EmotionRankCLAP: Bridging Natural Language Speaking Styles and Ordinal Speech Emotion via Rank-N-Contrast

Shreeram Suresh Chandra[1,2], *Lucas Goncalves*[3], *Junchen Lu*[4], *Carlos Busso*[5], *Berrak Sisman*[1]

[1]Center for Language and Speech Processing (CLSP), Johns Hopkins University, USA
[2]The University of Texas at Dallas, USA [3]Amazon, USA [4]NUS, Singapore
[5]Language Technologies Institute (LTI), Carnegie Mellon University, USA

busso@cmu.edu, sisman@jhu.edu

Abstract

Current emotion-based *contrastive language-audio pretraining* (CLAP) methods typically learn by naïvely aligning audio samples with corresponding text prompts. Consequently, this approach fails to capture the ordinal nature of emotions, hindering inter-emotion understanding and often resulting in a wide modality gap between the audio and text embeddings due to insufficient alignment. To handle these drawbacks, we introduce EmotionRankCLAP, a supervised contrastive learning approach that uses dimensional attributes of emotional speech and natural language prompts to jointly capture fine-grained emotion variations and improve cross-modal alignment. Our approach utilizes a Rank-N-Contrast objective to learn ordered relationships by contrasting samples based on their rankings in the valence-arousal space. EmotionRankCLAP outperforms existing emotion-CLAP methods in modeling emotion ordinality across modalities, measured via a cross-modal retrieval task.[1]

Index Terms: emotion, ordinality, contrastive language-audio pretraining, speaking style descriptions

1. Introduction

The expression and perception of human emotion are inherently continuous in nature [1]. Emotions also possess an ordinal nature, as humans are more adept at detecting relative changes in expression rather than identifying absolute emotional states [2]. However, existing paralinguistic models that attempt to capture the ordinality of speech emotion primarily rely on dimensional attribute annotations [3,4], which limit their ability to fully represent the nuanced structure of emotional expression. We believe that fine-grained and ordinal nature of speech emotion can be more effectively captured with natural language descriptions.

Natural language supervision has emerged as a promising approach for enhancing audio and speech understanding. In particular, *contrastive language-audio pretraining* (CLAP) [5] has gained popularity as a method for aligning audio with natural language prompts. By sharing a common representation space across modalities, CLAP enables tasks such as zero-shot captioning [6], classification [7], and cross-modal retrieval [8].

CLAP has also been adopted extensively for emotion tasks including *speech emotion recognition* (SER) [9], emotional *text-to-speech* (TTS) [10] and *emotion audio retrieval* (EAR) [11]. GemoCLAP [9] focuses on building a discriminative representation space for SER using categorical labels. ParaCLAP [12] and CLAP with prompt-augmentation [11] improve supervision by describing acoustic properties of emotional audio. The work most similar to ours, CLAP4emo [13], generates pseudo-captions using pre-trained *large language models* (LLMs) based on categorical emotion annotations of speech utterances. With the current approach of using only categorical emotions, intra-class variability is overlooked—for instance, all speech-text pairs labeled as "happiness" are treated identically, ignoring differences in intensity or expression. Likewise, inter-class relationships are not captured, such as the fact that "disgust" and "fear" are more closely related than "happiness" and "fear".

A key limitation in existing CLAP-based models is their reliance on the diagonal-constraint-based *symmetric cross-entropy* (SCE) loss [14], which presents two major drawbacks. First, at the batch level, this loss function fails to capture inter-emotion relationships across modalities. Since emotions are inherently ordinal, aligning each speech-text pair in isolation overlooks the structured relationships between different emotional states. Secondly, while emotion-based CLAP models leverage emotion annotations in text prompt design, they retain the loss formulation of CLIP [14], designed originally for self-supervised training, leading to a modality gap between text and audio embeddings at the end of training. Here, the modality gap refers to the insufficient overlap of embedding spaces of different modalities, a well-documented issue in cross-modal learning frameworks [15]. We argue that this modality gap can be effectively reduced in a supervised setting by incorporating dimensional emotional attributes in speech.

To address these limitations, we adopt Rank-N-Contrast [16], a contrastive learning objective specifically designed to learn ordered representations by ranking samples relative to their positions in the target label space. This objective ensures that the learned representations maintain the intended ordinal structure, aligning with the target rankings. While extensively studied in regression tasks, its application to cross-modal representation learning, particularly for capturing the ordinality of emotions, remains unexplored. In this work, we introduce EmotionRankCLAP, a novel supervised contrastive learning strategy that uses dimensional emotional attributes to learn a continuous emotion embedding space with the cross-modal formulation of Rank-N-Contrast. Our key contributions are as follows:

- We propose leveraging the ordinal nature of emotions to learn a fine-grained emotion embedding space, using the Rank-N-Contrast objective;
- We show that using Rank-N-Contrast as an alternative to symmetric cross entropy loss improves cross-modal alignment, bringing the distributions of the audio embeddings and text embeddings closer together;
- We formulate a cross-modal retrieval task that checks the emotion ordinal consistency of the audio and text embeddings - and we show EmotionRankCLAP outperforms other emotion-based CLAP models in this test.

[1]https://kodhandarama.github.io/emotionrankclap.github.io/

- We generate and release natural-language emotional speaking style descriptions based on dimensional emotion attributes from the MSP-Podcast corpus [17] (release 1.12) to bridge the speech and text modalities in the CLAP model.

To the best of our knowledge, this study is the first to leverage the ordinal nature of speech emotions to align the continuums of dimensional speech emotion and natural language speaking style descriptions.

2. Related work

2.1. Cross-modal contrastive learning

Contrastive learning has proven to be an effective approach for aligning multiple modalities in shared representation spaces [5, 14, 18]. While unsupervised contrastive learning relies solely on modality co-occurrence, it can lead to imprecise alignments without capturing task-specific semantic relationships, prompting the exploration of supervised settings [19, 20]. By incorporating supervision, contrastive learning frameworks can better capture fine-grained inter-modality relationships, making them particularly effective for emotion-related tasks. Inspired by these strategies, we propose a cross-modal version of Rank-N-Contrast, leveraging dimensional emotional attributes as an additional supervision signal to improve speech-text alignment.

2.2. Natural language description of speech emotion

Emotion annotations have traditionally been limited to manually annotated categorical labels or dimensional attributes. However, recent advancements have shifted the focus towards using natural language, allowing for more descriptive representations of speech emotion. This has been made possible thanks to caption generation capability of LLMs [21]. This capability has been adapted into multimodal SER models [22,23] to generate pseudo-captions. Speech language models like SECap [24] and AlignCap [25] present a paradigm shift away from SER and towards speech emotion captioning via speech language models. Similarly, emotional TTS models are increasingly prioritizing controllability by using natural speaking style prompts rather than relying solely on categorical emotion labels [26]. Our approach leverages an LLM to generate speaking style descriptions in the absence of speech datasets with captions.

3. EmotionRankCLAP

We propose EmotionRankCLAP, a supervised cross-modal contrastive learning framework to align emotional speech with natural language speaking style descriptions in a shared embedding space, leveraging the ordinal nature of speech emotions through a Rank-N-Contrast learning objective.

3.1. Problem Formulation

Let $\{X_i^a, X_i^t\}$ for $i \in \{1, ..., N\}$ be a batch of <speech, text> pairs. Input from audio and text modalities are first encoded via two separate encoders, $f^a(.)$ and $f^t(.)$, yielding embeddings:

$$\hat{X}_i^a = f^a(X_i^a); \quad \hat{X}_i^t = f^t(X_i^t), \tag{1}$$

where $\hat{X}^a \in \mathbb{R}^{N \times V}$ and $\hat{X}^t \in \mathbb{R}^{N \times U}$. We employ a pre-trained, frozen WavLM-based dimensional SER model.[2] [27] as the audio encoder $f^a(.)$, extracting 1024-dimensional embeddings via attentive statistics pooling across the temporal dimension from the last transformer layer. The text encoder $f^t(.)$ is a pre-trained, frozen DistilRoBERTa model [3] [28], using the final-layer [CLS] token as a 768-dimensional embedding. These representations are then projected to the same dimension $D = 512$:

$$\hat{E}_i^a = proj^a(\hat{X}_i^a); \quad \hat{E}_i^t = proj^t(\hat{X}_i^t), \tag{2}$$

where $\hat{E}^a, \hat{E}^t \in \mathbb{R}^{N \times D}$ are the projected embeddings, and $proj^a$ and $proj^t$ are modules with a linear transformation followed with ReLU activation. The goal of EmotionRankCLAP is to align the two modalities in the same embedding space while preserving ordinality to capture the dimensional nature of emotion in both text descriptions and speech.

3.2. Supervised contrastive learning with Rank-N-Contrast

Emotions are inherently continuous and ordinal, meaning that within any batch of emotional speech and its corresponding speaking style descriptions, a structured relationship exists between each possible pair, totaling $N \times N$ cross-modal pairs. To learn this structured relationship, we adopt Rank-N-Contrast, which contrasts samples based on their rankings in valence-arousal label space.

In the proposed formulation, we jointly model the ordinality of valence and arousal by considering them together in the label space. Valence reflects the sentiment expressed in the utterance, ranging from negative to positive. Arousal indicates the level of activation, with values spanning from calm to highly active.

For a given audio embedding anchor $\hat{E}_i^a$, the likelihood of association with a text embedding $\hat{E}_j^t$ depends on the relative distance of their labels in the valence-arousal space. Emotional distance is assessed by the L_2 distance between ($valence_i^a$, $arousal_i^a$) and ($valence_j^t$, $arousal_j^t$), where closer samples are considered more alike. Here, i and j denote sample indices.

Let $S_{i,j} := \{\hat{E}_k^t \mid d(\hat{E}_i^a, \hat{E}_k^t) > d(\hat{E}_i^a, \hat{E}_j^t)\}$ denote the set of text embeddings that are of higher rank than $\hat{E}_j^t$ in terms of label distance with respect to $\hat{E}_i^a$, where $d(\cdot, \cdot)$ is the L_2 distance measure between two labels in the valence-arousal plane.

Then the normalized likelihood of $\hat{E}_j^t$ given $\hat{E}_i^a$ and $S_{i,j}$ can be written as

$$P(\hat{E}_j^t \mid \hat{E}_i^a, S_{i,j}) = \frac{\exp(\text{sim}(\hat{E}_i^a, \hat{E}_j^t)/\tau)}{\sum_{\hat{E}_k^t \in S_{i,j}} \exp(\text{sim}\hat{E}_i^a, \hat{E}_k^t)/\tau)}, \tag{3}$$

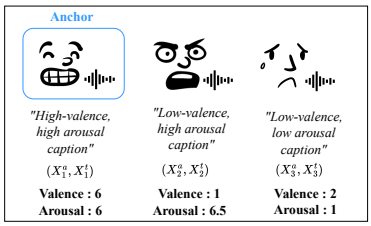
where $S_{i,j}$ represents the set of all $\hat{E}_k^t$ that satisfy the ranking condition with respect to $\hat{E}_i^a$ and $\hat{E}_j^t$. This set contains the corresponding negative pairs for the positive pair $\hat{E}_i^a, \hat{E}_j^t$. The similarity function $\text{sim}(x, y) = \frac{x^T y}{\|x\| \cdot \|y\|}$ calculates the cosine similarity between cross-modal features and τ denotes the temperature parameter. Defining this objective over all samples in a batch, we get the Rank-N-Contrast cross- modal loss:

$$\mathcal{L}_{\text{RNC-CM}} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} -\log P(\hat{E}_j^t \mid \hat{E}_i^a, S_{i,j}). \tag{4}$$

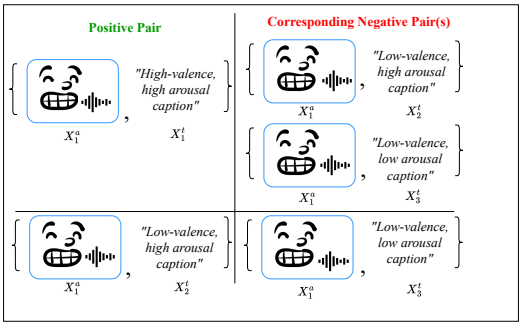
The loss function $\mathcal{L}_{\text{RNC-CM}}$ exploits the continuous structure of the valence-arousal label space to ensure that emotional speech samples and speaking style descriptions with similar valence-arousal values also remain close in the learned representation space. The Rank-N-Contrast formulation enhances cross-modal alignment by leveraging all $N \times N$ speech-text pairs within a

[2] https://huggingface.co/3loi/SER-Odyssey-Baseline-WavLM-Multi-Attributes

[3] https://huggingface.co/j-hartmann/emotion-english-distilroberta-base



(a) Given batch of speech-text pairs



(b) Cross-modal pair construction with Rank-N-Contrast

Figure 1: *Illustration of Rank-N-Contrast in a cross-modal setting. The anchor is boxed in blue. (a) A batch of speech-text pairs along with their valence-arousal labels. (b) Positive and negative pair selection via Rank-N-Contrast criteria.*

batch to form positive-negative pairs based on a ranking criterion. Each positive pair is assigned corresponding negative pairs according to their similarity ranking, ensuring a structured contrastive learning process. In contrast, SCE uses only N positive pairs per batch, limiting cross-modal alignment.

3.3. Illustrative example of positive/negative pair selection

We consider a batch of three speech-text pairs (X_i^a, X_i^t) $(i \in \{1, 2, 3\})$ with corresponding valence-arousal annotations as shown in Figure 1(a). As a demonstration of the positive/negative pair selection, we set the first speech utterance X_1^a as the anchor. Figure 1(b) illustrates two positive pairs and their corresponding negative pairs.

When considering the pair (X_1^a, X_1^t) as positive, $d(X_1^a, X_1^t) = 0$ as both share the same label. This makes X_2^t and X_3^t negative samples since $d(X_1^a, X_2^t) > 0$ and $d(X_1^a, X_3^t) > 0$. Similarly, when X_2^t forms a positive pair with X_1^a, X_3^t is a negative sample since $d(X_1^a, X_3^t) > d(X_1^a, X_2^t)$. In this case, X_1^t is not a negative sample since $d(X_1^a, X_1^t) < d(X_1^a, X_2^t)$.

Thus, structured relationships emerge: closer positive pairs tend to have more negative samples, reinforcing their closeness, while distant positive pairs have fewer negative samples, reducing their attraction. For a batch of N, we iterate over each X_i^a for $i = 1, \ldots, N$, forming N relationships per anchor, resulting in $N \times N$ structured relationships.

3.4. Generation of speaking style descriptions

Existing speech emotion datasets are primarily designed for categorical and dimensional emotion recognition [17, 29], providing annotations in terms of categorical labels and dimensional attributes (valence, arousal, dominance). In contrast, speech emotion captioning remains an emerging field, with limited datasets featuring manually annotated speaking style descriptions. To bridge this gap, we make use of an LLM [30] to generate pseudo-captions based on valence and arousal. While this work focuses on these two attributes, incorporating dominance and other factors is left for future exploration. Figure 2 contains the prompt used to generate the natural language speaking style descriptions using OpenAI's o1 model.

> "Given the following scale of emotions - valence (1-very negative; 7-very positive), arousal (1-very calm; 7-very active), write a sentence describing a speaking style that is {*VALENCE*} on valence, {*AROUSAL*} on arousal. Do not use any numbers in the sentence. The sentence should start with: The person is speaking ..."

Figure 2: *Prompt used to generate emotional style descriptions based on valence-arousal values.*

4. Experiments

In this section, we discuss the experimentation settings, the baselines and the evaluations used to probe the properties of the cross-modal embeddings.

4.1. Experimental setup

Dataset: We use the MSP-Podcast v1.12 corpus [17] for training, validation, and testing. Collected from real-world podcasts, it features significant acoustic variability, diverse speakers, and a broad range of emotional expressions, making it particularly challenging. We filter out samples with categorical emotion labels 'X' (no agreement) and 'O' (other), resulting in 90,022 training, 25,258 development, and 34,963 test samples (using only test 1 set). The large test set provides a comprehensive coverage of speaking styles. Each speech utterance is annotated with (valence, arousal, dominance) based on annotations provided by at least five annotators. We utilize the average score across annotators.

Baselines:

- **CLAP-template:** This model is trained with the CLAP framework (SCE loss) using the text prompt: ''speech has {categorical label} emotion" as input to the text encoder.
- **CLAP4emo [13]:** This model replaces the pre-defined prompts in CLAP-template with natural language style descriptions generated with the help of ChatGPT [30] and an NRC lexicon [31]. The captions for this model are generated following the pipeline described in their paper.
- **CLAP-SCE (A-V):** An ablation model trained with CLAP framework under SCE loss, where we use captions generated with dimensional emotional attributes instead of categorical emotion. The difference between this method and the proposed method is the loss function (SCE vs RNC). Here, (A-V) indicates that we use dimensional emotional attributes to generate the captions for this method.
- **SupConCLAP (A-V):** Another ablation baseline where we replace the SCE loss in CLAP-SCE (A-V) with SupCon [19], using categorical emotion labels to define the similarity matrix between text and audio embeddings.
- **ParaCLAP [12]:** This model is trained to align emotional audio with acoustic properties like pitch, jitter, shimmer, articulation rate using natural language descriptions. We extract acoustic information using Parselmouth-Praat [32].

Training details: All input waveforms are resampled to 16 kHz and cropped or zero-padded to 10 seconds. Text inputs are truncated to a maximum of 512 tokens. To ensure fair comparisons, all models share the same text and audio encoder architecture, with jointly trained projection layers. We train using the *Adam* optimizer (learning rate 1×10^{-4}), a learnable temperature (initialized at 1.0), and a batch size of 64 on an

Table 1: *Comparison of methods on cross-modal alignment. * denotes statistically significant improvement over all baselines (two-tailed p-test, p<0.05).*

Method	MMD ↓	Wass. Dist. ↓
CLAP-SCE (A-V)	0.096±.002	0.180±.003
SupConCLAP (A-V)	0.4436±.003	0.4172±.003
EmotionRankCLAP (A-V)	***0.087±.001**	***0.065±.007**

Table 2: *Comparison of cross-modal retrieval methods. * indicates a statistically significant improvement over all baselines (two-tailed p-test, p<0.05). AOC and VOC denote arousal and valence ordinal consistency, while KT represents Kendall's Tau.*

Method	AOC (KT) ↑	VOC (KT) ↑
CLAP-template	0.171±.26	0.466±.14
CLAP4emo [10]	0.284±.20	0.533 ±.14
ParaCLAP [8]	0.283±.20	0.217±.20
CLAP-SCE (A-V)	0.492±.15	0.505±.13
SupConCLAP (A-V)	0.346±.19	0.530±.15
EmotionRankCLAP (A-V)	***0.552±.12**	***0.616±.13**

NVIDIA L4 GPU. All models are implemented in PyTorch and trained for 15 epochs, selecting the checkpoint with the lowest validation loss.

4.2. Evaluations and results

4.2.1. Cross-modal alignment

This evaluation tests the overlap of audio and text embedding spaces. We conduct 30 trials, each randomly sampling 5000 speech-text pairs from the MSP Podcast test-1 set. The audio embeddings are extracted from speech utterances, and text embeddings are extracted from the natural language descriptions. We measure *maximum mean discrepancy* (MMD) [33] with a *radial basis function* (RBF) kernel and Wasserstein distance [34] both of which quantify alignment between the embedding distributions, where lower scores indicate better alignment. In this test, we only consider baselines which are trained with the same caption data as the proposed method, and we report the mean and standard deviation across the 30 trials. As shown in Table 1, EmotionRankCLAP significantly outperforms the baselines by achieving the lowest MMD and Wasserstein distance scores, highlighting Rank-N-Contrast's superior cross-modal alignment compared to SCE and SupCon.

4.2.2. Cross-Modality Emotion Ordinality Test

In this evaluation, we examine how well the audio and text embedding spaces preserve ordinal consistency for dimensional emotional attributes. Specifically, ordinal consistency here means that speaking style descriptions indicating higher (or lower) valence (or arousal) should align more closely with speech utterances that exhibit correspondingly higher (or lower) valence (or arousal) levels. We design a cross-modal retrieval task to probe this property. Using the prompt in Figure 2, we generate 100 lists of speaking style descriptions, each containing 14 descriptions. We evaluate two properties: *valence ordinal consistency* (VOC) and *arousal ordinal consistency* (AOC). For VOC, we fix the arousal value in each list and vary valence from 0.5 to 7 in steps of 0.5. The fixed arousal value is incremented by 0.5 across lists, spanning the range [0.5,7], and resets to 0.5 after reaching 7. Conversely, for AOC, we fix the valence value in each list and vary arousal similarly. After generating these lists, we use the trained CLAP model as a retrieval system to find the most similar speech utterances for each textual description. The model encodes the speaking style prompt into a text embedding and retrieves the speech utterance with the closest audio embedding based on cosine similarity. We evaluate this property using the Kendall's Tau coefficient (KT) [35] between the valence (or arousal) values used to generate the speaking style descriptions and the valence (or arousal) values of the retrieved speech utterances, as shown in Figure 3, and report the mean and standard deviation across 100 lists. To prevent redundant retrievals, each item is retrieved only once.

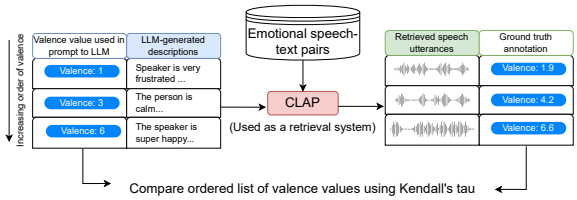


Figure 3: *Cross-Modality Emotion Ordinality Test: This figure shows a three-sample example for valence ordinal consistency, while the actual evaluation uses 14 samples per list, repeated across 100 lists for both valence and arousal.*

We observe that models trained using caption data generated with dimensional attribute guidance (denoted as (A-V)) are more consistent across both VOC and AOC tests. This result highlights the importance of incorporating dimensional attributes when generating speaking style captions, as it helps in enhancing fine-grained cross-modal retrieval and in maintaining ordinal consistency. Interestingly, models trained with captions based on categorical emotions (CLAP-template and CLAP4emo) are competitive in VOC tests, but their performance degrades in AOC tests. Overall, EmotionRankCLAP achieves a significantly higher KT coefficient in both settings—VOC (lists with varying valence) and AOC (lists with varying arousal), as shown in Table. 2. This result demonstrates that our proposed cross-modal Rank-N-Contrast loss along with the use of captions generated with dimensional attribute guidance better preserves the ordinal structure of valence and arousal in the embedding spaces.

5. Conclusions

This work proposes EmotionRankCLAP, a supervised contrastive learning approach that leverages the ordinal nature of emotions to learn a cross-modal representation space to align dimensional speech emotions with corresponding speaking style descriptions. We generate natural language speaking style descriptions using dimensional attributes of speech emotion and we show that this is crucial in preserving emotion ordinality. We show that the proposed cross-modal formulation of Rank-N-Contrast loss improves cross-modal alignment between text and audio embedding spaces. We also design a cross-modal retrieval task to check ordinal consistency between the embedding spaces, and show that EmotionRankCLAP preserves ordinal nature of both valence and arousal better compared to other emotion-based CLAP models. In the future, we will explore other speech emotion tasks that take advantage of close cross-modal alignment and ordinal structure in the embedding space.

6. Acknowledgment

This work is supported by NSF CAREER award IIS-2338979.

7. References

[1] J. A. Russell, "Core affect and the psychological construction of emotion." *Psychological review*, vol. 110, no. 1, p. 145, 2003.

[2] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2018.

[3] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4995–4999.

[4] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE transactions on affective computing*, vol. 5, no. 3, pp. 314–326, 2014.

[5] 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.

[6] S. Deshmukh, B. Elizalde, D. Emmanouilidou, B. Raj, R. Singh, and H. Wang, "Training audio captioning models without audio," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 371–375.

[7] S. Ghosh, S. Kumar, C. K. R. Evuru, O. Nieto, R. Duraiswami, and D. Manocha, "Reclap: Improving zero shot audio classification by describing sounds," *CoRR*, 2024.

[8] S. Deshmukh, B. Elizalde, and H. Wang, "Audio retrieval with wavtext5k and clap training," in *Interspeech 2023*, 2023, pp. 2948–2952.

[9] Y. Pan, Y. Hu, Y. Yang, W. Fei, J. Yao, H. Lu, L. Ma, and J. Zhao, "Gemo-clap: Gender-attribute-enhanced contrastive language-audio pretraining for accurate speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 10 021–10 025.

[10] X. Jing, K. Zhou, A. Triantafyllopoulos, and B. W. Schuller, "Enhancing emotional text-to-speech controllability with natural language guidance through contrastive learning and diffusion models," *arXiv preprint arXiv:2409.06451*, 2024.

[11] H. Dhamyal, B. Elizalde, S. Deshmukh, H. Wang, B. Raj, and R. Singh, "Prompting audios using acoustic properties for emotion representation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 936–11 940.

[12] X. Jing, A. Triantafyllopoulos, and B. Schuller, "Paraclap – towards a general language-audio model for computational paralinguistic tasks," in *Interspeech 2024*, 2024, pp. 1155–1159.

[13] W.-C. Lin, S. Ghaffarzadegan, L. Bondi, A. Kumar, S. Das, and H.-H. Wu, "Clap4emo: Chatgpt-assisted speech emotion retrieval with natural language supervision," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 791–11 795.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[15] C. Yaras, S. Chen, P. Wang, and Q. Qu, "Explaining and mitigating the modality gap in contrastive multimodal learning," *arXiv preprint arXiv:2412.07909*, 2024.

[16] K. Zha, P. Cao, J. Son, Y. Yang, and D. Katabi, "Rank-n-contrast: learning continuous representations for regression," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[17] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 471–483, 2017.

[18] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "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.

[19] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[20] S. Stewart, K. Avramidis, T. Feng, and S. Narayanan, "Emotion-aligned contrastive learning between images and music," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8135–8139.

[21] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[22] S. Dutta and S. Ganapathy, "Llm supervised pre-training for multimodal emotion recognition in conversations," *arXiv preprint arXiv:2501.11468*, 2025.

[23] H. Wu, H.-C. Chou, K.-W. Chang, L. Goncalves, J. Du, J.-S. R. Jang, C.-C. Lee, and H.-Y. Lee, "Empower typed descriptions by large language models for speech emotion recognition," in *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2024, pp. 1–6.

[24] Y. Xu, H. Chen, J. Yu, Q. Huang, Z. Wu, S.-X. Zhang, G. Li, Y. Luo, and R. Gu, "Secap: Speech emotion captioning with large language model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 17, 2024, pp. 19 323–19 331.

[25] Z. Liang, H. Shi, and H. Chen, "Aligncap: Aligning speech emotion captioning to human preferences," *arXiv preprint arXiv:2410.19134*, 2024.

[26] D. Yang, S. Liu, R. Huang, C. Weng, and H. Meng, "Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

[27] L. Goncalves, A. N. Salman, A. R. Naini, L. M. Velazquez, T. Thebaud, L. P. Garcia, N. Dehak, B. Sisman, and C. Busso, "Odyssey 2024-speech emotion recognition challenge: Dataset, baseline framework, and results," *Development*, vol. 10, no. 9,290, pp. 4–54, 2024.

[28] J. Hartmann, "Emotion english distilroberta-base," https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/, 2022.

[29] 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.

[30] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[31] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Computational Intelligence*, vol. 29, no. 3, pp. 436–465, 2013.

[32] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing Parselmouth: A Python interface to Praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[33] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. Mar, pp. 723–773, 2012.

[34] G. Peyré and M. Cuturi, "Computational optimal transport," *Foundations and Trends in Machine Learning*, vol. 11, no. 5-6, pp. 355–607, 2019.

[35] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938. [Online]. Available: https://doi.org/10.2307/2332226