $\underline{\text{underline}}$ indicates the second-best performance.

Model	**GigaSpeech**						**TextrolSpeech**						**Zhvoice**					
	Audio-to-Text			Text-to-Audio			Audio-to-Text			Text-to-Audio			Audio-to-Text			Text-to-Audio		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CLAP-Laion	21.1	53.3	71.0	21.1	54.1	69.8	19.6	57.8	80.9	23.1	57.8	$\underline{84.4}$	40.1	74.3	**87.3**	39.2	73.0	85.7
CLAP-Microsoft	$\underline{24.6}$	**58.4**	**72.8**	23.6	$\underline{57.6}$	**72.5**	20.1	$\underline{60.3}$	75.9	21.6	62.3	79.9	38.4	$\underline{77.2}$	85.2	35.4	$\underline{75.1}$	**86.1**
AudioCLIP	24.1	$\underline{57.8}$	70.6	$\underline{24.7}$	56.4	71.6	20.1	54.8	76.4	24.6	60.3	81.4	**43.9**	75.1	83.5	**45.2**	74.3	85.7
RA-CLAP	-	-	-	-	-	-	$\underline{23.5}$	59.0	$\underline{82.0}$	$\underline{25.5}$	$\underline{64.5}$	81.0	-	-	-	-	-	-
MGA-CLAP	22.4	52.0	67.1	20.3	52.2	67.0	23.1	55.8	79.4	22.1	**64.8**	78.4	22.4	53.6	67.1	19.4	51.5	65.4
FS-CLAP(ours)	**25.5**	56.8	$\underline{71.7}$	**24.8**	**57.9**	$\underline{72.2}$	**28.6**	**63.3**	**82.4**	**28.6**	64.3	**84.9**	$\underline{42.6}$	**77.6**	**87.3**	$\underline{43.5}$	**76.0**	**86.1**

the nuanced emotional connections required for tasks like ESSR, especially in data-scarce scenarios.

Fusion-Separation Module. To achieve fine-grained cross-modal modeling, our proposed architecture integrates a Fusion Module followed by a Separation Module (SM).

Table 2: Performance comparison by mAP@10.

Model	**GigaSpeech**		**TextrolSpeech**		**Zhvoice**	
	Audio-to-Text	Text-to-Audio	Audio-to-Text	Text-to-Audio	Audio-to-Text	Text-to-Audio
CLAP-Laion	35.0	34.9	36.6	39.6	54.5	54.0
CLAP-Microsoft	$\underline{38.5}$	37.7	36.4	39.2	54.2	51.6
AudioCLIP	38.1	$\underline{38.5}$	36.4	40.1	$\underline{56.2}$	$\underline{57.1}$
RA-CLAP	-	-	$\underline{39.9}$	$\underline{43.5}$	-	-
MGA-CLAP	34.9	33.8	37.8	39.3	35.0	32.8
FS-CLAP(ours)	**38.6**	**39.0**	**43.4**	**44.8**	**57.6**	**57.9**

We propose a Fusion Module to integrate fine-grained features from both modalities into a holistic representation, addressing the representation collapse caused by the dominance of a single modality. To achieve this, the module first obtain features extracted by the unimodal encoders. These are the audio features $h_a = \{\hat{e}^i \in \mathbb{R}^D | i = 1 \dots N\}$ and the text features $h_t = \{\hat{e}^j \in \mathbb{R}^D | j = 1 \dots M\}$. These two feature sequences are concatenated into a combined sequence, $h_f = \{\hat{e}^k \in \mathbb{R}^D | k = 1 \dots (N+M)\}$, which serves as the input to the Fusion Module. This module, featuring two multi-head residual self-attention blocks, processes h_f to model complex inter-modal dependencies and output deeply fused features. And then the Separation Module takes the fused features h_f as input. It separates them back into their original modal streams and refines them using parallel Feed-Forward Network (FFN). Finally, an attention pooling operation extracts the global representations required for contrastive learning.

3.2. Hard Negative Guided Contrastive Loss

MGA-CLAP [10] introduced the HN Loss for audio-language tasks, inspired by the effectiveness of hard negatives in contrastive learning [12]. This loss uses dynamic weighting to efficiently focus optimization on hard negatives without extra data processing. The loss function is reformulated as follows:

$$\begin{aligned}\mathcal{L}_{\text{HN}} = -(&\sum_{i=1}^{B} \log \frac{e^{\langle \tilde{p}_i, \tilde{q}_i \rangle / \tau}}{e^{\langle \tilde{p}_i, \tilde{q}_i \rangle / \tau} + \sum_{j, j \neq i} \alpha_{i,j} e^{\langle \tilde{p}_i, \tilde{q}_j \rangle / \tau}} \\ &+ \sum_{i=1}^{B} \log \frac{e^{\langle \tilde{q}_i, \tilde{p}_i \rangle / \tau}}{e^{\langle \tilde{q}_i, \tilde{p}_i \rangle / \tau} + \sum_{j, j \neq i} \beta_{i,j} e^{\langle \tilde{q}_i, \tilde{p}_j \rangle / \tau}})\end{aligned} \tag{1}$$

where $\alpha_{i,j}$, $\beta_{i,j}$ represent modality-specific difficulty scores for unpaired audio-to-text and text-to-audio samples, respectively. They dynamically up-weight hard negative pairs while down-weighting easy ones, focusing the model's training on the most informative examples. The formula is written as:

$$\alpha_{i,j} = \frac{B e^{\gamma < \tilde{p}_i, \tilde{q}_j > / \tau}}{\sum_k e^{\gamma < \tilde{p}_i, \tilde{q}_k > / \tau}}, \quad \beta_{i,j} = \frac{B e^{\gamma < \tilde{q}_i, \tilde{p}_j > / \tau}}{\sum_k e^{\gamma < \tilde{q}_i, \tilde{p}_k > / \tau}} \tag{2}$$

where γ is the scaling ratio, with a default value of 0.15.

We regularize both audio and text features to mitigate multi-modal alignment deviation caused by instability within single-modality features. The total loss function as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{HN}} + \frac{\mathbb{E}\left[|\mathbf{h}_a|\right]}{\|\mathbf{h}_a\|_2} + \frac{\mathbb{E}\left[|\mathbf{h}_t|\right]}{\|\mathbf{h}_t\|_2} \tag{3}$$

where $\mathbf{h}_a$ and $\mathbf{h}_t$ are the audio and text feature vectors, respectively, and $\mathbb{E}$ denotes the mean absolute value of a vector's elements.

3.3. Data Construction Pipeline

Fig. 1 illustrates our pipeline for rapidly building a dataset of emotion description text-audio pairs to expand our database.

Video downloading. We collected 666 hours of content from 333 popular short dramas in internet, ensuring each had at least 50 episodes.

Vocal separation & Speaker recognition. First, we extract the audio from the video and use Demucs [13] to separate vocals from background sounds. For speaker diarization, we downsample the audio to 24kHz and apply the Paraformer-large model [14] with CAM++ [15] clustering to obtain sentence-level speaker labels.

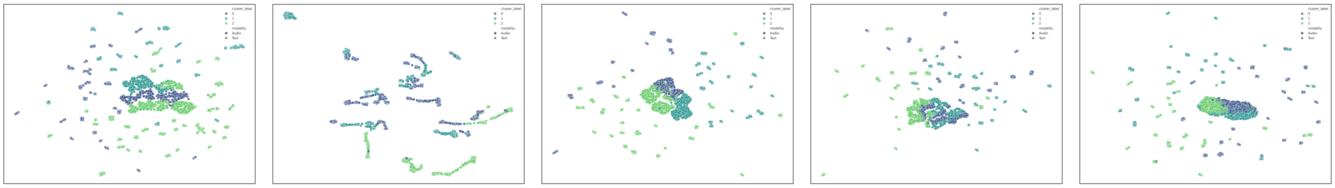
(a) (b) (c) (d) (e)

Fig. 2: Qualitative analysis on the PromptSpeech dataset. (a)-(e) show the UMAP visualization results for Ours, CLAP-Laion, CLAP-Microsoft, AudioCLIP, and MGA-CLAP, respectively.

Speaker aggregation & Audio captioning. To correct speaker diarization errors where multiple people are assigned a single label, we quantify the consistency of each speaker cluster. Our method calculates a mean similarity score for each speaker based on the pairwise cosine similarity of all their assigned audio embeddings. The formula for cosine similarity is as follows:

$$\cos(\theta_{i,j}) = \frac{\vec{v}_i \cdot \vec{v}_j}{\|\vec{v}_i\| \times \|\vec{v}_j\|} \tag{4}$$

where $\vec{v}_i$ and $\vec{v}_j$ represent the feature vectors of two different audio segments, $\theta_{i,j}$ represents the angle between the two vectors, and $\|\vec{v}_i\|$ and $\|\vec{v}_j\|$ represent the L2 norm of the two feature vectors, respectively.

Then, we calculate the mean similarity of a single audio clip with the other audio clips in the same category. The formula is written as:

$$\bar{s}_i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\cos(\theta_{i,j}) + 1}{2} \tag{5}$$

We first filter for relevant audio (mean similarity ≥ 0.65), then annotate the remaining clips with emotion descriptions using SenseVoice [16].

4. EXPERIMENTS

4.1. Datasets

PromptSpeech (PS) [17]: Contains 38 hours of speech (26,588 samples) paired with text prompts describing style and content, including both synthetic and real audio sourced from LibriTTS. TextrolSpeech (TS) [18]: Provides 330 hours of speech annotated with 236,220 natural text descriptions, compiled from various public emotional speech datasets. GigaSpeech-M [19]: A large-scale, 10,000-hour English speech corpus sourced from diverse audio like audiobooks and podcasts, covering various topics and both read and spontaneous speaking styles. Zhvoice [19]: A 900-hour Chinese corpus compiled from 8 open-source datasets. We used a subset of 506,850 utterances for our experiments.

Table 3: Ablation study, where FU, SE and HN represent the Fusion module, Separation module, and HN loss, respectively.

			TextrolSpeech				Zhvoice			
			Audio-to-Text		Text-to-Audio		Audio-to-Text		Text-to-Audio	
FU	**SE**	**HN**	mAP@10	R@1	mAP@10	R@1	mAP@10	R@1	mAP@10	R@1
×	×	×	36.6	19.6	39.6	23.1	54.5	40.1	54.0	39.2
✓	×	×	40.8	25.6	43.8	**29.7**	57.4	**44.3**	57.7	**43.9**
×	✓	×	38.5	22.6	41.6	24.1	57.3	41.4	56.1	40.9
×	×	✓	40.4	25.6	41.3	25.6	56.5	42.2	55.3	40.9
✓	✓	×	40.8	23.6	43.6	27.1	55.4	40.5	56.4	41.4
✓	✓	✓	**43.4**	**28.6**	**44.8**	28.6	**57.6**	42.6	**57.9**	43.5

4.2. Implementation Details

In all experiments, every model was trained for 10 epochs using the AdamW optimizer with a batch size of 96, 2 gradient accumulation steps, and a weight decay of 0.01. We used the Cosine Annealing Learning Rate strategy with 1 warmup epoch. The peak learning rates for different modules were set as follows 4e-5 for the CLAP-Laion [**?**] and 1e-4 for the Fusion-Separation part. The final model checkpoint was selected based on its performance on the validation set, and its performance was then evaluated on the corresponding test set.

4.3. Results and Ablation study

As shown in Table 1 and Table 2, our FS-CLAP model consistently achieves top-two results against SOTA baselines. For fair comparison, all models were fine-tuned under the same settings. Baseline CLAP models [**?**, 8, 7, 10, 9] suffer from severe imbalance on the PromptSpeech dataset [17]. UMAP visualizations (Fig. 2 (a)-(e)) directly demonstrate our model's superior feature alignment. We argue this balanced representation is fundamentally important, and our model's stronger overall performance on larger datasets (TS, GigaSpeech, and Zhvoice) validates this capability. Furthermore, we conducted ablation experiments to verify the effectiveness of the sub-modules. As shown in Table 3, the combination of the Fusion-Separation module yields better performance.

5. CONCLUSION

In this paper, we introduce FS-CLAP, a novel framework for ESSR task that tackles the challenges of insufficient representation alignment and representation collapse. Our method establishes a new fine-grained alignment paradigm by leveraging full-sequence features, a fusion-separation module, and a contrastive loss emphasizing hard negatives. Comprehensive experiments and ablation studies validate that FS-CLAP achieves state-of-the-art performance and demonstrate the efficacy of its components. Practically, FS-CLAP can enhance Text-to-Speech systems by retrieving more suitable emotional style prompts. We also release our standardized data pipeline for future research.

6. REFERENCES

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., “Learning transferable visual models from natural language supervision,” in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[2] Yusong Wu, Ke Chen, and etc. Zhang, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[3] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” *Advances in neural information processing systems*, vol. 33, pp. 12449–12460, 2020.

[4] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shi-Xiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu, “Secap: Speech emotion captioning with large language model,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024, vol. 38, pp. 19323–19331.

[5] Kazuki Yamauchi, Yusuke Ijima, and Yuki Saito, “Stylecap: Automatic speaking-style captioning from speech based on speech and language self-supervised learning models,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11261–11265.

[6] Shreeram Suresh Chandra, Lucas Goncalves, Junchen Lu, Carlos Busso, and Berrak Sisman, “Emotion-rankclap: Bridging natural language speaking styles and ordinal speech emotion via rank-n-contrast,” *arXiv preprint arXiv:2505.23732*, 2025.

[7] Haoqin Sun, Jingguang Tian, Jiaming Zhou, Hui Wang, Jiabei He, Shiwan Zhao, Xiangyu Kong, Desheng Hu, Xinkang Xu, Xinhui Hu, et al., “Raclap: Relation-augmented emotional speaking style contrastive language-audio pretraining for speech retrieval,” *arXiv preprint arXiv:2505.19437*, 2025.

[8] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang, “Natural language supervision for general-purpose audio representations,” 2023.

[9] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, “Audioclip: Extending clip to image, text and audio,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[10] Yiming Li, Zhifang Guo, Xiangdong Wang, and Hong Liu, “Advancing multi-grained alignment for contrastive language-audio pre-training,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7356–7365.

[11] Yi Meng, Xiang Li, Zhiyong Wu, Tingtian Li, Zixun Sun, Xinyu Xiao, Chi Sun, Hui Zhan, and Helen Meng, “Calm: contrastive cross-modal speaking style modeling for expressive text-to-speech synthesis,” *arXiv preprint arXiv:2308.16021*, 2023.

[12] Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka, “Contrastive learning with hard negative samples,” *arXiv preprint arXiv:2010.04592*, 2020.

[13] Alexandre Défossez, “Hybrid spectrogram and waveform source separation,” in *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021.

[14] Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan, “Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition,” *arXiv preprint arXiv:2206.08317*, 2022.

[15] Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen, “Cam++: A fast and efficient network for speaker verification using context-aware masking,” *arXiv preprint arXiv:2303.00332*, 2023.

[16] Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, et al., “Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms,” *arXiv preprint arXiv:2407.04051*, 2024.

[17] Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan, “Prompttts: Controllable text-to-speech with text descriptions,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao, “Textrolspeech: A text style control speech corpus with codec language text-to-speech models,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10301–10305.

[19] Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu, “Speechcraft: A fine-grained expressive speech dataset with natural language description,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1255–1264.