

Multi-Dimensional Ordinal Embedding for Attribute Modeling in Speech Emotion Recognition

Abinay Reddy Naini

Language Technologies Institute (LTI)
Carnegie Mellon University, Pittsburgh, USA
The University of Texas at Dallas, Richardson, USA
anaini@andrew.cmu.edu

Carlos Busso

Language Technologies Institute (LTI)
Carnegie Mellon University
Pittsburgh, USA
cbusso@andrew.cmu.edu

Abstract—Predicting emotional attributes such as arousal, valence, and dominance from speech typically involves modeling absolute scores independently. However, emotional expressions inherently exhibit ordinal relationships, suggesting preference learning frameworks may yield more robust results. While previous studies have separately applied pairwise and list-wise preference learning for *speech emotion recognition* (SER), the potential benefits of jointly learning multiple emotional attributes using preference-based multitask frameworks remain unexplored. This study proposes a novel multi-dimensional ordinal embedding approach inspired by the RankNet formulation, explicitly modeling subtle ordinal differences across arousal, valence, and dominance within a unified embedding space. Our framework uses a shared neural architecture with separate output layers for each emotional attribute, trained using a loss function that combines pairwise differences across these dimensions. We evaluate the effectiveness of our model on the MSP-Podcast dataset and conduct cross-domain experiments on the MSP-IMPROV and BiIC-Podcast corpora. Our results demonstrate significant performance gains over traditional pairwise preference learning methods. Importantly, the multitask framework generalizes better across domains, validating the advantage of jointly learning interdependent emotional attributes.

Index Terms—Speech emotion recognition, multi-task learning, preference learning

I. INTRODUCTION

Speech emotion recognition (SER) plays a fundamental role in human-computer interaction, facilitating emotionally intelligent systems in diverse applications such as healthcare, education, automotive industries, and customer service [1]–[3]. Traditionally, SER systems have adopted either categorical labels, which assign discrete emotional classes such as happiness or sadness [4], or dimensional labels that represent emotions as continuous values on scales such as arousal, valence, and dominance [5]–[10]. While continuous attribute prediction is widely adopted, psychological evidence indicates that humans are more adept at making relative or comparative judgments rather than assigning absolute values to subjective phenomena [11]–[13]. Preference learning leverages this inherent human capability by focusing on the ordinal nature of emotions, providing more reliable labels and improved robustness in emotional modeling [14]–[19]. Despite its benefits, most SER preference-learning frameworks have traditionally modeled

each emotional attribute independently, overlooking potential correlations among emotional dimensions. Parthasarathy and Busso [20] demonstrated in emotion regression tasks using absolute scores that multi-task learning across attributes was effective. We also expect this observation to be true for preference learning, where the comparisons across attributes and samples can lead to more robust models.

However, training independent models for each emotional attribute comes with inherent limitations. Firstly, separate models disregard the intrinsic correlation and interactions among emotional dimensions, limiting their ability to capture nuanced emotional states accurately [20]. Secondly, the independent training approach inevitably increases computational complexity, as multiple models need to be trained and maintained separately. Additionally, this disjoint training methodology may introduce inconsistencies or conflicts across models due to differences in training data distributions. By simultaneously modeling all three emotional attributes of arousal, valence, and dominance into a single, unified embedding space, our proposed method enables the model to inherently understand the relative position of each speech sample within the emotional attribute space. This unified ordinal embedding approach leverages inter-attribute relationships, enhancing the model’s capability to retrieve speech files situated within specific regions of the emotional attribute space more effectively than attribute-specific models.

Building on these insights, we propose the novel *multi-dimensional ordinal embedding* (MOE) formulation, inspired by RankNet [21]. Our proposed method employs a shared neural architecture with separate output layers dedicated to each emotional attribute, enabling it to explicitly learn subtle ordinal differences across arousal, valence, and dominance within a unified embedding space. A key part of the formulation is the definition of the relative labels. We construct preference labels based on consensus scores across all attributes, ensuring a meaningful margin of difference in emotional ratings. Specifically, we select pairs of speech samples with consensus ratings differing by at least a predefined margin across all three emotional attributes simultaneously (e.g., *sample 1* is more positive, and dominant than *sample 2* but less active). This approach ensures that the training set is robust and reflects significant emotional differences while also including samples

This study was funded by the NSF under grant CNS-2016719.

with subtle emotional differences, facilitating effective learning. The embedding function is trained using a loss formulation that integrates the ordinal preferences across attributes through a joint optimization procedure, efficiently capturing the subtle differences and inter-dependencies between arousal, valence, and dominance.

Our experiments, conducted on the MSP-Podcast dataset, validate the efficacy of the proposed multi-dimensional ordinal embedding approach. Compared to traditional attribute-specific pairwise preference learning methods, MOE consistently delivered statistically significant improvements across all three emotional dimensions, confirming its superior ability to leverage shared attribute representations. Furthermore, cross-domain evaluations on the MSP-IMPROV and BIIC-Podcast datasets indicated notable generalization improvements over the baseline pairwise preference learning methods, highlighting the robustness of our unified ordinal embedding framework. Additionally, our analysis revealed improved precision in retrieving speech samples with specific emotional attributes compared to attribute-specific preference and absolute attribute prediction models. We further performed extensive ablation studies to underscore the effectiveness of the integrated ordinal embedding strategy, clearly demonstrating the advantage of jointly learning multiple emotional attributes.

II. RELATED WORK

Preference learning has emerged as a powerful approach within SER, leveraging the natural human ability to discern relative emotional intensities rather than absolute values [11], [12]. Traditional approaches for emotional annotation involve assigning absolute emotional ratings, which can lead to inconsistencies due to subjective interpretations and anchoring biases based on previously encountered stimuli [22]–[24]. Preference learning methods circumvent these issues by using relative emotional comparisons, making the annotations more reliable and robust. Since emotional databases are often annotated with absolute scores, a key challenge is forming the relative labels between samples needed for preference learning formulations. For categorical emotions, Cao et al. [18], [25] proposed ranking methods using RankSVM, establishing clear preferences among different emotional classes. For example, a happy ranker will consider a sentence with label *happiness* to be preferred over another sample with a different emotion. Lotfian and Busso [19] further expanded this paradigm by employing inter-evaluator agreements and intra-class confusions, effectively capturing the annotators’ consensus and disagreement to enhance the reliability of preference-based labels. Other strategies to form relative labels for emotional attributes are using trends across multiple annotators for two sentences [26], exploring anchoring on consecutive annotations [27], and creating margin over the consensus attribute values [28]. Additionally, Han et al. [16] proposed *consistent rank logits* (CORAL) for ordinal classification within SER tasks, effectively capturing the ordinal relationships within individual emotional attributes. While these methods have demonstrated promising results, they primarily focus on single emotional

dimensions independently, leaving room for exploring multi-dimensional preference learning.

Multi-task learning (MTL) aims to jointly model multiple related tasks, leveraging shared representations to enhance performance and generalization. Recent research has shown that emotional attributes such as arousal, valence, and dominance are interconnected and can benefit from joint learning strategies [5]–[7]. Parthasarathy and Busso [20] showed that jointly modeling these emotional attributes through shared neural network layers significantly improves the predictive accuracy and robustness compared to single-task regression models to predict absolute scores. Moreover, multi-task approaches often reveal intrinsic relationships among attributes, helping the models better capture the nuanced emotional states [5], [6]. Despite the clear advantages demonstrated in regression-based MTL approaches, integrating preference learning with MTL for SER remains relatively unexplored. Addressing this gap, our work introduces a multi-task preference learning framework, extending the benefits of preference-based modeling across multiple emotional dimensions simultaneously.

III. METHODOLOGY

The goal of our approach is to effectively model the emotional attributes of speech by jointly considering their ordinal nature. Previous work primarily addressed each attribute independently, limiting the ability to exploit the inherent connections among emotional dimensions. To address this limitation, we propose a novel multi-task learning framework inspired by the RankNet formulation, enabling simultaneous ordinal modeling of arousal, valence, and dominance. The cost function jointly considers the trends across the emotional attributes, leading to clear performance improvement as demonstrated in Section V.

A. RankNet Formulation

RankNet was introduced by Burges [21]. It is a well-established method for pairwise preference learning, typically applied through gradient descent optimization. Given two samples (x_i, x_j) with corresponding feature vectors (Φ_i, Φ_j) , the model maps these vectors to preference scores using a function $f(\cdot)$, as illustrated in Figure 1(a):

$$s_i = f(\Phi_i), \quad s_j = f(\Phi_j) \quad (1)$$

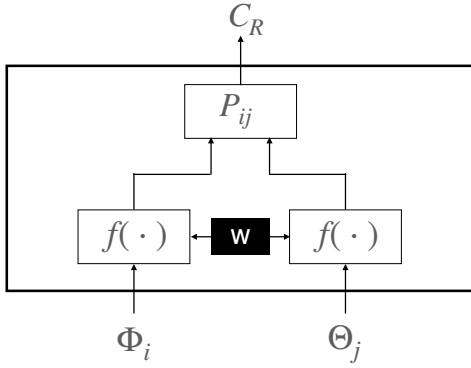
The probability of preferring sample x_i over x_j is represented using a sigmoid function:

$$P_{ij} = \frac{1}{1 + e^{-\sigma(s_i - s_j)}} \quad (2)$$

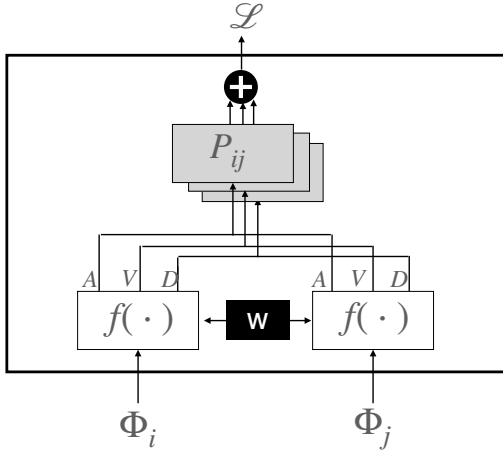
The RankNet cost function ($\mathcal{C}_{\mathcal{R}}$) optimizes the model by minimizing the cross-entropy loss between the predicted probability P_{ij} and the expected preference \bar{P}_{ij} , where $\bar{P}_{ij} = 1$ if x_i is preferred and $\bar{P}_{ij} = 0$ otherwise:

$$\mathcal{C}_{\mathcal{R}} = -\bar{P}_{ij} \log P_{ij} - (1 - \bar{P}_{ij}) \log(1 - P_{ij}) \quad (3)$$

This cost simplifies to $\mathcal{C}_{\mathcal{R}} = \log(1 + e^{-\sigma(s_i - s_j)})$, when $\bar{P}_{ij} = 1$, and $\mathcal{C}_{\mathcal{R}} = \log(1 + e^{-\sigma(s_j - s_i)})$ when $\bar{P}_{ij} = 0$.



(a) RankNet



(b) Proposed MOE Approach

Fig. 1: (a) Block diagram of the RankNet-based preference learning for SER. (b) Block diagram of the proposed Multi-dimensional Ordinal Embedding strategy for SER. A, V, D represent arousal, valence, and dominance components.

B. Proposed Multi-dimensional Ordinal Embedding (MOE)

Building on RankNet, our proposed Multi-dimensional Ordinal Embedding framework simultaneously models multiple emotional dimensions (arousal, valence, and dominance) as shown in Figure 1(b). For each pair of speech samples (x_i, x_j) with feature vectors (Φ_i, Φ_j) , our model provides separate preference scores for each emotional attribute $(s_i^{\text{aro}}, s_i^{\text{val}}, s_i^{\text{dom}}$ and $s_j^{\text{aro}}, s_j^{\text{val}}, s_j^{\text{dom}}$) using the function $f(\cdot)$:

$$\begin{aligned} [s_i^{\text{aro}}, s_i^{\text{val}}, s_i^{\text{dom}}] &= f(\Phi_i) \\ [s_j^{\text{aro}}, s_j^{\text{val}}, s_j^{\text{dom}}] &= f(\Phi_j) \end{aligned}$$

For each emotional attribute, we compute pairwise score differences as shown as the A, V, D components in Figure 1(b):

$$\begin{aligned} O_{\text{aro}} &= s_i^{\text{aro}} - s_j^{\text{aro}} \\ O_{\text{val}} &= s_i^{\text{val}} - s_j^{\text{val}} \\ O_{\text{dom}} &= s_i^{\text{dom}} - s_j^{\text{dom}} \end{aligned}$$

Following the RankNet-inspired pairwise loss formulation, we define individual loss components for each emotional dimension:

$$\mathcal{L}(\text{aro/val/dom}) = \log \left(1 + e^{-\sigma O(\text{aro/val/dom})} \right) \quad (4)$$

We can consider two formulations for the summation shown in Figure 1(b), combining three attribute-specific RankNet loss components. A simple formulation to train the proposed approach is to consider a linear combination of the individual loss components for arousal, valence, and dominance, without any log approximation. This strategy results in the following total loss function:

$$\mathcal{L}_{\text{Linear}} = \mathcal{L}_{\text{aro}} + \mathcal{L}_{\text{val}} + \mathcal{L}_{\text{dom}} \quad (5)$$

The limitation of this strategy is that it treats the emotional dimensions independently, without capturing their joint interactions within a unified optimization objective. To effectively integrate these dimensions into a unified training objective, we combine the individual loss components using a simplified log approximation inspired by the listwise ranking formulation proposed in [29]:

$$\mathcal{L}_{\text{multi}} = \log \left(1 + e^{-\sigma O_{\text{aro}}} + e^{-\sigma O_{\text{val}}} + e^{-\sigma O_{\text{dom}}} \right) \quad (6)$$

This joint loss function allows the model to exploit shared representations, capturing the interdependencies among emotional dimensions. Our approach improves robustness and generalization, effectively leveraging complementary emotional information during training.

C. Preference Learning Labels

To establish preference labels for training our proposed multi-dimensional ordinal embedding model, we adopted a consensus-based strategy inspired by the works of Lotfian and Busso [28]. Our methodology involves determining relative emotional preferences between pairs of speech samples, focusing simultaneously on three emotional attributes: arousal, valence, and dominance. For each pair of samples (x_i, x_j) , we first calculate consensus scores by averaging evaluator ratings across each emotional dimension, denoted as $\hat{e}_{\text{aro}}^{x_i}, \hat{e}_{\text{val}}^{x_i}, \hat{e}_{\text{dom}}^{x_i}$ and $\hat{e}_{\text{aro}}^{x_j}, \hat{e}_{\text{val}}^{x_j}, \hat{e}_{\text{dom}}^{x_j}$, respectively.

Preference labels are defined based on the absolute differences between these consensus scores for each dimension, ensuring that a pair is only considered if the absolute differences across all three dimensions exceed a predefined margin m . Formally, the preference between two samples is established as follows:

$$|\hat{e}_{(\text{aro/val/dom})}^{x_i} - \hat{e}_{(\text{aro/val/dom})}^{x_j}| \geq m \quad (7)$$

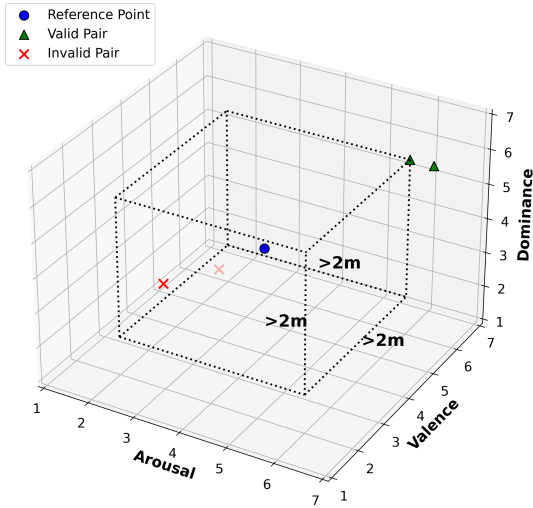


Fig. 2: Illustration of valid preference pair selection where differences in arousal, valence, and dominance all exceed margin m , as indicated by dotted cube. The reference point (blue circle) is at the center of the cube. Samples inside the cube (red crosses) are not valid.

TABLE I: Number of eligible sample pairs in the MSP-Podcast training set that satisfy the margin m condition simultaneously across arousal, valence, and dominance.

Margin m	0.5	1.0	1.5	2.0
Number of Pairs	2.33B	941.6M	249.1M	74.1M

Only pairs simultaneously satisfying these criteria for arousal, valence, and dominance are retained. Each retained pair is then assigned three binary labels (1 or 0), indicating the direction of preference for each attribute (1 if x_i is preferred over x_j , 0 otherwise). We adopted the margin $m = 1.5$ when evaluator ratings are scaled from 1 to 7. This rigorous filtering ensures that our training set comprises pairs with substantial differences, enhancing the robustness and reliability of our preference-based training process. An illustration of this selection strategy is provided in Figure 2, where each dimension satisfies the minimum difference criterion. The dotted projections emphasize that the valid pair lies beyond the margin m in all three emotional directions.

Table I summarizes the number of valid preference pairs as a function of m obtained using this filtering strategy. The table uses the MSP-Podcast dataset’s training split, which contains 112,712 segments (see Sec. IV-A). The counts reflect the total number of sample pairs where the emotional differences for arousal, valence, and dominance simultaneously exceed the defined margin, ensuring clear and meaningful preference relationships. This table highlights that the selection strategy produces more than enough training samples even with a relatively high margin such as $m = 2$ on a Likert scale of 1 to 7. This table demonstrates the practicality of our margin-based filtering approach in constructing a high-quality set of

emotionally distinguishable training pairs.

IV. EXPERIMENTAL SETUP

A. Emotional Databases

We use release 1.12 of the MSP-Podcast corpus [30], a publicly available dataset consisting of over 324 hours of speech data annotated for emotional attributes along with primary and secondary emotions. The audio samples originate from diverse platforms under Creative Commons licenses, covering a wide range of topics, including science, politics, entertainment, finance, and art. Strict quality control measures were applied to filter out segments containing background noise, music, or overlapping speech. Each segment was annotated by at least five annotators across the emotional dimensions of arousal, valence, and dominance, capturing these emotional attributes ranging from calm to active, negative to positive, and weak to strong, respectively. Although primary and secondary emotional categories are also available, they were not considered in this study. The dataset is divided into train (112,712 segments), development (31,961 segments), and test (44,395 segments) sets, aiming for speaker independence across partitions. Relative preference pairs were generated strictly within these sets for all the experiments.

Additionally, we employ the MSP-IMPROV dataset [31] as one of our cross-domain test datasets. The MSP-IMPROV corpus consists of dyadic interactions involving 12 actors, comprising 8,438 speaking turns annotated by at least five annotators for arousal, valence, and dominance. The recordings were made in a controlled environment, providing an ideal scenario for examining domain mismatch. We used data from six actors from the first three sessions as the test set, reserving the remaining sessions for training or adaptation purposes. Furthermore, we use the BIIC-Podcast corpus [32], which consists of 157 hours of speech samples collected from Taiwanese Mandarin podcasts, following a similar annotation methodology as the MSP-Podcast corpus. The BIIC-Podcast corpus also includes annotations for primary and secondary emotional categories alongside the three emotional attributes and serves as an additional cross-domain target dataset in another language.

B. Implementation Details

For our experiments, we extracted feature representations using the WavLM-Large model [33], a powerful transformer-based architecture optimized for speech tasks. We utilized the pre-trained WavLM-Large model from HuggingFace [34], and fine-tuned the entire model (without pruning any transformer layers) using a downstream task designed explicitly for speech emotion recognition. The fine-tuning process employed the MSP-Podcast corpus, optimizing with the Adam optimizer [35] using a learning rate of $1e-5$ over 10 epochs. After fine-tuning, we froze all parameters of the WavLM-Large model and used the resulting 1,024-dimensional output vector as our feature representation. The downstream architecture that follows the frozen WavLM encoder consists of a fully connected feed-forward neural network with three layers. The

first two layers contain 1,024 and 512 neurons, respectively, followed by an output layer with 3 neurons corresponding to the arousal, valence, and dominance dimensions. This network serves as the scoring function $f(\cdot)$ used in the proposed multi-dimensional ordinal embedding framework (Fig. 1(b)).

We utilized two high-performance resources for training and testing the models. We performed fine-tuning and model training on an EC2 g5.4xlarge instance equipped with an NVIDIA A10G GPU, while other experiments were conducted on an NVIDIA GeForce RTX 3090 GPU. We employed the Adam optimizer [35] with a learning rate of $1e-4$, ensuring efficient model convergence.

C. Metrics

Below we describe the metrics used for evaluating the performance of the proposed method in all experiments:

- **Kendall’s Tau (KT) coefficient:** KT is a non-parametric rank correlation metric that measures the ordinal association between two ranked variables. It is commonly used in preference learning to assess how well the predicted rankings align with the ground-truth orderings. To compute the KT metric, we randomly sampled 200 test samples at a time, repeated this process 20 times, and reported the average performance across these trials.
- **Precision at $K\%$:** Precision at $K\%$ indicates the percentage of correctly retrieved samples that match the target emotional attribute within the top $K\%$ of the samples ranked by the model. Specifically, Precision at 10% represents the proportion of true positives in the top 10% of the retrieved results, measuring the retrieval effectiveness of emotionally relevant samples.

D. Baselines

We compare the proposed multi-dimensional ordinal embedding (MOE) approach against three baseline methods:

- **Absolute Attribute Prediction (ABS):** This baseline involves training a regression model using the same architecture as the preference learning network to directly predict the absolute emotional attribute scores. The predicted scores are then used to rank the speech samples, and the model’s performance is assessed using precision-based retrieval metrics. This method provides a contrast between preference-based and traditional score-based modeling for emotional attributes.
- **Pairwise Preference Learning (Pairwise):** This baseline trains separate models for each emotional attribute (arousal, valence, and dominance) using standard pairwise preference learning. Each model receives training pairs labeled with relative preferences for one specific attribute. The performance of this baseline is evaluated independently for each emotional dimension.
- **Multi-dimensional Ordinal Embedding with Linear Loss (MOE_{Linear}):** This is a simplified version of our proposed MOE framework, where the final loss is a direct linear summation of the attribute-specific RankNet losses for arousal, valence, and dominance. Unlike the

TABLE II: *Kendall’s Tau* (KT) coefficient comparison for arousal, valence, and dominance using different preference learning methods on the MSP-Podcast dataset. MOE refers to our proposed Multi-dimensional Ordinal Embedding approach. MOE_{Linear} is a variant using a linear loss combination instead of the log-approximation defined in Eq. (6). The symbols ^{*}†‡, ^{*}† and/or ^{*} indicate statistically significant improvement over the MOE_{Linear}, Pairwise, and ABS methods, respectively.

Method	Arousal	Valence	Dominance
ABS	0.481	0.492	0.361
Pairwise	0.521 [*]	0.508 [*]	0.412 [*]
MOE _{Linear}	0.538 ^{*†}	0.527 ^{*†}	0.418 [*]
MOE	0.553^{*†‡}	0.548^{*†‡}	0.433^{*†}

main MOE formulation that uses a log-based joint loss approximation (Equation 6), MOE_{Linear} treats each dimension equally and adds their losses directly (Equation 5). This baseline helps us evaluate the effectiveness of our joint optimization strategy in modeling shared emotional representations.

V. RESULTS & DISCUSSIONS

Using the preference label extraction method described in Section III-C, we obtained approximately 200,000 pairs of samples, each with simultaneous preference labels across arousal, valence, and dominance. We compared our *multi-dimensional ordinal embedding* (MOE) approach against a total of three baseline methods: (1) traditional pairwise preference learning, which separately models each emotional attribute using 200,000 sample pairs per attribute, (2) a regression-based absolute attribute prediction (ABS) framework that directly predicts the continuous attribute scores, and (3) a variant of our model, termed MOE_{Linear}, which uses a linear combination of the individual loss components without the log approximation, as presented in Equation 5. Table II summarizes the comparison among these approaches using *Kendall’s Tau* (KT) coefficient as the evaluation metric, which measures the ordinal correlation between predicted and ground-truth rankings. We employ a one-tailed t-test to assess statistical significance, considering improvements significant at a p -value less than 0.05. A result with ^{*}†‡, ^{*}† and/or ^{*} indicate that the method is statistically significantly better than the MOE_{Linear}, Pairwise, and ABS methods, respectively.

Results shown in Table II highlight that our proposed MOE method significantly outperforms all baseline methods, demonstrating the effectiveness of simultaneously modeling multiple emotional dimensions within a unified framework. Specifically, the MOE approach achieved relative improvements of approximately 6.1%, 7.9%, and 5.1% for arousal, valence, and dominance, respectively, compared to the traditional pairwise preference learning. The MOE_{Linear} variant, which uses a linear combination of attribute-wise losses, performed better than both the ABS and pairwise baselines, but still underperformed compared to the full MOE model. These results emphasize the advantage of the proposed log-approximation formulation from Equation 6, which better captures the interdependencies

TABLE III: Kendall’s Tau (KT) coefficient for cross-domain evaluation on MSP-IMPROV and BIIC-Podcast datasets for arousal, valence, and dominance attributes. Each row block compares multiple methods including Pairwise, ABS, MOE_{Linear}, and the proposed MOE method. The symbols *†‡, *† and/or * indicate statistically significant improvement over the MOE_{Linear}, Pairwise, and ABS methods, respectively.

Method	Arousal	Valence	Dominance
MSP-IMPROV Corpus			
ABS	0.502	0.448	0.426
Pairwise	0.542*	0.551*	0.446*
MOE _{Linear}	0.551*†	0.569*†	0.447*
MOE	0.567*†‡	0.582*†‡	0.454*†
BIIC-Podcast Corpus			
ABS	0.409	0.304	0.257
Pairwise	0.431*	0.321*	0.296*
MOE _{Linear}	0.445*†	0.332*†	0.291*
MOE	0.468*†‡	0.349*†‡	0.292*

among emotional attributes. Moreover, the MOE_{Linear} approach incurs approximately 71% more training time than the MOE model, further underscoring the efficiency of our proposed training strategy. This computational efficiency stems from the log-approximation formulation in MOE, which simplifies gradient computations by avoiding the need to backpropagate through separate loss branches for each attribute.

To evaluate the generalization of the proposed MOE method, we conducted cross-domain experiments on the MSP-IMPROV and BIIC-Podcast datasets. The models are trained with the MSP-Podcast corpus, evaluating the models in the two other databases without any adaptation. As shown in Table III, MOE consistently outperformed all other baselines, including ABS, Pairwise, and MOE_{Linear}, across most emotional attributes and domains. On the MSP-IMPROV corpus, MOE achieved statistically significant improvements across all three emotional dimensions, with gains particularly notable in valence. For the BIIC-Podcast corpus, MOE showed strong improvements in arousal and valence, while performing comparably to other methods on dominance. These results demonstrate that the shared emotional representations learned by MOE during training on the MSP-Podcast corpus transfer effectively to unseen domains. This observation reinforces our hypothesis that learning a unified, multi-attribute ordinal embedding leads to more robust and generalizable emotion recognition systems.

We also want to evaluate our strategy in emotional speech retrieval tasks. For this purpose, we estimate the precision at 10%. The precision results presented in Figure 3 compare the retrieval performance of four different approaches: absolute attribute prediction (ABS), attribute-specific pairwise preference learning (Pairwise), the linear variant of our proposed method (MOE_{Linear}), and the full multi-dimensional ordinal embedding approach (MOE). Across all emotional attributes (arousal, valence, and dominance), MOE consistently achieves the highest precision, clearly outperforming MOE_{Linear}, Pairwise, and ABS methods. Specifically, the improvement of MOE over Pairwise and MOE_{Linear} highlights the ben-

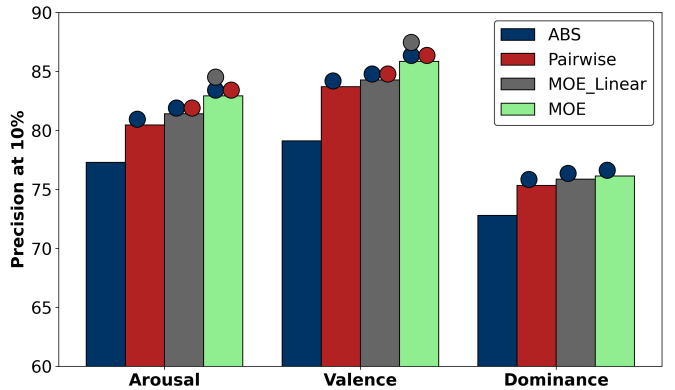
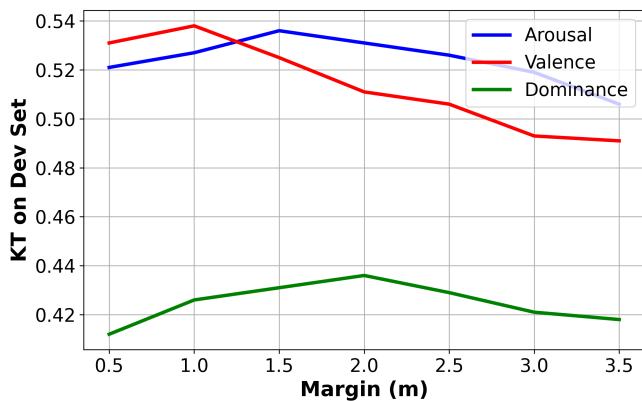


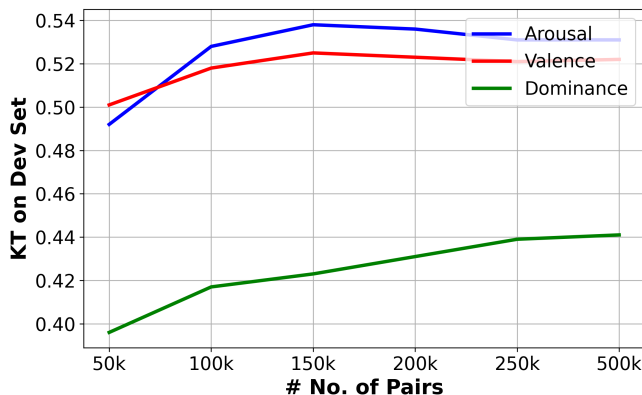
Fig. 3: Precision at 10% comparison for absolute attribute prediction (ABS), pairwise preference learning (Pairwise) with attribute-specific training, the linear variant of multi-dimensional ordinal embedding (MOE_{Linear}), and the proposed multi-dimensional ordinal embedding approach (MOE). A colored circle on top of a bar indicates that the performance of the corresponding method is statistically significant relative to the method denoted by the circle’s color.

efits of jointly modeling ordinal relationships across multiple emotional attributes in a non-linear space, rather than relying on independent or linear formulations. Furthermore, the notable performance gap between MOE-based methods and ABS confirms that preference-based ordinal modeling is more effective than directly predicting absolute scores, leading to more accurate and reliable retrieval of emotionally relevant speech samples.

The ablation results illustrated in Figure 4(a) demonstrate how varying the margin (m) during preference label extraction affects model performance on the development set of the MSP-Podcast. We use Kendall’s Tau (KT) as the performance metric for this evaluation. The plot highlights that the optimal margin for achieving peak performance varies across emotional attributes. Specifically, arousal and valence achieve their highest KT scores at relatively lower margins ($m = 1.5$ and $m = 1.0$, respectively), while dominance reaches its optimal performance at a higher margin ($m = 2.0$). These results indicate that using smaller margins for dominance labels might introduce more annotation noise, negatively impacting the reliability of the preference pairs. Therefore, careful margin selection tailored to each emotional attribute is crucial to reducing noisy labels and optimizing model accuracy. Figure 4(b) presents the impact of the number of training pairs on the KT performance on the development set. The results clearly show attribute-specific differences in performance trends. Valence achieves near-optimal KT performance with fewer training pairs (around 150k), highlighting that preferences in valence might be easier for the model to capture efficiently with less training data. Conversely, dominance continues improving steadily and achieves its best performance with a larger number of training pairs (500k). This result suggests dominance is more subtle or complex and requires more



(a) Effect of margin on KT score.



(b) Effect of number of training pairs on KT score.

Fig. 4: Ablation studies on development set performance.

data for effective modeling. Arousal shows rapid performance improvement initially and reaches its optimal performance at approximately 150k pairs, after which additional training pairs do not significantly benefit performance. These insights emphasize the importance of attribute-specific training pair selection to ensure efficient use of available data.

VI. CONCLUSIONS

This study introduced a novel Multi-dimensional Ordinal Embedding (MOE) framework for speech emotion recognition, explicitly capturing subtle ordinal differences across the emotional attributes of arousal, valence, and dominance within a unified embedding space. Unlike conventional preference learning approaches that model each attribute independently, the proposed method leverages a joint preference formulation to learn shared representations across attributes, significantly enhancing the model’s ability to capture complex emotional dynamics. Evaluations on the MSP-Podcast dataset demonstrated statistically significant improvements across all emotional attributes compared to traditional attribute-specific pairwise learning methods. Cross-domain evaluations on the MSP-IMPROV and BIIC-Podcast corpora further highlighted the robustness and generalization capabilities of MOE, indicating that jointly modeling emotional dimensions facilitates better

transfer to unseen domains. Ablation studies provided deeper insights, emphasizing the importance of carefully selecting margins and training pair quantities tailored to each emotional attribute. The findings also showed superior precision performance for the ordinal embedding approach compared to attribute-specific and absolute-value prediction methods. Collectively, our results validate that jointly modeling emotional attributes through multi-dimensional ordinal embedding significantly improves both predictive performance and cross-domain robustness. Future work will explore extending the MOE framework to incorporate multi-modal cues, enabling richer modeling of emotional context in emotion recognition.

VII. ETHICAL IMPACT STATEMENT

This work focuses on modeling human emotions using speech signals through a preference learning framework. While the proposed method seeks to improve the robustness and generalizability of emotion recognition systems, we recognize the potential ethical implications associated with deploying models that interpret or react to human affect. Emotional analysis technologies may be used to personalize user experiences and increase accessibility in positive ways, but they can also pose risks of manipulation, surveillance, or emotional profiling, particularly without user consent.

Our methodology relies on publicly available datasets such as MSP-Podcast, MSP-IMPROV, and BIIC-Podcast, which are collected under ethical protocols, and we do not have access to personally identifiable information or annotator identities. Furthermore, while our approach could eventually enable more adaptive and emotionally aware AI systems, we emphasize that any real-world deployment of such systems should be governed by strict ethical guidelines regarding transparency, fairness, and user privacy. We advocate for responsible use of affective technologies and encourage future work to explicitly assess and mitigate any biases or unintended consequences arising from emotion modeling.

REFERENCES

- [1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, “Emotion recognition in human-computer interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, January 2001.
- [2] R. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.
- [3] Lucas Goncalves, Ali N. Salman, Abinay Reddy Naini, Laureano Moro-Velázquez, Thomas Thebaud, Paola Garcia, Najim Dehak, Berrak Sisman, and Carlos Busso, “Odyssey 2024 - speech emotion recognition challenge: Dataset, baseline framework, and results,” in *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024, pp. 247–254.
- [4] R. Lotfian and C. Busso, “Formulating emotion perception as a probabilistic model with application to categorical emotion classification,” in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 415–420.
- [5] M. Abdelwahab and C. Busso, “Study of dense network approaches for speech emotion recognition,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2018)*, Calgary, AB, Canada, April 2018, pp. 5084–5088, IEEE.

- [6] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B.W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10745–10759, September 2023.
- [7] M. Abdelwahab and C. Busso, "Active learning for speech emotion recognition using deep neural network," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2019)*, Cambridge, UK, September 2019, pp. 441–447.
- [8] J.R.J. Fontaine, K.R. Scherer, E.B. Roesch, and P.C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.
- [9] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1215–1227, April-June 2023.
- [10] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Not all features are equal: Selection of robust features for speech emotion recognition in noisy environments," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022)*, Singapore, May 2022, pp. 6447–6451.
- [11] 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, January-March 2021.
- [12] G.N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 248–255.
- [13] S. Parthasarathy and C. Busso, "Predicting emotionally salient regions using qualitative agreement of deep neural network regressors," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 402–416, April-June 2021.
- [14] 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. 2, pp. 314–326, July-September 2014.
- [15] S. Parthasarathy, R. Lotfian, and C. Busso, "Ranking emotional attributes with deep neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 4995–4999.
- [16] W. Han, T. Jiang, Y. Li, B. Schuller, and H. Ruan, "Ordinal learning for emotion recognition in customer service calls," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, Barcelona, Spain, May 2020, pp. 6494–6498.
- [17] Abinay Reddy Naini, Mary A. Kohler, and Carlos Busso, "Unsupervised domain adaptation for preference learning based speech emotion recognition," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.
- [18] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, January 2015.
- [19] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 490–494.
- [20] S. Parthasarathy and C. Busso, "Jointly predicting arousal, valence and dominance with multi-task learning," in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.
- [21] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *International conference on Machine learning (ICML 2005)*, Bonn, Germany, August 2005, pp. 89–96.
- [22] G. N. Yannakakis and H. P. Martinez, "Grounding truth via ordinal annotation," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2015)*, Xi'an, China, September 2015, pp. 574–580.
- [23] L. Martinez-Lucas, A. Salman, S.-G. Leem, S.G. Upadhyay, C.-C. Lee, and C. Busso, "Analyzing the effect of affective priming on emotional annotations," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023, pp. 1–8.
- [24] L. Martinez-Lucas, A.N. Salman, S.-G. Leem, W.-S. Chien, S.G. Upadhyay, C.-C. Lee, and C. Busso, "Affective priming in emotional annotations and its effect on speech emotion recognition," *IEEE Transactions on Affective Computing*, 2025.
- [25] H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 358–361.
- [26] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 252–256.
- [27] A. Reddy Naini, A. Salman, and C. Busso, "Preference learning labels by anchoring on consecutive annotations," in *Interspeech 2023*, Dublin, Ireland, August 2023, pp. 1898–1902.
- [28] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5205–5209.
- [29] A. Reddy Naini, F. Diaz, and C. Busso, "Ranklist – a listwise preference learning framework for predicting subjective preferences," *ArXiv e-prints (arXiv:2508.09826)*, pp. 1–9, August 2025.
- [30] 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, October-December 2019.
- [31] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost, "MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception," *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 67–80, January-March 2017.
- [32] S.G. Upadhyay, W.-S. Chien, B.-H. Su, L. Goncalves, Y.-T. Wu, A.N. Salman, C. Busso, and C.-C. Lee, "An intelligent infrastructure toward large scale naturalistic affective speech corpora collection," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2023)*, Cambridge, MA, USA, September 2023, pp. 1–8.
- [33] A. T. Liu, S.-W. Li, and H.-Y. Lee, "TERA: Self-supervised learning of transformer encoder representation for speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 2351–2366, July 2021.
- [34] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, and Q. Lhoest and A.M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.
- [35] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, May 2015, pp. 1–13.