Preference Learning Labels by Anchoring on Consecutive Annotations

Abinay Reddy Naini, Ali N. Salman and Carlos Busso

Multimodal Signal Processing (MSP) Lab., Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA

AbinayReddy.Naini@utdallas.edu, ali.salman@utdallas.edu, busso@utdallas.edu

Abstract

An important task in human-computer interaction is to rank speech samples according to their expressive content. A preference learning framework is appropriate for obtaining an emotional rank for a set of speech samples. However, obtaining reliable labels for training a preference learning framework is a challenging task. Most existing databases provide sentence-level absolute attribute scores annotated by multiple raters, which have to be transformed to obtain preference labels. Previous studies have shown that evaluators anchor their absolute assessments on previously annotated samples. Hence, this study proposes a novel formulation for obtaining preference learning labels by only considering annotation trends assigned by a rater to consecutive samples within an evaluation session. The experiments show that the use of the proposed anchor-based ordinal labels leads to significantly better performance than models trained using existing alternative labels.

Index Terms: Speech emotion recognition, Preference learning

1. Introduction

Understanding paralinguistic information from speech such as emotional content can facilitate *human-computer interaction* (HCI) to be more attentive to the user's needs [1, 2]. Within the area of *speech emotion recognition* (SER), an emerging approach is the use of preference learning formulations [3–9], leveraging the ordinal nature of emotions [10]. Preference learning is an appealing solution for a speech-emotional retrieval system aiming to rank the given speech samples according to their emotions [11,12]. It has many applications in healthcare, security and defense, education, and entertainment.

Ranking-based SER systems can retrieve speech samples according to either emotional categories such as happiness, sadness, and anger (e.g., is *sample A* happier than *sample B*?) [6, 8, 13], or based on emotional attributes such as arousal, valence, and dominance (e.g., is *sample A* more active than *sample B*?) [3, 4, 14, 15]. In this study, we focus on the emotional attributes of arousal, valence, and dominance. A typical SER system is trained to predict the emotional attribute scores using labels derived from subjective evaluation from multiple annotators [16–21]. It is well-known in cognitive psychology that humans struggle when asked to provide an absolute judgment of a subjective concept such as emotion, leading to low inter-evaluator agreement [22]. Hence, a preference learning strategy is more suitable to train an SER system, relying on trends in the annotations to derive relative labels.

The crucial part of training an SER system using preference learning is to obtain reliable ordinal labels indicating preference between samples. Directly obtaining ordinal labels with subjective evaluations is very difficult due to the number of preferences between pairs of samples ($N(N-1)/2$ for N samples in the database). Hence, existing methods in the literature have obtained preference labels by using the available absolute scores provided by multiple annotators [6, 8, 9, 11, 15]. A simpler way of obtaining relative labels for emotional attributes is by taking the difference in the average of scores assigned to two samples [9], or by leveraging trends across annotations assigned to two different samples [15]. Even though it is possible to obtain an extensive set of training pairs using these two methods, they do not leverage a crucial observation related to how we perceive emotions: we anchor our judgments on previous experiences [10, 12, 23]. In the context of emotional perceptual evaluations, this observation implies that the labels are strongly influenced by the emotional content observed on previous samples annotated by the human raters. This sequential dependency in the labels is ignored by these methods.

This study hypothesizes that considering the preference between annotations assigned by a worker to immediately consecutive sentences during a perceptual evaluation session can lead to more reliable relative labels than considering trends across absolute scores. We propose a novel formulation to obtain ordinal labels from existing attribute scores that leverages the anchoring process by considering preference across consecutive annotations. We only consider a preference label for a pair of samples if a worker annotates two consecutive samples with different scores. We also evaluate cases where multiple evaluators consistently agree on the given preference. Even though the approach creates a sparse set of relative labels, the experimental evaluation on the MSP-Podcast corpus demonstrates that the proposed labels lead to improvements in performance for arousal, valence, and dominance, compared with alternative approaches to obtain relative labels from absolute scores. The results also show that we can achieve good performance for valence using a smaller set of training samples.

2. Background

Yannakakis et al. [10, 12] detailed the benefits of using preference learning models for emotion recognition problems. This study provides findings across different fields suggesting that humans perceive emotion by anchoring their judgment on previous emotional experiences. Helson [23] suggested that people's perceptions of the world are shaped by their past experiences and the degree of change they have undergone. Miller [22] explored the limitations of human memory and processing capacity. The study argued that human perception is constrained by the number of items that can be held in working memory. However, studies have shown that our experiences are based on a previous short memory. We hypothesize that ordinal labels derived from consecutive ratings can lead to more reliable labels.

There are several approaches that have used preference learning for SER. The labels used to train preference learning models correspond to relative labels between pairs of samples, indicating which sample of the two is preferred with respect to a given emotional descriptor (e.g., happier, angrier, more positive, more active). For categorical emotions, Cao et al. [6,8] proposed a ranking method using RankSVM. The preference between sentences was defined by imposing that all sentences labeled with a target emotion (e.g., happiness) were preferred over sentences annotated with a different emotion (e.g., anger). Lotfian and Busso [7] proposed a preference learning framework without relying on consensus labels by using inter-evaluator agreement and intra-class confusion between the emotions. Han et al. [5] used a *consistent rank logits* (CORAL)-based method to jointly train multiple ordinal binary SER tasks for improving consistency across sub-classification tasks. For emotional attributes, Martinez et al. [3] showed that better generalization can be achieved by using a rank-based transformation of emotional attributes than by grouping them into classes. They proposed a preference learning framework using relative labels drawn after selecting samples with lower and higher values of an attribute. Samples above a threshold were considered as the *high* class and samples below a threshold were considered as the *low* class. Then, the preferences were established between samples of the low and high classes. Our study focuses on relative labels for emotional attributes. As baselines, we consider two approaches that have been proposed to transform absolute scores into relative labels: the absolute difference between consensus labels (Sec. 2.1) and the QA-based approach (Sec. 2.2). This section describes these two approaches and the RankNet framework [24] (Sec. 2.3), which is the approach used to evaluate our proposed labels.

2.1. Ordinal Labels using the Consensus Labels (ABS)

The most straightforward approach for establishing preference between samples x_i and x_j using emotional attributes is to compare the difference between the average scores provided for each sample. We define $\hat{e}^{x_i}$ and $\hat{e}^{x_j}$ as the average scores provided to samples x_i and x_j. The approach defines a preference if $|\hat{e}^{x_i} - \hat{e}^{x_j}| \geq m$, where m is a margin set as a hyper-parameter, which avoids defining a preference when the difference is minimal. This approach was used by Lotfian and Busso [9] to build preference models, discussing practical considerations for setting this margin. They showed that a margin of $m = 2.4$ (when rescaled to the range from 1 to 7 used in this study) led to the best preference learning labels. Hence, we adopt this margin to implement this approach. We refer to this approach as the ABS labels.

2.2. Ordinal labels with Qualitative Agreement (QA)

Instead of relying on consensus labels, Parthasarathy and Busso [15,25] proposed strategies based on the *qualitative agreement* (QA) method [26] aiming to create more reliable relative labels by identifying trends across evaluators. The approach captures the relative trends across the individual annotations provided to two samples, instead of relying on the consensus label. Figure 1(a) illustrates the QA-based method [15] based on the Likert scale (1: low, 7: high) scores for each sentence. Consider two sentences (sentence 1, sentence 2) annotated by N_1 and N_2 independent annotators. Then, a matrix of size $N_1 \times N_2$ is obtained by comparing all individual annotations between the pair of sentences. The matrix contains trends between the annotations, indicated by $\uparrow, \downarrow$, or equal $(=)$. The symbols $\uparrow$ and $\downarrow$ indicate if the label assigned to sentence 1 is greater or lower than the label assigned to sentence 2, respectively. These trends are established when the differences in the emotional attribute scores provided by the corresponding raters are greater than a margin. As implied in Figure 1(a), we set this margin to 1. Finally, a preference between the pair of sentences is decided by aggregating the trends in the matrix. If one sentence is consistently preferred over the other, we establish a preference. This decision is implemented with a threshold, which we set to 60% in this paper. As an illustration, Figure 1(a) shows that sentence 1 is preferred over sentence 2, since 65% of the trends are $\uparrow$ (13 $\uparrow$, 2 $\downarrow$, 5 $=$). This approach can be evaluated between each pair of sentences. Pairs of samples with a preference less than the threshold (60%) are discarded from the set of relative labels used to train and evaluate the preference learning models. We refer to this approach as the QA labels.

2.3. The RankNet Framework

This study relies on the RankNet-based implementation for preference learning. The RankNet algorithm, which was initially presented by Burges [24], employs a probabilistic cost function to train a model to differentiate between pairs of data points using gradient descent. Given the feature vectors Φ_i and Φ_j of two samples (x_i, x_j), a feature representation function $f(\cdot)$ is used to extract the corresponding preference scores given by $s_i = f(\Phi_i)$, and $s_j = f(\Phi_j)$. Using the preference scores, the probability that one sample (x_i) is preferred over the other sample (x_j) is modeled by a sigmoid function as follows:

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

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

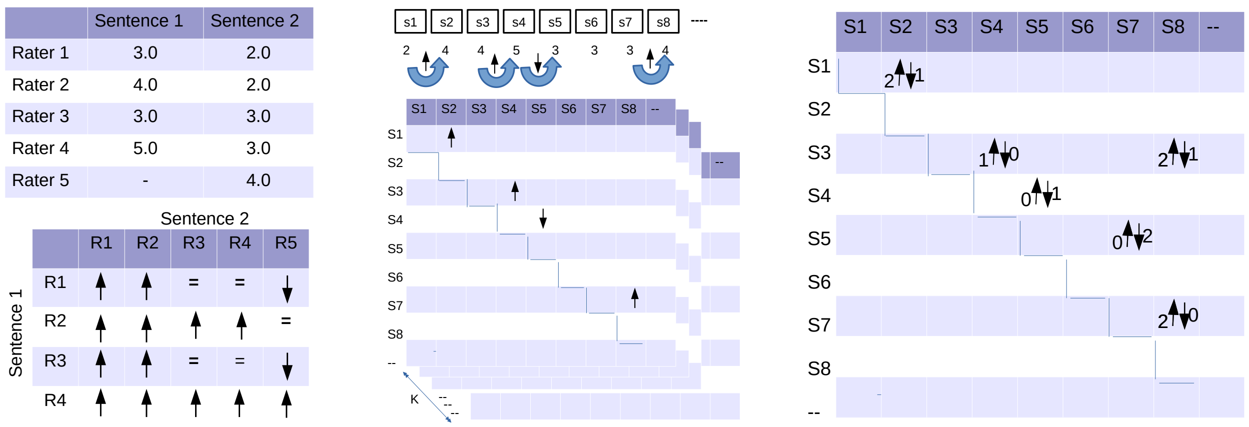
$$\mathcal{L} = -\bar{P_{ij}} \log P_{ij} - (1 - \bar{P_{ij}}) \log(1 - P_{ij}). \tag{2}$$

The loss $\mathcal{L}$ simplifies to $\mathcal{L} = \log(1 + \exp^{-\sigma(s_i - s_j)})$ when $\bar{P_{ij}} = 1$, and $\mathcal{L} = \log(1 + \exp^{-\sigma(s_j - s_i)})$ when $\bar{P_{ij}} = 0$. We implement the feature representation $f(\cdot)$ in the RankNet with two fully connected layers.

3. Proposed Anchor-Based Ordinal Labels

In this section, we describe the proposed anchor-based method to obtain ordinal labels using *consecutive annotations* (CLs). Figure 1 illustrates the two-step approach. Our approach considers the sequential order of all the annotations provided by a particular evaluator, leveraging the thesis that the sentences previously annotated by the evaluator serve as anchors for the emotional perception of the next sentence.

In the first step, we obtain individual CL matrices, as shown in Figure 1(b). We create one matrix per evaluator. We assume that each evaluator completed several sessions. In each session, the evaluator annotated several sentences in order. The top section of Figure 1(b) illustrates a session with eight sentences. Within a session, we consider preference between a pair of samples that are consecutively annotated if and only if the evaluator provided different scores to them. In the example in Figure 1(b), we only have 4 of these cases: sentences 1 and 2 ($\uparrow$), sentences



	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

Sentence 2

Sentence 1	R1	R2	R3	R4	R5
R1	↑	↑	=	=	↓
R2	↑	↑	↑	↑	=
R3	↑	↑	=	=	↓
R4	↑	↑	↑	↑	↑

s1 s2 s3 s4 s5 s6 s7 s8 ----
2 4 4 5 3 3 3 4
S1 S2 S3 S4 S5 S6 S7 S8 --
K
S1 S2 S3 S4 S5 S6 S7 S8 --
2↑↓1
1↑↓0
2↑↓1
0↑↓1
0↑↓2
2↑↓0

(a) QA (b) individual CL matrix of annotators (c) Cumulative CL matrix

Figure 1: *(a) The figure shows the QA-based approach to obtain relative labels from sentence-level annotations (baseline QA labels). Figures (b) and (c) show the proposed CL approach. (b) The figure describes the CL matrix to derive the anchor-based ordinal labels, where K represents the total number of annotators. (c) The figure shows the cumulative matrix obtained after combining CL matrices.*

3 and 4 (↑), sentences 4 and 5 (↓), and sentences 7 and 8 (↑). These trends are used to create the entries of the CL matrices, where the rows and columns of the matrices represent the corresponding pair of sentences (see the entries in the matrix where the four trends in the example of Fig. 1 are placed). With all the preferences from the evaluator, we obtain an $N \times N$ sparse upper triangular matrix, where N represents the total number of samples in the dataset. We indicate a preference between a pair of consecutive samples by $\uparrow$ or $\downarrow$ depending on the scores. If we have K evaluators, this step will generate K CL matrices.

In the second step, all K CL matrices are combined, creating a matrix of size $N \times N$. This matrix aggregates the trends across evaluators, collecting the number of $\uparrow$ and $\downarrow$ trends for each pair of sentences in the database. Figure 1(c) illustrates this process. For example, the entry $2\uparrow\downarrow 1$ in the position (3,8) indicates that at least three evaluators consecutively annotated sentences 3 and 8. Two evaluators preferred sentence 8, and one evaluator preferred sentence 3 with respect to a given emotional attribute. We consider pairs of sentences for training and evaluating the preference learning model if the overall preference for one of the sentences is at least 75%.

Given that we only consider sentences that are consecutively annotated, the resulting matrix is very sparse. However, given that the evaluators consciously or unconsciously made a direct comparison between the pairs of sentences while annotating the samples, these ordinal labels are more reliable for preference learning formulations.

4. Experimental Setting

4.1. The MSP-Podcast Corpus

Our study uses release 1.10 of the MSP-Podcast corpus [27], which is a publicly available database containing over 166 hours of speech annotated with emotional labels. The data is sourced from various audio-sharing websites with Creative Commons licenses. The recordings cover a diverse range of topics, such as science, politics, entertainment, finance, and art. To ensure the quality of the data, all of the speaking turns were filtered to avoid background music, noise, and any overlapped speech. Each turn of the dataset was annotated for the attributes of arousal (calm versus active), valence (negative versus positive), and dominance (weak versus strong) by at least five annotators. The recordings were also annotated with primary and secondary emotional categories, but these are not used in this study.

We consider the order in which the annotations were collected to implement our approach, focusing on consecutive labels. We use the train (63,076 segments), development (10,999 segments), and test (16,903 segments) sets provided by the corpus. These partitions aimed to have speaker independent sets for the train, development, and test sets. We only use relative pairs from samples belonging to the same sets (i.e., we have three separate tables to create the relative labels).

4.2. Feature Extraction

In all our experiments, we extract the Wav2Vec2.0-large feature representation [28], which is used as the input for the feature representation block. This network uses a *convolutional neural network* (CNN) followed by a transformer-based feature encoder to produce a sequence of contextualized speech representations. The Wav2Vec2.0-large model is built using 24 transformer blocks with a model dimension of 1,024. Each vector has a receptive field of 20 ms. For the implementation, we used the pre-trained Wav2Vec2.0 large model from the HuggingFace library [29]. Then, we prune the top 12 transformer blocks and fine-tune the model. This strategy was suggested by Wagner et al. [18]. For fine-tuning the model, we use the train set of the MSP-Podcast corpus, relying on the Adam optimizer [30] 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. Implementation

The baseline relative labels are the ABS (Sec. 2.1) and QA (Sec. 2.2) labels. For the proposed method, we considered three cases. The first case (CL1) considers all the pairs with a preference of more than 75% agreement. The second case (CL2) considers only the pairs which are preferred by at least two annotators with more than 75% agreement. The third case (CL3) considers all pairs which are annotated by at least three annotators with a similar 75% agreement threshold. CL2 and CL3 provide more restrictive sets for the relative labels than the ones provided by CL1, where CL3 is the most restrictive set.

Table 1: *Number of training pairs, measured in thousands, generated by the different implementations of the CL approach. It also shows the coverage, measured in percentage, of the samples in the train set included in the generated relative labels.*

	CL1			CL2			CL3		
	A	V	D	A	V	D	A	V	D
Pairs[K]	204	193	174	43	47	39	21	21	19
Cov.[%]	96.5	95.6	96.8	85.2	88.5	82.7	61.3	64.6	59.8

Table 2: Kendall's Tau *(KT) coefficient of the baselines and proposed methods for arousal, valence, and dominance. Each row provides the results using the ordinal labels indicated by the first entry of the row. Each column indicates how the ordinal labels are obtained for the test set. The symbols* * *and* † *indicate that using a given label leads to significant improvement compared to using the baseline QA, and ABS labels, respectively.*

	ABS	QA	CL1	CL2	CL3
			Arousal		
ABS	0.482	0.496	0.489	0.494	0.497
QA	0.491	0.512	0.481	0.486	0.485
CL1	0.501	0.521*	0.526*†	0.534*†	0.535*†
CL2	**0.504***	**0.527***†	**0.533***†	**0.539***†	0.537*†
CL3	0.498	0.513	0.518*†	0.535*†	**0.539***†
			Valence		
ABS	0.301	0.292	0.284	0.289	0.292
QA	0.311	0.316*	0.298	0.304*	0.302
CL1	0.308	0.321*	0.331*†	0.334*†	0.331*†
CL2	**0.315***	**0.330***†	0.346*†	0.348*†	0.341*†
CL3	0.314*	0.329*†	**0.349***†	**0.351***†	**0.345***†
			Dominance		
ABS	0.380	0.376	0.364	0.369	0.373
QA	0.388	0.393	0.382	0.393*	0.397*
CL1	**0.398**	0.395	0.417*†	0.419*	0.426*†
CL2	0.395	0.402*	**0.428***†	**0.430***†	0.424*†
CL3	0.389	**0.406***	0.416*†	0.426*†	**0.432***†

While CL1 has a higher number of training pairs compared to CL2, and CL3, the lighter restrictions can lead to noisier labels (a trade-off between quantity and quality of the labels). We compared all three implementations of the proposed and baseline labels using the RankNet framework (Sec. 2.3). Due to the differences in the label extraction process across methods, we assess the model trained with the proposed and baseline labels with all the test sets, creating matched and mismatched conditions, avoiding unfair comparisons.

The fully connected layers in the RankNet framework are implemented using two hidden layers with 128 nodes. All the models are initialized with random values and trained for 20 epochs with a learning rate set to 0.00001. We consider the best model based on performance on the development set of the MSP-Podcast corpus. The best model is then evaluated on the test set. We implemented all the models using Tensorflow 2.0, with an NVIDIA GeForce RTX 3090 GPU.

5. Experimental Results

As an ordinal formulation, the RankNet models are trained to rank the samples of the test set according to an emotional attribute. We evaluate the performance using the *Kendall's Tau* (KT) coefficient, which estimates the order of the rank provided by a method. To mitigate the computational burden of analyzing all possible testing pairs, a random subset of 200 samples is selected at a time to estimate the performance. We repeat this process 20 times, reporting the average KT results in Table 2. We evaluate if the results are statistically significant using the one-tailed t-test, asserting significance at a p-value < 0.05. The models using the ABS labels are trained using all possible pairs. For the QA method, we randomly sampled 200k samples from the training set to build the models. Each of the models using the proposed sets of labels (CL1, CL2, CL3) is trained using all possible pairs of samples that satisfy the conditions imposed by each implementation. Table 1 shows the number of training pairs obtained for each criterion, along with the percentage of the training data included in the relative labels (coverage). Even though a limitation of our approach is the sparseness in the labels, since we only consider samples that are consecutively annotated, we have more than 19,000 pairs to train the models even in the worse case (i.e., dominance using CL3). More than 95% of all the training samples are included in the relative labels when we use CL1. With CL2, we still use more than 82% of the training samples.

Table 2 shows the mean KT values across all the proposed and baseline methods when tested on relative labels obtained using all these five methods for arousal, valence, and dominance. For all the attributes, we observe better performances using the proposed method, even on the test sets obtained using the baseline methods. These results indicate that using the proposed anchor-based ordinal labels results in better preference learning models. Among the proposed implementations, we observe better performance using CL2 for most cases in the retrieval of arousal and dominance. CL3 leads to better performance for valence, indicating that a more *pure* training set is necessary. It is reasonable that a harder task, such as predicting valence from speech [31, 32], performs better with less noisy labels even if the size of the training set is reduced. With the anchor-based ordinal labels, the best-performing method showed average relative improvements of $\sim 11\%$ (arousal), $\sim 15\%$ (valence), and $\sim 10\%$ (dominance) compared to the best-performing baseline method. When using the test labels obtained with the baseline approaches, we also observe average relative improvements of $\sim 3\%$ (arousal), $\sim 4.7\%$ (valence), and $\sim 4.5\%$ (dominance) compared to the best models trained with the baseline labels.

6. Conclusions

This study explored the importance of obtaining reliable ordinal labels to train a preference learning framework in speech-emotional retrieval tasks. We considered ordinal labels using consecutive annotations from annotators, which resulted in less noisy and more reliable relative labels. Training preference learning models with these labels led to better performance than training the models with alternative strategies to derive relative labels. We also explored the trade-off between quality and quantity in the implementation of the proposed anchor-based ordinal labels. For arousal and dominance, the best approach was the CL2 implementation, where at least two annotators must agree on the trend. For valence, however, increasing the restriction to improve the quality of the labels led to the best performance, even though the size of the train set decreased. In the future, we want to explore similar strategies to deal with ordinal labels for categorical emotions.

7. Acknowledgement

This work was supported by NSF grants CNS-2016719.

8. References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. 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*. Cambridge, MA, USA: MIT Press, 1997.

[3] H. Martinez, G. 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] G. 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.

[11] S. Parthasarathy, R. Cowie, and C. Busso, "Using agreement on direction of change to build rank-based emotion classifiers," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2108–2121, November 2016.

[12] G. 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] 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.

[14] A. Reddy Naini, M. 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.

[15] 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.

[16] 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: IEEE, April 2018, pp. 5084–5088.

[17] S. Parthasarathy and C. Busso, "Semi-supervised speech emotion recognition with ladder networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2697–2709, September 2020.

[18] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *ArXiv e-prints (arXiv:2203.07378)*, pp. 1–25, March 2022.

[19] 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.

[20] K. Sridhar, W.-C. Lin, and C. Busso, "Generative approach using soft-labels to learn uncertainty in predicting emotional attributes," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2021)*, Nara, Japan, September-October 2021, pp. 1–8.

[21] 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.

[22] G. A. Miller, "The magical number seven, plus or minus two: some limits on our capacity for processing information," *Psychological Review*, vol. 63, no. 2, pp. 81–97, March 1956.

[23] H. Helson, *Adaptation-level Theory*. Harper and Row, January 1964.

[24] 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.

[25] 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.

[26] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," Belfast, Northern Ireland, UK, September 2010, SEMAINE Report D6b. [Online]. Available: http://semaine-project.eu

[27] 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.

[28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems (NeurIPS 2020)*, vol. 33, Virtual, December 2020, pp. 12 449–12 460.

[29] 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. L. 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.

[30] D. 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.

[31] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.

[32] K. Sridhar and C. Busso, "Unsupervised personalization of an emotion recognition system: The unique properties of the externalization of valence in speech," *IEEE Transactions on Affective Computing*, vol. 13, no. 4, pp. 1959–1972, October-December 2022.