COMBINING RELATIVE AND ABSOLUTE LEARNING FORMULATIONS TO PREDICT EMOTIONAL ATTRIBUTES FROM SPEECH

Abinay Reddy Naini, Shruthi Subramanium, Seong-Gyun Leem, and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
{abinayreddy.naini, shruthi.subramanium, seong.leem, busso}@utdallas.edu

ABSTRACT

Predicting absolute scores is the most common *speech-emotion recognition* (SER) task when predicting emotional attributes (i.e., valence, arousal, and dominance). However, studies have shown that emotion has an ordinal nature where it is more reliable to establish a preference between speech samples (e.g., one sample is more positive than the other). This paper pursues a novel direction to combine absolute and relative learning formulations for SER. The proposed multitask formulation can simultaneously estimate preference between speech samples and predict their absolute score, providing a flexible tool to analyze emotional content in speech. Both tasks mutually complement each other, allowing the model to outperform SER systems that are exclusively trained to either predict absolute scores or estimate preferences. The multitask weights can be set according to the intended applications, prioritizing one task while slightly compromising the performance of the other task.

Index Terms— Speech emotion recognition, Multi-task learning, Preference learning.

1. INTRODUCTION

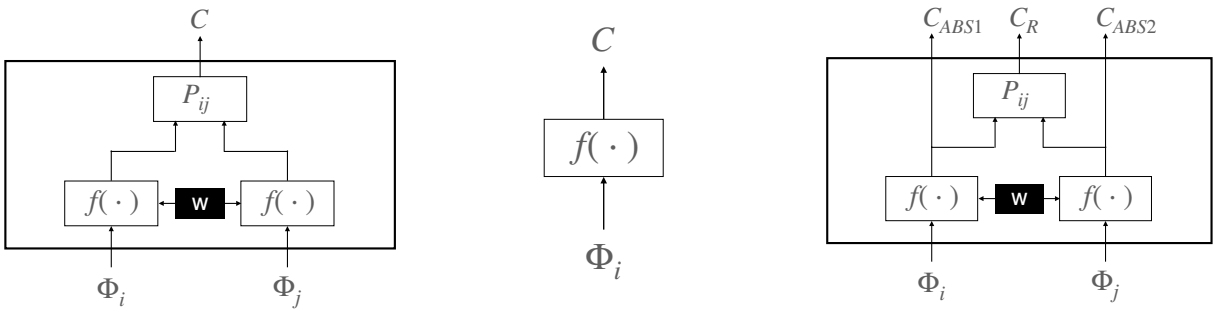
Automatic emotion recognition from speech has many advantages over other modalities given the ubiquitousness of speech-based interfaces, increasing its applicability across many domains including health care, education, gaming, security and defense, and customer service [1–3]. A system that can automatically characterize the emotional content of speech can be an important tool to analyze massive amounts of data coming from multi-media domains. While a typical *speech emotion recognition* (SER) system is often formulated as a recognition problem of emotional classes such as happiness, sadness, and anger [4, 5], a popular alternative is to predict emotional attributes, such as arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) [6–8]. Emotional attributes provide an appealing approach to represent the emotional content [9]. For example, it can provide a more detailed emotional characterization to contrast similar samples that may belong to the same emotional class (e.g., different shades of happiness).

With emotional attributes, the SER problem is commonly formulated as a regression task that aims to predict the value of the target attribute. A less popular approach is the formulation of an SER task as a preference learning problem, where the goal is to determine preference between samples with respect to one emotional attribute (e.g., one sentence is more *positive* than the other) [10–17]. This approach leverages the ordinal nature of emotions [18], where the labels, and therefore the models, are more reliable by capturing relative trends between the emotional content of the samples, rather than their absolute values [19–21]. However, from a practical perspective, it is more desirable for some applications to have a SER system that can predict the absolute value of the emotional attribute, rather than relative comparisons. This study demonstrates that both formulations are not mutually exclusive and that by combining them, we can achieve a flexible model that can predict absolute scores and establish preferences between samples.

This study presents a novel formulation that combines preference learning and regression formulations in a *multitask learning* (MTL) architecture. The SER model simultaneously predicts an emotional attribute score, along with a preference score that can be used to rank-order speech sentences according to a given emotional attribute. Our formulation relies on the RankNet framework [22], which is a popular method for training a preference learning model. The approach creates a feature transformation that is applied to a pair of speech samples to determine preference among them in terms of an emotional attribute. The contribution of this study is to use the feature transformation in RankNet to simultaneously predict the absolute score of the two speech samples. Our approach increases its robustness and performance by combining the regression loss and the preference learning loss. Furthermore, it offers the flexibility to simultaneously predict emotional attributes and establish preferences between speech samples, increasing the range of applications that this approach can be used.

This work was supported by Laboratory for Analytic Sciences (LAS)

979-8-3503-0689-7/23/$31.00 ©2023 IEEE



(a) RankNet framework for preference learning (b) Framework for absolute score (c) Proposed multi-task framework

Fig. 1: Framework considered in this study. (a) RankNet framework for training a preference learning-based model, (b) prediction model to estimate the absolute emotional attribute score of a sample, and (c) proposed multi-task framework formulation to train a robust functional block that can predict both relative preferences and absolute attribute scores.

The experimental evaluation on the MSP-Podcast corpus demonstrates that the SER model trained using the proposed MTL framework can perform better than models trained for either regression or preference learning tasks. The results are consistently observed for arousal, valence, and dominance. The approach also provides a principled way to prioritize one task during the training process, while slightly decreasing the performance of the other task, leading to improved performance for the primary task. In this setting, the secondary task contributes to better regularization, increasing the robustness of our proposed formulation. The novel contribution of the paper is to leverage the use of preference labels in the prediction of absolute scores, which are more robust and consistent than absolute labels, as demonstrated by previous studies on preference learning. To the best of our knowledge, this study is the first attempt to train an SER deep learning model in a multi-task framework using both preference and absolute attribute score labels.

2. RELATED WORK

Yannakakis et al. [18] argued that emotions should be studied and represented in an ordinal manner since we are better at detecting relative emotional trends rather than absolute values. Consciously or unconsciously, we anchor our emotional perception on previous emotional experiences [23]. Therefore, labels generated by annotators who assess the absolute value of a stimulus are less reliable than labels that describe relative emotional trends [24]. These findings have led to several research directions to build more reliable emotional labels to train preference learning methods that aim at rank-order a set of samples according to an emotional dimension. (e.g., happier, angrier, more positive, more active). These approaches involve training preference learning models with labels that represent the relative ranking of pairs of samples, indicating which sample is preferred over the other with regard to the target emotional descriptor.

For categorical emotions, Cao et al. [14, 16] proposed a ranking method using RankSVM. This method establishes preferences between sentences by stipulating that all sentences labeled with a target emotion (e.g., happiness) are preferred over sentences annotated with a different emotion (e.g., anger). Lotfian and Busso [15] introduced a preference learning framework that does not rely on consensus labels. Instead, it utilizes inter-evaluator agreement and intra-class confusion to define preferences between emotions. Han et al. [12] employed a method based on *consistent rank logits* (CORAL) to jointly train multiple ordinal binary SER tasks, aiming to enhance consistency across sub-classification tasks. Most of these approaches have focused on obtaining a reliable preference learning block, which can be used in information retrieval tasks. However, less emphasis has been given to making SER models robust to various SER-related applications.

Cao et al. [16] improved the accuracy of a categorical emotional classification task by combining it with the preference score obtained using RankSVM models. The approach relied on a ranker for each emotional category. In other domains, Kim et al. [25] recently improved the text classification accuracy by using preference scores between pair of text documents as the auxiliary task. In both these approaches, preference labels are used as auxiliary information to improve classification accuracy. However, there is a need to make these approaches robust to preference learning tasks along with the primary classification/regression task.

3. METHODOLOGY

Figure 1(c) shows our proposed framework that builds upon the RankNet framework (Fig. 1(a)). This section describes the proposed MTL formulation for the SER task to simultaneously rank and predict emotional attributes, including the building blocks of the proposed approach.

3.1. The RankNet Framework

A popular machine-learning approach for preference learning is using the RankNet-based implementation. The RankNet algorithm, originally proposed by Burges [22], employs a probabilistic cost function to train a model that distinguishes preference between pairs of data points through gradient descent as shown in Figure 1(a). For two samples (x_i, x_j) with feature vectors Φ_i and Φ_j, RankNet creates a feature representation function $f(\cdot)$ to extract the preference scores: $s_i = f(\Phi_i)$ and $s_j = f(\Phi_j)$. To model the probability of preferring one sample (x_i) over the other (x_j), RankNet uses a sigmoid function, defined as follows:

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

During training, the function $f(\cdot)$ is trained using the preferences between sample pairs as ground truth labels. If sample x_i is preferred over sample x_j, the expected probability $\bar{P_{ij}}$ is set to 1; otherwise, it is set to 0. The cost function $(\mathcal{C}_\mathcal{R})$ used to optimize the parameters of the function $f(\cdot)$ is the cross-entropy between the expected probability $\bar{P_{ij}}$ and the actual probability P_{ij}.

$$\mathcal{C}_\mathcal{R} = -\bar{P_{ij}} \log P_{ij} - (1 - \bar{P_{ij}}) \log(1 - P_{ij}). \tag{2}$$

The loss $\mathcal{C}_\mathcal{R}$ simplifies to $\mathcal{C}_\mathcal{R} = \log(1 + \exp^{-\sigma(s_i - s_j)})$ when $\bar{P_{ij}} = 1$, and $\mathcal{C}_\mathcal{R} = \log(1 + \exp^{-\sigma(s_j - s_i)})$ when $\bar{P_{ij}} = 0$. Notice that the function $f(\cdot)$ can be arbitrarily built, providing flexibility to the framework.

3.2. Proposed Multi-task Framework

Figure 1(c) shows the proposed multi-task formulation for training a model that simultaneously establishes preferences between pairs of speech files and predicts their absolute attribute level. Similar to the RankNet formulation, we consider feature vectors (Φ_i and Φ_j) from a pair of speech samples (x_i, x_j). The proposed approach employs a feature representation function $f(\cdot)$ to extract emotional scores from the corresponding pair of sentences, given by $e_i = f(\Phi_i)$ and $e_j = f(\Phi_j)$. The key idea in the proposed approach is to force the values for e_i and e_j not only to preserve their relative differences (i.e., RankNet formulation) but also to be close to their absolute scores (i.e., regression formulation).

We estimate two losses during training. The first loss is the RankNet loss for preference learning, and the second loss is the *concordance correlation coefficient* (CCC) loss for the regression task. The resulting cost function consists of three components: $\mathcal{C}_\mathcal{R}$, $\mathcal{C}_{ABS1}$, and $\mathcal{C}_{ABS2}$. Similar to RankNet, we model the probability of preferring one sample (x_i) over the other (x_j) by a sigmoid function as follows:

$$P_{ij} = \frac{1}{1 + e^{-\sigma(e_i - e_j)}}. \tag{3}$$

We obtain the RankNet cost component $\mathcal{C}_\mathcal{R}$ from equation 2. For the regression tasks, we obtain the loss components $\mathcal{C}_{ABS1}$ and $\mathcal{C}_{ABS2}$ using CCC. The CCC maximizes the Pearson's correlation between the true and predicted values while minimizing the difference between their means. Hence, the overall loss for the proposed MTL formulation is given by

$$\mathcal{C} = \alpha_{PL}\mathcal{C}_\mathcal{R} + (\alpha_{ABS})(\frac{1}{2}\mathcal{C}_{ABS1} + \frac{1}{2}\mathcal{C}_{ABS2}), \tag{4}$$

where α_{PL}, and α_{ABS} are normalized hyperparameters that are set to achieve good performance in both tasks or to prioritize one of the tasks, as desired. During the training of the function $f(\cdot)$, our approach requires pairs of samples with relative and absolute labels. For the relative labels, our approach needs to know which sample is preferred over the other. The preference score (1 or 0) indicates which sentence is preferred (see Sec. 3.3 for details on the relative labels used in this study). For the absolute labels, our approach needs to know the consensus emotional attribute scores as ground truth for both samples.

During inference, the emotional scores e_i obtained for a test speech sample x_i can be used to rank order the sample among other sentences in the test set with a simple sort operation. The emotional score e_i also provides the absolute prediction for the emotional attribute.

3.3. Ordinal Labels with Qualitative Agreement (QA)

One of the challenges in employing preference-learning formulations is obtaining ground truth for relative labels with preference between samples since most existing datasets are annotated with absolute scores for the emotional attributes. A common approach is to obtain these relative scores by observing trends in the absolute scores. Among these methods, we select the *qualitative agreement* (QA)-based method proposed by Parthasarathy and Busso [20]. This popular approach captures the relative trends across the individual annotations provided to two samples, instead of relying on the consensus labels. Figure 2 illustrates the QA-based method [20]. Consider a database annotated with a Likert scale (1: low, 7: high) for each sentence by several evaluators. The illustration considers two sentences, namely Sentence 1 and Sentence 2, which are annotated by N_1 and N_2 independent annotators, respectively. In the example in Figure 2, Sentence 1 has four raters, and Sentence 2 has five raters. We obtain a matrix of size $N_1 \times N_2$ by comparing the individual annotations between the pair of sentences. This matrix captures the trends between the annotations, represented by the symbols $\uparrow$ (upward trend), when the annotation for Sentence 1 is higher than the annotation for Sentence 2, $\downarrow$ (downward trend) when the annotation for Sentence 1 is lower than the annotation of Sentence 2, and $=$ (equality) when both annotations are equal. These trends are established when the differences in emotional attribute scores provided by the respective raters exceed

	Sentence 1	Sentence 2
Rater 1	3.0	2.0
Rater 2	4.0	2.0
Rater 3	3.0	3.0
Rater 4	5.0	3.0
Rater 5	-	4.0

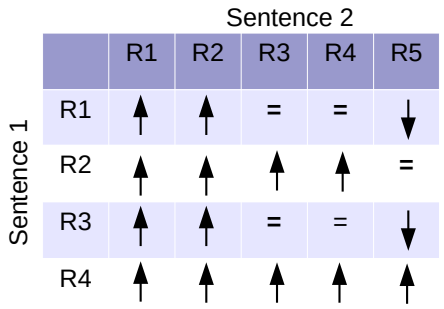


Fig. 2: The figure shows the QA-based approach to obtain relative labels from sentence-level annotations following the strategy proposed in Parthasarathy and Busso [20]. The scores provided to two sentences by multiple raters are compared, establishing trends that are used to establish which sentence is preferred over the other.

a predefined margin. In this study, as depicted in Figure 2, we set this margin to 1. To establish a preference among a pair of sentences, we aggregate the trends in the qualitative matrix. We establish a preference if one sentence is consistently preferred over the other. This decision is made by applying a threshold over the proportion of the trends ($\uparrow$, $\downarrow$ and $=$). We set this threshold to 60% (i.e., 60% of the trends obtained in the qualitative matrix should indicate that one sentence is preferred over the other). For instance, in Figure 1(c), Sentence 1 is preferred over Sentence 2 since 65% of the trends indicate an upward trend (13 $\uparrow$, 2 $\downarrow$, 5 $=$). This approach can be applied to evaluate preferences between each pair of sentences. Only pairs of sentences that satisfy this criterion are used to train the preference learning model.

4. EXPERIMENTAL SETTING

4.1. The MSP-Podcast Corpus

For our study, we utilize release 1.10 of the MSP-Podcast corpus [26], which is a publicly available database comprising more than 166 hours of speech annotated with emotional labels. The corpus was obtained from various audio-sharing websites that offer content under Creative Commons licenses. The recordings encompass a wide range of topics, including science, politics, entertainment, finance, and art. To ensure data quality, all speaking turns in the dataset were carefully filtered to exclude background music, noise, and any instance of overlapped speech. Each turn in the dataset was annotated by a minimum of five annotators for the attributes of arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong). The evaluators rated each attribute with a scale ranging from 1 to 7. The consensus score is obtained by averaging the values assigned by all the raters. Additionally, the recordings were annotated with primary and secondary emotional categories, but these annotations are not utilized in the context of this particular study.

4.2. Feature Extraction

In all our experiments, we extract the Wav2Vec2.0-large-robust model feature representation [27], which is used as the input for the feature representation block. This Wav2Vec2.0-large-robust model uses a *convolutional neural network* (CNN) followed by a transformer-based feature encoder consisting of 24 transformer blocks, pre-trained with diverse speech datasets. Each vector has a receptive field of 20 ms. For the implementation, we used the pre-trained Wav2Vec2.0-large-robust model from the HuggingFace library [28]. Then, we prune the top 12 transformer blocks and fine-tune the model. This strategy was suggested by Wagner et al. [7]. For the downstream head, we considered two fully connected layers of 1024 nodes each with *rectified linear unit* (ReLU) as the activation function. For fine-tuning the model, we use the train set of the MSP-Podcast corpus, relying on the Adam optimizer [29] with a learning rate set to 0.00001 for 10 epochs. We consider the average pooled vector obtained across all frames as the sentence-level representation.

4.3. Baselines

The evaluation of this approach considers two baselines, one for preference learning and one for absolute score predictions. The first baseline is the RankNet method illustrated in Figure 1(a). For constancy, the function $f(\cdot)$ is implemented with exactly the same architecture as the function $f(\cdot)$ in our proposed framework. We refer to this baseline as PL. The second baseline is an absolute attribute score prediction model described in Figure 1(b). Given a feature vector (Φ_i) obtained for a sample (x_i), the function $f(\cdot)$ is trained by optimizing the CCC loss between the predicted value of the attribute and its ground truth consensus label. The function $f(\cdot)$ is also implemented using the same architecture as our approach, for consistency. We refer to this baseline as ABS.

4.4. Implementation

The function $f(\cdot)$ takes the Wav2Vec2.0-large-robust feature representation (Sec. 4.2) as input for the baseline and our proposed MTL framework. The function $f(\cdot)$ is implemented us-

ing two fully connected layers, each consisting of 1024 nodes with a layer normalization along with a dropout of 0.5. For all models, random initialization is applied, and they are trained for 20 epochs using a learning rate of 0.00001. The selection of the best model is based on its performance on the development set of the MSP-Podcast corpus. Once identified, the best model is evaluated on the test set. The implementation of all models is carried out using Tensorflow 2.0, leveraging the computational power of an NVIDIA GeForce RTX 3090 GPU.

We optimize the proposed MTL formulation by varying the weight ratio in the objective function ($\alpha_{PL}, \alpha_{ABS}$) based on the performance on the development set. Based on the intended application, we considered three cases. In the first case (Case-1), we optimize the hyper-parameters α_{PL} and α_{ABS} to achieve the best performance for both tasks (i.e., preference learning and regression tasks). Both tasks are equally important. In the second case (Case-2), we optimize the hyper-parameters α_{PL} and α_{ABS} to achieve the best performance for the preference learning problem (primary task), even if the performance of the regression problem (secondary task) is slightly affected. For the third case (Case-3), we set the hyper-parameters α_{PL} and α_{ABS} to achieve the best performance for the regression problem, setting the preference learning problem as the secondary task. We consider Cases-2 and Cases-3 to observe the maximum performance of the proposed MTL-based model in each independent task when set as the primary problem.

5. EXPERIMENTAL RESULTS

As an approach based on ordinal formulation, the RankNet baseline models (PL) and the proposed multi-task learning model (MTL) are trained to rank the samples in the test set based on an emotional attribute. To assess the performance, we utilize the *Kendall's Tau* (KT) coefficient, which provides an estimation of the ordering provided by a given method. To alleviate the computational load of analyzing all possible pairs in the testing set, we randomly select a subset of 200 samples at a time to evaluate the performance. This process is repeated 20 times. Table 1 reports the average KT results across the 20 testing sets. To determine if the results are statistically significant, we employ a one-tailed t-test, considering significance at a p-value less than 0.05. To assess the proposed MTL formulation and the ABS baseline in predicting emotional attributes, we report the performance using CCC by randomly splitting the test set into 20 subsets of similar size. Then, we conduct a two-tailed t-test over the 20 subsets. We defined statistical significance at a p-value less than 0.05. Table 1 shows the comparison between the proposed MTL model, and the PL and ABS baselines trained using the same functional block $f(\cdot)$.

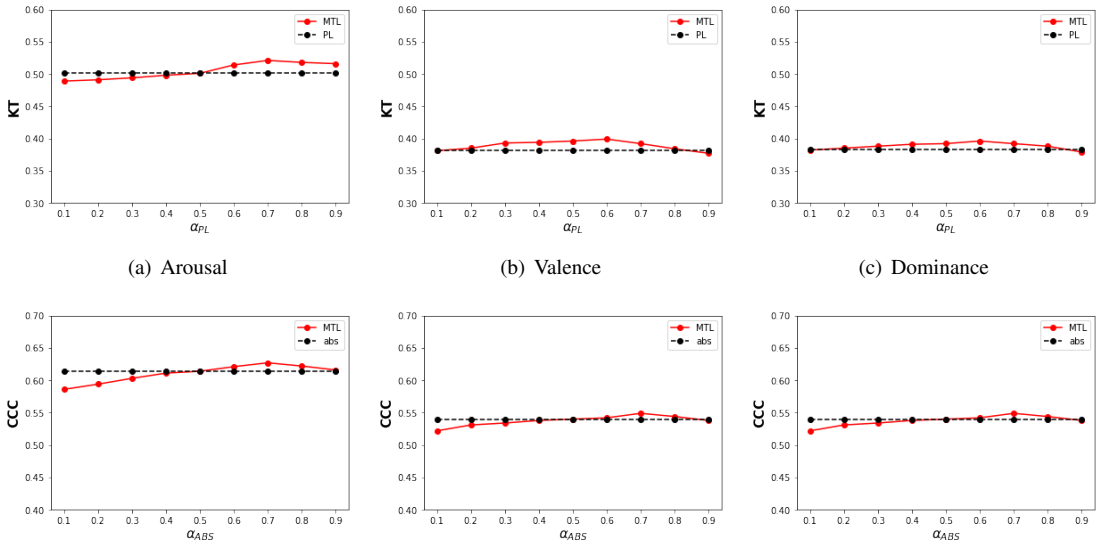
Figure 3 shows the MTL system performance in the development set when the corresponding weights are varied.

Table 1: *Kendall's Tau* (KT) coefficient and *concordance correlation coefficient* (CCC) of the baselines and proposed methods for arousal, valence, and dominance. The table reports results for Case-1, Case-2 and Case-3 (Sec. 4.4). MTL: proposed multi-task learning framework, PL: preference learning framework, ABS: absolute attribute prediction. The symbol * indicates that using the proposed framework leads to significant improvement over the corresponding baseline method.

	Arousal	Valence	Dominance
Case-1 (For MTL task)			
$(\alpha_{PL}, \alpha_{ABS}) =$	(0.45,0.55)	(0.3,0.7)	(0.4,0.6)
MTL [KT]	**0.507**	**0.393***	**0.391***
PL [KT]	0.502	0.381	0.383
MTL [CCC]	**0.619**	**0.546***	**0.542**
ABS [CCC]	0.614	0.531	0.540
Case-2 (For PL task)			
$(\alpha_{PL}, \alpha_{ABS}) =$	(0.7,0.3)	(0.6,0.4)	(0.6,0.4)
MTL [KT]	**0.521***	**0.399***	**0.396***
PL [KT]	0.502	0.381	0.383
Case-3 (For ABS task)			
$(\alpha_{PL}, \alpha_{ABS}) =$	(0.3,0.7)	(0.2,0.8)	(0.3,0.7)
MTL [CCC]	**0.627***	**0.557***	**0.549***
ABS [CCC]	0.614	0.531	0.540

We consider the sum of both weight coefficients (α_{PL}, and α_{ABS}) to equal 1. Hence, a high value for one task weight coefficient results in a lower weight for the other task. As a reference, the Figure also shows a straight dotted black line with the results obtained by the PL and ABS baselines. For all three attributes, we consistently observe the best performance for each task when the corresponding weight is above 0.5 but less than 1. This result indicates that adding the secondary task improves the performance of the primary task in the MTL system, validating the importance of jointly considering preference learning and regression tasks.

Table 1 shows the results for Case-1, Case-2 and Case-3. First, we analyze the results for in Case-1, when both tasks are equally important. Table 1 shows the comparison of the results obtained by the MTL model and both the PL and ABS frameworks, using KT and CCC, respectively. The proposed MTL-based model performs significantly better than the PL-based model for valence, and dominance. The results for the proposed model are slightly better for arousal but the difference is not statistically significant. When compared to the ABS based-model for the regression task, our proposed approach leads to significantly better CCC performance for valence. The results of our proposed approach are slightly better for arousal and dominance. Particularly, the proposed MTL system led to a relative improvement in the valence prediction of $\sim 3.14\%$ (KT) and $\sim 2.82\%$ (CCC) compared to the results obtained by the PL and ABS baselines, respectively.



(a) Arousal (b) Valence (c) Dominance

(d) Arousal (e) Valence (f) Dominance

Fig. 3: Performance in the development set of the proposed *multi-task learning* (MTL) system for different weight values in the objective function (α_{PL}, α_{ABS}). The approach is compared with the results obtained with the results obtained with the preference learning (PL), and absolute attribute prediction (ABS) baselines. The RankNet cost weight is α_{PL}, and the CCC weight is α_{ABS}, with $\alpha_{PL} + \alpha_{ABS} = 1$.

It is interesting to see that the proposed MTL formulation resulted in a model that performs either significantly better or similarly compared to the PL and ABS models.

We also analyze the cases when one of the formulations is the primary task and the other is the secondary task. Table 1 shows the performance of the MTL system when it is optimized to maximize the performance on the preference learning-based task, without focusing on the regression tasks (Case-2). We observe a higher weight given to the RankNet cost component while training could further improve the MTL system performance. The table shows that the optimal value for α_{PL} is between 0.6 and 0.7, depending on the emotional attribute. This setting leads to relative improvements of $\sim$ 3.78% (arousal), $\sim$ 4.72% (valence), and $\sim$ 3.39% (dominance) compared to the PL baselines. All these differences are statistically significant. We observe similar trends when the primary task is the regression task (Case-3). Table 1 shows significant improvements in predicting the absolute attribute scores compared to the ABS baseline, with relative improvements of $\sim$ 2.12% (arousal), $\sim$ 4.89% (valence), and $\sim$ 1.67% (dominance). These results are observed when the weight α_{ABS} is set between 0.7 and 0.8. When we compare the results of our proposed approach for Case-1 with the results for Case-2 and Case-3, we observe that the performance can be increased with relative improvements between 1.28% and 2.76% by adjusting the cost function weights. These results reveal that α_{PL} and α_{ABS} can influence the MTL model performance, and they can be set according to the intended target application.

6. CONCLUSIONS

This study explored a novel multi-task formulation for speech emotion recognition that combines in a unified framework a preference learning and an absolute label prediction formulation. The approach preserves the relative preference between speech samples while predicting the actual emotional attribute score. We observed that this formulation can simultaneously predict absolute attribute scores along with preference labels with performance that are higher than single-task SER systems that are exclusively built to complete one of these tasks. We further showed that the multi-task weights can further improve the performance according to the intended application while slightly compromising the performance on the secondary task. The proposed multi-task formulation provides a flexible and robust SER model that can simultaneously quantify the absolute scores for an attribute and establish preferences between speech samples with respect to a given emotional attribute. This formulation is ideal for practical applications that require retrieval of emotional speech from a large speech repository.

A future research direction is to explore alternative objective functions that will improve the performance of both preference learning and absolute prediction tasks. We will also explore other strategies to estimate relative labels from absolute labels, such as the one proposed by Naini et al. [21].

7. REFERENCES

[1] Y. Lei and H. Cao, "Audio-visual emotion recognition with preference learning based on intended and multimodal perceived labels," *IEEE Transactions on Affective Computing*, vol. Early Access, 2023.

[2] 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.

[3] R. Picard, *Affective Computing*, MIT Press, Cambridge, MA, USA, 1997.

[4] R. Lotfian and C. Busso, "Predicting categorical emotions by jointly learning primary and secondary emotions through multitask learning," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 951–955.

[5] K. W. Gamage, V. Sethu, and E. Ambikairajah, "Salience based lexical features for emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 5830–5834.

[6] 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.

[7] 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. Early Access, 2023.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] A. Reddy Naini, M.A. Kohler, and C. Busso, "Unsupervised domain adaptation for preference learning based speech emotion recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, Rhodes island, Greece, June 2023, pp. 1–5.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[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.

[24] 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.

[25] J. Kim, J. Shin, and D. Kang, "Prefer to classify: Improving text classifiers via auxiliary preference learning," in *International Conference on Machine Learning (ICML 2023)*, Honolulu, HI, USA, July 2023, pp. 1–12.

[26] 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.

[27] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," *ArXiv e-prints (arXiv:2104.01027)*, pp. 1–9, April 2021.

[28] 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 amd A.M. Rush, "HuggingFace's transformers: State-of-the-art natural language processing," *ArXiv e-prints (arXiv:1910.03771v5)*, pp. 1–8, October 2019.

[29] 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.