UNSUPERVISED DOMAIN ADAPTATION FOR PREFERENCE LEARNING BASED SPEECH EMOTION RECOGNITION

Abinay Reddy Naini[⋆], *Mary A. Kohler*[‡], *and Carlos Busso*[⋆]

[⋆]Department of Electrical and Computer Engineering, The University of Texas at Dallas
[‡]Laboratory for Analytic Sciences, North Carolina State University

ABSTRACT

Retrieving speech samples that have specific expressive content has many applications. It is desirable to build a preference learning framework that ranks speech samples according to emotional attribute values that generalize well to new domains. A popular architecture for preference learning is the RankNet framework, which uses a function to obtain the preference between pairs of speech sentences. This study explores implementing this function with alternative feature representations that are explicitly selected to reduce the mismatch between source and target domains. In particular, we implement our preference-learning based *speech emotion recognition* (SER) system using ladder networks and adversarial domain adaptation. The study also proposes a novel combination of these two unsupervised domain adaptation strategies. The experimental results in cross-corpus evaluations using the MSP-Podcast and MSP-IMPROV datasets reveal that the proposed adversarial domain adaptation on a ladder network-based feature representation performs the best across different conditions. The results also show that preference learning leads to better precision for retrieval tasks than comparable SER systems built to directly predict absolute emotional attribute scores.

***Index Terms*—** Speech emotion recognition, Preference learning, Domain adaptation

1. INTRODUCTION

Recognizing emotions plays an important role in developing advanced interface systems, which enhance human-computer interactions [1, 2]. While a typical *speech emotion recognition* (SER) system classifies speech into emotional categories such as happiness, sadness, and anger [3, 4], an appealing alternative is to predict emotional attributes such as arousal, valence, and dominance [5–7]. Extensive research has been done to improve SER systems, where the task is to predict emotional attributes with models trained with labels derived from subjective evaluation from multiple annotators [8, 9]. However, achieving annotations with a high inter-evaluator agreement is difficult due to the differences in perceiving emotions across people [10]. One promising approach is to rank emotional behaviors using preference learning strategies [11–14].

A preference learning system is designed to determine the relative emotional level between a pair of samples concerning a particular emotion or attribute. The results can be used to determine the ranking order of several test samples. Studies have shown that annotators tend to agree better on relative trends [15–18], creating better labels to train preference learning models [11, 19]. Furthermore, a preference learning framework can be more suitable for retrieval tasks, providing the ideal tool for screening massive speech repositories in search for samples with target emotional content. Among the most popular methods is to use the RankNet framework [20] to train a preference learning model [12, 21]. This approach creates a feature transformation that is applied to a pair of speech samples to determine the preference with respect to a given emotional descriptor. The function in the RankNet formulation is general and can be implemented with different machine-learning strategies. This study implements this function with a key objective in mind: generalization to a new domain, collected under different settings.

One of the barriers for deploying SER systems is generalization across different domains. The mismatches between domains include acoustic conditions (microphone, environmental noise), idiosyncratic differences, inconsistency in the distribution of the emotional content conveyed in the recordings, and differences in language and culture. We explore two unsupervised approaches that have been successfully implemented on SER: ladder network [22], using intermediate layer representation reconstructions, and adversarial *domain adaptation* (DA) [23], using domain classification implemented with a gradient reversal. To the best of our knowledge, this is the first attempt to make a preference learning-based SER model robust to different domains. The results demonstrate that building the RankNet function with these training strategies leads to better performance in cross-corpus evaluations than models that are not adapted to the target domain. An important contribution of this study is the finding that these unsupervised adaptation frameworks are complementary, where we show that a model that combines these approaches achieves the best performance when tested on an unlabeled domain. The results also demonstrate that the preference learning frameworks lead to better performance for retrieval tasks than models implemented to predict the absolute attribute scores in the speech files.

2. RELATED WORK

This section describes previous studies using preference learning for affective computing tasks. Yannakakis et al. [16, 18] detailed the benefits of using preference learning models for emotion recognition problems. For categorical emotions, Cao et al. [13, 24] proposed a ranking method for categorical emotions 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 [25] 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. [21] used *consistent rank logits* (CORAL)-based method to jointly train multiple ordinal binary SER tasks for improving consistency across subclassification tasks. For emotional attributes, Martinez et al. [11] showed that better generalization can be achieved with a rank-based transformation of emotional attributes than by grouping them into classes. Lotfian and Busso [14] discussed practical considerations on

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

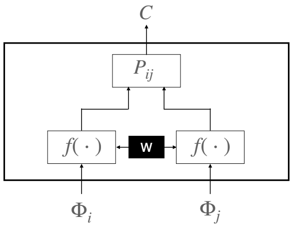


Fig. 1: Block diagram of RankNet. This study explores different implementations for the function $f(\cdot)$.

using preference learning in ranking speech according to emotional attributes, including the margin in the value of the target attribute required to prefer one sample over another. Instead of relying on consensus labels, Parthasarathy and Busso [19, 26] proposed strategies based on the qualitative agreement method [17] to create more reliable rank labels by identifying trends across evaluators.

In terms of frameworks for preference learning in SER tasks, Parthasarathy et al. [12] compared several alternative methods, including RankSVM, RankNet, and DNNRegression, showing that RankNet led to better performance.

3. PROPOSED PREFERENCE LEARNING FRAMEWORK

Our study implements RankNet using state-of-the-art domain adaptation strategies to increase the generalization of the models when evaluated on a different domain. This section describes the RankNet framework, and the building blocks of our proposed implementation.

3.1. The RankNet Framework

Figure 1 shows the formulation for RankNet, which is the preference learning framework used in this study. The RankNet algorithm was introduced by Burges [20]. It uses a probabilistic cost function to learn preference between pairs of data points using gradient descent. A typical RankNet formulation consists of a function $f(\cdot)$ that serves as a feature representation for two samples. Given the feature vectors Φ_i and Φ_j of the two samples, it generates the scores $s_i = f(\Phi_i)$, and $s_j = f(\Phi_j)$. The probability that sample x_i is preferred over sample x_j, denoted by $x_i >> x_j$, is given by $P_{ij} = \frac{1}{1+e^{-\sigma(s_i - s_j)}}$. During training, the function $f(\cdot)$ is trained using the ground truth labels with preferences between sample pairs. 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 cost function $\mathcal{L} = -\bar{P_{ij}} \log P_{ij} - (1 - \bar{P_{ij}}) \log(1 - P_{ij})$ is the cross entropy between the expected probability $\bar{P_{ij}}$ and actual probability P_{ij}, which is used to train the parameters of the function $f(\cdot)$. $\mathcal{L}$ 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 $\bar{P_{ij}} = 0$.

3.2. Chunk-Based Segmentation (CBS)

We consider the *chunk-based segmentation* (CBS) proposed by Lin and Busso [27, 28] to obtain a fixed dimensional discriminative feature vector for variable-length speech segments. Figure 2(a) shows a diagram of the CBS process. The first step is to extract frame-level acoustic features for the entire sentence (Sec. 4.2). The features are split into chunks creating a fixed number of chunks with fixed duration regardless of the duration of the sentence. This dynamic chunk segmentation is achieved by changing the overlap between chunks according to the duration of the sentence. We obtain the chunk-level representation using a *long short-term memory network* (LSTM) on each chunk. Finally, the chunk-level representations are temporally aggregated to obtain a sentence-level aggregation, using the RNN-AttenVec model in Lin and Busso [27]. The final output is a discriminative sentence-level feature vector. The CBS approach is highly parallelizable and efficient given the fixed number of chunks, which have fixed durations.

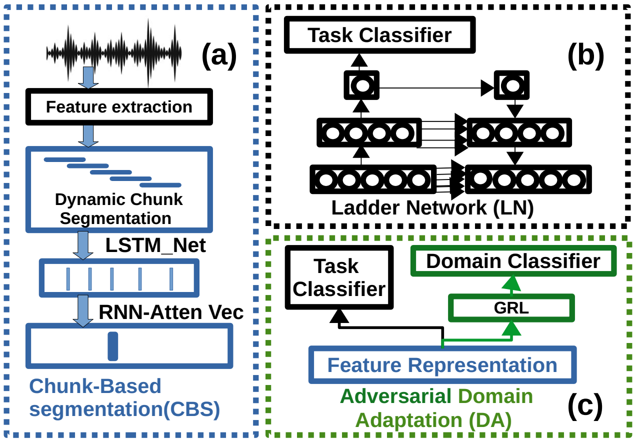


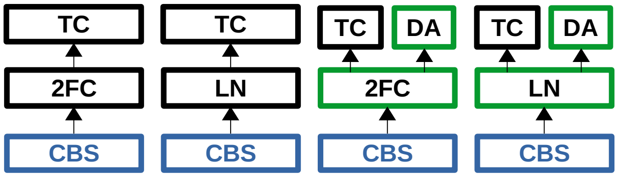
Fig. 2: Block diagram of the building blocks of our proposed preference learning approach. (a) Chunk-based segmentation, (b) ladder network, and (c) adversarial domain adaptation.

3.3. Ladder Network (LN)

The first strategy to increase the generalization of the model is the ladder network strategy [22]. We consider the ladder network implementation proposed by Parthasarathy and Busso [8] for SER. Figure 2(b) shows the diagram, which consists of an encoder and a decoder with horizontal lateral connections at each layer. Gaussian noise is added to the encoder, and the goal is to recover a clean version of the corresponding encoder. The goal of each decoder layer is to recover a clean version of the input given to the corresponding encoder layer, as the input to each encoder layer is corrupted with added Gaussian noise. The cost function consists of two weighted components: one for the task classifier, which is implemented at the end of the encoder, and one for the reconstruction losses. While both losses are considered for the labeled source domain, only the reconstruction losses are used for the target domain, which is assumed to be unlabeled. The ladder network is known to minimize the mismatch between domains, providing an effective training strategy for SER tasks [29].

3.4. Adversarial Domain Adaptation (DA)

The second strategy to improve the generalization of the preference learning model is adversarial *domain adaptation* (DA) [23]. Figure 2(c) shows the diagram, which follows the implementation presented in Abdelwahab and Busso [30]. The approach has a task classifier and a domain classifier. These two blocks share a common feature representation block. The goal of the domain classifier is to recognize if the data is coming from the source or target domain. In the adversarial DA approach, we introduce a *gradient reversal layer* (GRL) between the domain classifier and the feature representation network, which forces the domain classifier to produce maximum classification error in classifying the source and target domains. This strategy generates a feature representation network that creates similar responses for data from the source and target domains, effectively



(a) CBS+2FC (b) CBS+LN (c) CBS+DA (d) CBS+LN+DA

Fig. 3: Proposed architectures used to build function $f(\cdot)$ in RankNet. TC: *task classifier*, FC: *fully connected* layers.

reducing the mismatch between domains. The losses of the task and domain classifiers are considered when the network is trained using labeled data from the source domain. However, only the loss of the domain classifier is considered when the model is trained with the unlabeled target domain. By alternately considering the labeled data from the source domain and unlabeled data from the target domain, the model ensures maximum task classification accuracy, while making samples from both domains indistinguishable.

3.5. Proposed Architectures

Figure 3 shows the four architectures for $f(\cdot)$ that we consider, which are all implemented with the CBS as the initial feature representation. Figure 3(a) illustrates the first model, referred to as CBS+2FC, which includes two fully connected layers on top of the CBS. This model serves as a baseline to understand the benefits of domain adaptation. Figures 3(b) and 3(c) show the implementation of the two domain adaptation strategies. CBS+LN considers a *ladder network* (LN), in place of the 2FC layers, and CBS+DA adds the adversarial *domain adaptation* (DA) to the CBS+2FC framework. A contribution of this study is the combination of LN and DA. LN uses reconstruction losses and DA uses GRL to create a feature representation that has similar responses for samples across domains. These strategies use two different approaches to provide complementary robustness that increases the generalization of the models. This model is referred to as CBS+LN+DA, which is illustrated in Figure 3(d). The adversarial DA is added as an additional task on top of the encoder of the LN.

4. EXPERIMENTAL SETTING

4.1. Emotional Databases

We consider the MSP-Podcast corpus [31] as the source domain. We use release 1.10 consisting of 104,267 speaking turns from different audio recordings with Creative Commons licenses. The training set includes 63,076 speaking turns. All the speaking turns are obtained after removing background music, noise, and speech overlaps. Each turn is annotated by five annotators for emotional attributes, and primary and secondary emotional categories. This study uses emotional attributes for arousal, valence, and dominance.

We consider the MSP-IMPROV dataset [32] as the target domain. The MSP-IMPROV corpus consists of dyadic interactions between 12 actors consisting of a total of 8,438 speaking turns. The MSP-IMPROV database also provides annotations for the emotional attributes of arousal, valence, and dominance. Each speech file is annotated by at least five annotators. Unlike the MSP-Podcast corpus, the audio was recorded in a closed environment, making it perfect for testing the domain mismatch. Data from the six actors from the first three sessions are used as the test set. We have reserved the remaining sessions to train (matched condition) or adapt (mismatched condition) the proposed models.

4.2. Feature extraction

For all our experiments, we extracted the wav2vec2-large feature representation [33], which is used as input for the CBS. The wav2vec2-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 [34]. Then, we prune the top 12 transformer blocks, and fine-tuned the model using the MSP-Podcast corpus. For fine-tuning the model, we rely on the Adam optimizer [35] with a learning rate set to 0.00001 for 10 epochs.

We also present a baseline using the *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) [36], which includes 88 acoustic features.

4.3. Preference learning labels

We rely on the *qualitative agreement* (QA)-based method proposed by Parthasarathy and Busso [19] for obtaining preference labels. The approach identifies trends across individual annotations, instead of relying on the consensus label. First, it creates an $N_1 \times N_2$ matrix with pairwise comparisons across annotations assigned to two sentences, where N_1 and N_2 are the number of annotators for sentence 1 and 2. The elements of this matrix indicate if an annotation for sentence 1 is greater, equal or lower than an annotation for sentence 2. Then, the approach estimates the percentage for each preference. Similar to Parthasarathy and Busso [19], we consider a pair if 60% of the matrix's entries show that a sentence is preferred over the other.

4.4. Implementation

We consider three baselines to implement $f(\cdot)$. The first baseline is the CBS+2FC model (Fig. 3(a)). The second baseline extracts the eGeMAPS features as the input of a network with just two fully connected layers. We refer this model to as eGeMAPS+2FC. The third baseline corresponds to the CBS+LN model (Fig. 3(b)) trained in matched conditions (e.g., train and test on the MSP-IMPROV corpus). This approach is referred to as CBS+LN* in Table 1.

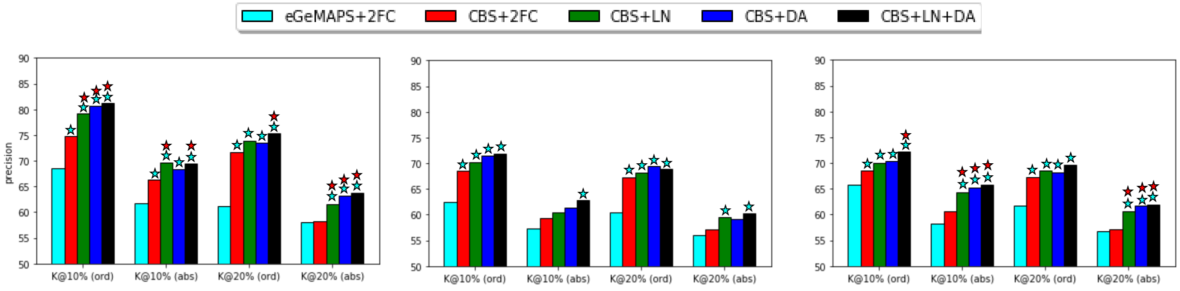
The FC block in Figures 3(a) and 3(c) is implemented with two hidden layers implemented with 128 nodes. The LN architecture is also implemented with two hidden layers with 128 nodes. The DA block in Figure 3(c) and 3(d) consists of two layers implemented with 64, and 32 nodes, respectively. Each of the models is trained for 20 epochs with a learning rate set to 0.0001. We consider the best model using its performance on the development set of the MSP-Podcast corpus. We have implemented all the models using Tensorflow 2.0 and Keras, with an NVIDIA GeForce RTX 3090 GPU.

5. EXPERIMENTAL RESULTS

As mentioned in Section 4.1, we train the models with the MSP-Podcast corpus, evaluating the performance on the MSP-IMPROV corpus. We evaluate if the results are statistically significant using the one-tailed t-test, asserting significance at p-value < 0.05.

One advantage of preference learning is that the number of training samples can be increased by including different pairs. If the training set has L samples, we can have $L(L-1)/2$ pairs. First, we evaluate the optimal amount of data needed to train the proposed preference learning framework using only the model CBS-LN and the MSP-Podcast corpus (120K, 240K, 480K, 720K, and 1M pairs). We observe that the top performance is achieved with $\sim$480k pairs, which we have used for training all the models.

First, we compare the performance of the preference learning methods for valence, arousal, and dominance. We report perfor-



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

Fig. 4: Precision at K achieved by different preference learning models in retrieving samples with low or high values of emotional attributes. The figure compares models implemented for both ordinal (ord), and absolute prediction (abs). A star on top of the bar indicates that a method is significantly better than the method represented by the color of the star.

Table 1: *Kendall's Tau* (KT) coefficient and accuracy of the baselines and proposed methods. The column "Labeled" indicates the data used to train the models. The column "Unlabeled" indicates the data used for the unsupervised domain adaptation.

Method	Labeled	Unlabeled	Arousal		Valence		Dominance	
			KT	Accuracy	KT	Accuracy	KT	Accuracy
CBS+LN*	MSP-IMPROV	MSP-IMPROV	0.617±0.013	84.7±1.42	0.375±0.018	66.6±1.34	0.558±0.015	81.2±1.42
eGeMAPS+2FC	MSP-Podcast	-	0.372±0.006	64.3±0.98	0.208±0.004	58.7±1.13	0.316±0.009	62.7±1.04
CBS+2FC	MSP-Podcast	-	0.417±0.011	71.9±0.87	0.258±0.010	61.8±0.91	0.392±0.006	69.2±0.85
CBS+LN	MSP-Podcast	MSP-IMPROV	0.484±0.005	79.2±1.22	0.313±0.007	63.9±1.26	0.447±0.011	72.3±1.01
CBS+DA	MSP-Podcast	MSP-IMPROV	0.462±0.010	78.5±1.04	0.301±0.013	63.7±1.71	0.442±0.009	71.5±1.17
CBS+LN+DA	MSP-Podcast	MSP-IMPROV	0.506±0.007	80.6±1.36	0.312±0.012	64.2±1.63	0.461±0.013	74.7±1.23

mance using both accuracy and the *Kendall's Tau* (KT) coefficient. Accuracy is defined as the percentage of the preferences in the test set correctly predicted by the models. The KT coefficient measures the order provided by models. To avoid computing the results across all possible testing pairs, we consider 200 randomly selected samples from the test set at a time to compute the metrics. We repeat this process 20 times, reporting the average and *standard deviation* (SD) of the performance across the 20 repetitions. Table 1 shows the performance comparison between all the baselines and the proposed models. It also shows the model CBS+LN* as a reference for matched conditions. As expected, CBS+LN* performs the best among all the methods given that it is trained and tested on the same dataset. Among the proposed preference learning models, CBS+LN+DA achieves the best performance in cross-corpus evaluations with at least 5% improvement over the second-best model in terms of the KT coefficient (p-value < 0.01). We also observe important improvements in the KT coefficient over the eGeMAPS+2FC baseline with $\sim$36% gain for arousal, $\sim$50% gain for valence, and $\sim$45% gain for dominance (p-value < 0.005). These results show the benefits of adapting the preference learning models using these unsupervised domain adaptation methods. The table also shows that using the Wav2Vec2 features with CBS leads to better results than using the eGeMAPS feature set. We also observed similar trends in accuracy improvements across the results. We do not observe a significant difference in performance between the two domain adaptation models (i.e., CBS+LN, and CBS+DA). The improvements observed while combining these models (e.g., CBS+LN+DA) indicate that these domain adaptation strategies are complementary, providing better generalization on new domains. The very low standard deviation values across all the results in Table 1 show that the models are fairly consistent with performance on multiple test samples.

We also assess the performance of the methods using *precision at K* (P@K). This metric estimates the precision when K percentage of the data is retrieved. For each attribute, we split the test set into *low* and *high* classes using the median split. We consider a success if the retrieved sample belongs to the correct class (i.e., class *high* if we are searching for samples with high value for the attribute). To understand the importance of preference learning in retrieval tasks, we also implement models that directly predict the scores of the emotional attributes. The absolute models follow the same architecture shown in Figure 3, adding a linear output layer on top of these functions. These models are trained with the average emotional attributes assigned by multiple annotators. The rank is generated based on the predicted emotional attribute provided by the models. Figure 4 shows the P@K results with $K = 10\%$ and $K = 20\%$. The preference learning-based models are marked with *(ord)*, and the absolute-based models with *(abs)*. All the proposed frameworks are significantly better than the eGeMAPS+2FC baseline. We observe better performance for preference-learning models compared to the corresponding absolute-based implementations. Overall, we observe that methods with domain adaptation (CBS+LN, CBS+DA, CBS+LN+DA) performed better than the models without any adaptation (eGeMAPS+2FC, CBS+2FC). This result is particularly clear for arousal, where the CBS+LN+DA model achieves significantly better results than both baseline models without domain adaptation.

6. CONCLUSIONS

This study explored the importance of domain adaptation in a preference learning framework, where the goal is to increase the generalization of the SER model. We considered combinations of ladder network and adversarial domain training strategies, implemented using the RankNet framework. The results showed the best performance when these domain adaptation methods were jointly combined, showing the complementary robustness added by these strategies. We also observed that the preference learning-based framework is better suited for speech emotion retrieval tasks, leading to better performance than methods trained to predict the absolute score of emotional attributes. In the future, we will explore how a preference learning framework can be used in cases where the goal is to predict the absolute score associated with emotional attributes.

7. 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] 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.

[4] 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.

[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," *ArXiv e-prints (arXiv:2203.07378)*, pp. 1–25, March 2022.

[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] 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.

[9] 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.

[10] V. Sethu, E. Mower Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, "The ambiguous world of emotion representation," *ArXiv e-prints (arXiv:1909.00360)*, pp. 1–19, May 2019.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] R. Cowie and G. McKeown, "Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme," September 2010, SEMAINE Report D6b.

[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, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Interspeech 2018*, Hyderabad, India, September 2018, pp. 252–256.

[20] 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.

[21] 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.

[22] A. Rasmusi, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems (NIPS 2015)*, Montreal, Canada, December 2015, pp. 3546–3554.

[23] Y. Ganin et al., "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, April 2016.

[24] 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.

[25] 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.

[26] 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.

[27] 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. Early Access, 2022.

[28] W.-C. Lin and C. Busso, "An efficient temporal modeling approach for speech emotion recognition by mapping varied duration sentences into fixed number of chunks," in *Interspeech 2020*, Shanghai, China, October 2020, pp. 2322–2326.

[29] S.-G. Leem, D. Fulford, J.-P. Onnela, D. Gard, and C. Busso, "Separation of emotional and reconstruction embeddings on ladder network to improve speech emotion recognition robustness in noisy conditions," in *Interspeech 2021*, Brno, Czech Republic, August-September 2021, pp. 2871–2875.

[30] M. Abdelwahab and C. Busso, "Domain adversarial for acoustic emotion recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2423–2435, December 2018.

[31] 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.

[32] 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.

[33] 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)*, Virtual, December 2020, vol. 33, pp. 12449–12460.

[34] T. Wolf et al., "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.

[36] F. Eyben et al., "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.