Ordinal learning for emotion recognition in customer service calls

Wenjing Han, Tao Jiang, Yan Li, Björn Schuller, Huabin Ruan

Angaben zur Veröffentlichung / Publication details:

Han, Wenjing, Tao Jiang, Yan Li, Björn Schuller, and Huabin Ruan. 2020. "Ordinal learning for emotion recognition in customer service calls." In *International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020), Barcelona, Spain, 4-8 May 2020*, edited by Ana I. Pérez-Neira, Xavier Mestre, Pau Closas, and Mónica Bugallo, 6494–98. New York, NY: IEEE. https://doi.org/10.1109/icassp40776.2020.9053648.

ORDINAL LEARNING FOR EMOTION RECOGNITION IN CUSTOMER SERVICE CALLS

Wenjing Han[1], *Tao Jiang*[1], *Yan Li*[1], *Björn Schuller*[2,3], *Huabin Ruan*[4*]

[1]Kuaishou Technology Corp., Beijing, China
[2]GLAM-Group on Language, Audio & Music, Imperial College London, UK
[3]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
[4]School of Life Sciences, Tsinghua University, Beijing, China

ABSTRACT

Approaches toward ordinal speech emotion recognition (SER) tasks are commonly based on the categorical classification algorithms, where the rank-order emotions are arbitrarily treated as independent categories. To employ the ordinal information between emotional ranks, we propose to model the ordinal SER tasks under a COnsistent RAnk Logits (CORAL) based deep learning framework. Specifically, a multi-class ordinal SER task is transformed into a series of binary SER sub-tasks predicting whether an utterance's emotion is larger than a rank. All the sub-tasks are jointly solved by one single network with a mislabelling cost defined as the the sum of the individual cross-entropy loss for each sub-task. Having the VGGish as our basic network structure, via minimizing above CORAL based cost, a VGGish-CORAL network is implemented in this contribution. Experimental results on a real-world call center dataset and the widely used IEMOCAP corpus demonstrate the effectiveness of VGGish-CORAL compared to the categorical VGGish.

Index Terms— speech emotion recognition, ordinal classification, consistent rank logits, VGGish

1. INTRODUCTION

As a frontline function, customer call support is of great importance for increasing a company's customer retention. To provide high-quality support, agents should not only answer customers' product-related questions professionally, but also address customers' negative emotions decently. Nowadays, with the rapid progression of artificial intelligence technology, there have been a growing trend to apply speech emotion recognition (SER) technique to estimate emotions of customers or agents from their conversations, thereby to help provide better call support [1, 2, 3, 4, 5]. To the same end, in this paper, we explore methods for developing an emotion estimator on real-world call center data using acoustic cues. Noticeably, we place special focus on detecting customers' rank-order negative behaviors. In contrast with the conventional *categorical* SER tasks that classify affective behaviors into nominal categories (e.g., *Happy*, *Sad*, *Angry*, etc.) [6, 7, 8], or the *continuous* SER tasks that recognize emotional behaviors described by real-valued attributes (e.g., *arousal*, *valence*, etc.) [9, 10, 11], we consider actually the *ordinal* SER task that recognizes emotional behaviors into ordinal labels measured on an interval scale (e.g., *1-Non-negative*, *2-Somewhat Negative*, *3-Obviously Negative* in this work).

Among previous works, ordinal SER tasks were often cast as categorical classification problems [5, 12, 13, 14], where the class labels were implicitly assumed to be independent to one another, despite they had a strong ordinal relationship in fact. Gradually, studies exploiting ordinal information in SER appeared in the literature [15, 16, 17, 18, 19]. A popular strategy in these studies was building a ranker modified from well-known classification algorithms, such as Support Vector Machine (SVM) based rankers in [17, 18] and Deep Neural Network (DNN) based ranker in [19], to predict the rank order of a set of samples on each emotional attribute. Indeed, and although this strategy can predict the relative emotional ranking between different samples, it cannot classify a sample's emotion into an absolute rank. Given this scenario, it is necessary to develop ordinal SER approaches that allow the utilization of ordinal relationship between emotion labels to improve prediction performance, plus the output of pre-set ranks to tell the absolute emotional levels at the same time.

Early works focusing on ordinal classification can be found in other domains [20, 21, 22]. Among them, the ordinal classification problem was normally reduced into a series of simpler binary classification sub-problems. A benefit of this kind of hybrid approaches is that new generalization bounds for ordinal classification can be easily derived from known bounds for binary classification, whereas a shortcoming is the unguaranteed consistency among binary classifiers, that is, the predictions for individual binary tasks can disagree. Lately, there were works that explore end-to-end approaches based on DNNs to address the ordinal classification problems and achieved effective performance in age estimation tasks [23, 24]. One approach worth noting is the so-called COnsistent RAnk Logits (CORAL) method. As approved in [24], in comparison with the other hybrid approaches, CORAL can not only simplify the model building procedure, but also theoretically guarantee

*Corresponding author. Email: ruanhuabin@tsinghua.edu.cn (H. R.). This work was supported by the National Natural Science Foundation of China (Grant No. 61802226).

the prediction consistency between sub-classification tasks.

Thus, in this paper, we aim to explore the use of CORAL based ordinal learning in estimating ordinal emotions. First, we choose the Audio Set VGGish [25] to be our basic network structure. Then, we utilize the CORAL formulation to guide label encoding and network learning. Specifically, instead of the typical one-hot label encoding for categorical classification, our emotional labels are encoded following a non-orthogonal recipe to reflect ranks. Taking advantage of the encoded labels, the conjunction of each output unit and all the previous layers can be viewed as a distinct binary SER classifier which is trained according to whether the emotional rank of an utterance is greater than a certain level. These binary SER classifiers share the same weight parameters excluding the bias units of the final layer. Such setting combined with a cross entropy based loss ensures greater penalty to larger classification errors, as well as the rank consistency among the binary classifiers. Moreover, a task importance weighting strategy is employed during the loss calculation in order to relief the data imbalance issue. Finally, the effectiveness is validated in a freshly collected middle-scale real-world call center database, together with the widely employed IEMOCAP corpus [26].

2. RELATED WORK

So far, less attention was paid on the ordinal SER tasks. Schuller *et al.*'s work in [12] and Deng *et al.*'s work in [13] can be recognized as bi-ordinal SER tasks actually. The former mapped emotion categories to *Low/High* arousal/valence and used SVM as the classifier. The latter mapped emotions to *Negative/Positive* valence and trained a sparse autoencoder for classification. In addition, there were also works coping with multi-class ordinal SER which mainly existed in the field of call center monitoring, though. Gupta *et al.* [3] predicted emotions on a 3-point scale, namely *Happy*, *Neutral*, and *Angry*, with Gaussian Mixture Models, from call center data. Lately, Li *et al.* [5] measured emotions in call center dialogs on a 5-point scale: *Clearly/Somewhat Positive/Negative* and *Neutral*. They combined several classifiers processing different acoustic and lexical features to achieve the final decision. We also notice that Zhang *et al.* [14] classified emotional attributes on a 3-point scale from the corpus IEMOCAP. A DNN based end-to-end framework was directly used on the raw audio signal. However, in the above works, no matter what kind of classifiers the authors constructed, no ordinal information between emotions were considered yet. Another possible but seldom used solution to model the ordinal SER task could be the regression-based approaches modelling *continuous* emotion, since the real-valued regression results could be quantized to ranks with multiple thresholds, however the threshold selection is tricky. In our approach, the boundary thresholds are actually transformed to the network's output biases, which can be well learned during the network training.

Recently, a group of researchers started to attach importance to the ordinal nature of emotions. In [27] , Yannakakis *et al.* recommended to annotate emotions in an ordinal way, and then model the ordinal labels with preference learning approaches. Following this, Parthasarathy *et al.* [19] used a deep learning ranker implemented with the RankNet algorithm to evaluate emotional preference between sentences in terms of attributes. Perhaps our focused problem is most similar to Parthasarathy *et al.*'s [19], but there is a fundamental difference we would like to underscore. They evaluated the emotional ranking between each pair of samples, but did not classify emotions directly into pre-set ranks described on an interval scale as our work requires.

Table 1. Data distribution, duration and Fleiss' κ for each emotional label in the call center database.

Label	Count	Duration [h]	κ
Non-negative	$2,317$	1.5	0.93
Somewhat Negative	$1,701$	1.1	0.68
Obviously Negative	519	0.4	0.75
In Total	$4,537$	3.0	0.79

3. DATA COLLECTION AND ANNOTATION

The database is created from recorded customer support calls in Chinese (8 kHz, mono). It consists of 129 conversation sessions in total. Each session involves a customer and an agent, but we only concern the speech from customer side in this work. Before emotion annotation, customer speech is segmented into utterances automatically. Then, each utterance is labeled by 10 annotators on a 3-point scale: *1-Non-negative*, *2-Somewhat Negative*, and *3-Obviously Negative*. Additionally, 2 more labels are designed for utterances that are not part of the emotion task: *Non-speech* for non-speech audio and *Non-understandable Speech* for poor-quality or heavily accented speech audio. To select utterances with consistent enough annotations, we follow the procedure described below:

(i) Discard the utterances marked as *Non-speech* or *Non-understandable Speech* by a majority of the annotators.

(ii) For the set of remaining utterances, delete those annotators' annotations whose Pearson's correlation with the averaged annotations are below 0.6.

(iii) For each remaining utterance, calculate its mean and standard deviation of remaining annotations. If an annotation is farther than one standard deviation from the mean, this annotation is discarded.

(iv) Discard also utterances without majority agreement.

This procedure results in a reduction from 5,270 to 4,537 utterances, each of which retains annotations from 4–9 annotators (4,129 with 9 annotations, 408 with 4-8 annotations). To analyze the reliability of agreement between the 9 annotators. Fleiss' κ is computed and shown in the 4*th* column of Table 1. The results reveal that annotators reach the highest agreement ($\kappa = 0.93$) on the annotation of *Non-negative* data. It is reasonable since *Non-negative*, which involves both *Positive* and *Neutral* actually, is more coarse-grained than *Somewhat Negative* and *Obviously Negative*. Therefore, this is relatively

easier for annotators to identify. One should also note that a κ of 0.79 on the entire database indicates good agreement between annotators. Distribution and duration of the retained data for each label are also shown in Table 1, where a heavy data imbalance can be noticed.

4. ORDINAL LEARNING WITH CORAL

Formally, let $\mathcal{S} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ be the training set consisting of N samples, where $\mathbf{x}_i \in \mathcal{X}$ is the i-th input utterance and $y_i \in \mathcal{Y} = \{r_1, r_2, \ldots, r_K\}$ is its corresponding emotional rank. For the K ranks we have $r_1 \prec r_2 \prec \cdots \prec r_K$, where $\prec$ denotes the ascending ordering. The goal of the ordinal SER task is to learn a mapping from utterances to emotional ranks $\mathcal{N}(\cdot) : \mathcal{X} \mapsto \mathcal{Y}$ such that a predefined loss function L is minimized. As mentioned, in order to improve the performance of ordinal SER, we explore the strategy of complementing a VGGish with the CORAL formulation (VGGish-CORAL). Figure 1 illustrates the framework evolution from the VGGish classifier (Figure 1(a)) to our proposed VGGish-CORAL ranker (Figure 1(b)). By contrast, two main modifications are made on label encoding and network learning respectively.

4.1. From VGGish to VGGish-CORAL

VGGish is a network pretrained on a large-scale Audio Set, and so potentially have stronger discriminative ability [25]. When it is applied for categorical classification tasks, a one-hot encoding method is commonly utilized. Unlike, we convert a rank label y_i into a vector consisting of $K-1$ binary labels $(y_i^1, \cdots, y_i^{K-1})$, where $y_i^k (k = 1, 2, \cdots, K-1)$ indicates whether y_i exceeds rank $r_k (i.e., y_i^k = 1\{y_i > r_k\})$. The boolean test $1\{\cdot\}$ equals 1 if the inner condition is true, and 0 otherwise. The underneath thinking is to transform the K-rank ordinal classification task into $K-1$ binary classification tasks. Providing the extended binary labels as goal outputs, we thus reconstruct the VGGish structure with $K-1$ units in the output layer (cf., Figure 1(b)). Each output unit corresponds to a distinct binary classification task. Then, the network can be viewed as a hybrid of $K-1$ binary classifiers.

According to the CORAL method [24], if let W denote the weight parameters of the neural network excluding the bias units of the final layer, b_k denote the bias corresponding to the k-th output unit, and $s(z) = 1/(1 + exp(-z))$ be the logistic sigmoid function, the predicted empirical probability for task k is defined as:

$$\hat{P}(y_i^k = 1) = s(g(\mathbf{x}_i, W) + b_k), \tag{1}$$

where $g(\mathbf{x}_i, W)$ is the input of the output units. Note that, in VGGish-CORAL, each task k shares the same weight parameters W but has independent bias units b_k. For model training, we minimize the loss function:

$$\begin{aligned} L = -\sum_{i=1}^{N}\sum_{k=1}^{K-1} \lambda^k [log(s(g(\mathbf{x}_i, W) + b_k)) y_i^k \\ + log(1 - s(g(\mathbf{x}_i, W) + b_k))(1 - y_i^k)], \end{aligned} \tag{2}$$

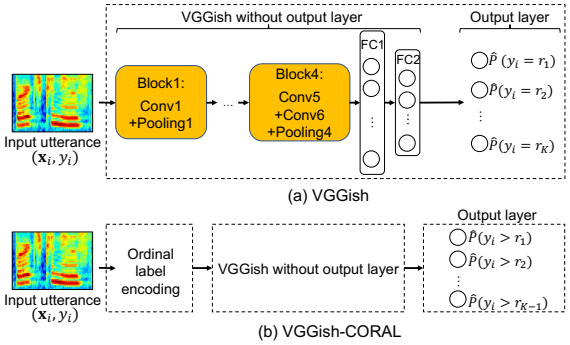

Fig. 1. Framework evolution from the VGGish based classifier (a) to our proposed VGGish-CORAL ranker (b).

which is the weighted cross entropy of the $K-1$ binary classifiers. A detailed description in regard to task weighs $\{\lambda_k\}_{k=1}^{K-1}$ is presented in the next subsection. Here, loss L is designed to weight larger classification errors more, because more of the individual cross entropy terms corresponding to binary classifiers will be violated.

Based on the binary task responses, the predicted rank for an input $\mathbf{x}_i$ is obtained via:

$$h(\mathbf{x}_i) = r_q, q = 1 + \sum_{k=1}^{K-1} f_k(\mathbf{x}_i), \tag{3}$$

where $f_k(\mathbf{x}_i) \in \{0, 1\}$ is the prediction of the k-th binary classifier, and defined as:

$$f_k(\mathbf{x}_i) = 1\{\hat{P}(y_i^k = 1) > 0.5\}. \tag{4}$$

By minimizing the loss L defined in Eq. (2), the $\{f_k\}_{k=1}^{K-1}$ are *rank-monotonic*, i.e., $f_1(\mathbf{x}_i) \leq f_2(\mathbf{x}_i) \leq \cdots \leq f_{K-1}(\mathbf{x}_i)$, which makes sure that the predictions are consistent. For more information regarding the theoretical demonstration for the classifier consistency, please refer to [24].

4.2. Task importance weighting

Let $S_k = \sum_{i=1}^{N} 1\{y_i^k = 1\}$ be the number of instances whose ranks exceed r_k. Note that, by the rank ordering, we have $S_1 \geq S_2 \geq \cdots \geq S_{K-1}$. Let $M_k = \max(S_k, N - S_k)$ be the size of the class with more instances in each binary task. Our importance of the k-th task is defined as the scaled $\sqrt{M_k}$:

$$\lambda^k = \frac{\sqrt{M_k}}{\max_{1 \leq i \leq K-1} \sqrt{M_i}}. \tag{5}$$

Under this weighting scheme, the label imbalance for each binary classification task after extending the original ranks into binary label vectors is taken into account.

5. EXPERIMENTAL EVALUATION

In this section, we evaluate our approach for ordinal SER on the call center dataset, as well as the highly popular IEMOCAP corpus (16 kHz, mono) [26] for reproducible experiments.

5.1. Data preparation

The call center dataset is speaker-independently divided into three partitions with a 8:1:1 split (i.e., 3,655 utterances for the

training set, 458 for the development set, 424 for the test set). We use the development set for hyperparameter tuning and early stopping, and the test set for results reporting. When it comes to the IEMOCAP, since this corpus provides only categorical labels and real-valued attributes originally, a label transformation is required in advance to fit the ordinal SER task (cf., Section 4). Specifically, having a special focus on valence prediction, we discretize the real-valued valence ratings ranging from 1 to 5 (1-negative, 5-positive) to the integer-valued ranks used in call center data annotation: *1-Non-negative* contains ratings in the range (3, 5], *2-Somewhat Negative* contains rating in the range (2, 3], and the *3-Obviously Negative* contains ratings in the range [1, 2]. By this, the set of leveraged 5,531 utterances from 5 emotional categories (i.e., *Angry*, *Excited*, *Happy*, *Neutral*, *Sad*) consists of 1,944 utterances for *1-Non-negative*, 2,022 utterances for *2-Somewhat Negative*, and 1,565 utterances for *3-Obivously Negative*, respectively. Moreover, IEMOCAP signal is downsampled to 8 kHz for alignment with the call center data. Then, we perform a leave-one-speaker-out cross validation scheme on it.

5.2. Setup details

Before feeding into the networks, each utterance is fixed to 8 s by zero padding if shorter, while random cropping if longer. Then, Mel-scaled spectrograms are extracted as network input. We use Librosa [28] for extraction with a window size of 256, a hop size of 128, and a number of Mel bands of 96. Given the 8 kHz sampling rate of raw signal, the network input vector is of shape 96x501 for each utterance. As a baseline system, we use a pre-trained VGGish [25] for classification as shown in Figure 1(a). It contains 4 blocks of convolutional and max pooling layers, 2 fully-connected layers, and 1 softmax output layer with 3 units corresponding to 3 emotional ranks. For more detailed information on the VGGish structure, please refer to [25]. For comparison, our implemented VGGish-CORAL adopts the same structured layers with the VGGish above, except that the output layer has 2 sigmoid units (always 1 less than the number of emotional ranks) and the loss function is CORAL based as described in Section 4. Both, the unweighed average recall (UAR) and the root mean squared error (RMSE), along with the equal error rate (EER) for *Obviously Negative* detection, are used for evaluation.

5.3. Results and analysis

Table 2 summarizes the obtained results from different methods. When integrating the CORAL formulation into VGGish (VGGish-CORAL), one can note that the system performance is generally improved on both the call center dataset and the IEMOCAP corpus. Higher UARs and lower RMSEs suggest that the VGGish-CORAL does not only assign more labels correctly, but also tends to assign numerically closer labels in those incorrect classification cases. This is because, with the CORAL strategy, larger classification errors are inherently given greater penalty during the network training, whereas simply doing categorical classification of the rank-order labels does not have this feature. EERs are further investigated to evaluate the detection of *Obviously Negative*, since the customers with *Obviously Negative* emotion should be treated with more attention in the call center scenario. As observed in Table 2, the VGGish-CORAL method achieves lower EERs than the baseline system. That is, the CORAL strategy reaches a lower false alarm rate while a lower false rejection rate.

Table 2. Performance comparison (UAR [%]: unweighted average recall, RMSE: root mean squared error, EER [%]: equal error rate) between the different methods. IEMOCAP-Val denotes the valence prediction on IEMOCAP. VGGish is the baseline method. VGGish-CORAL is our proposed method. TIW denotes that the Task Importance Weighting strategy is added during network training. Note that the presented EERs correspond to *Obviously Negative* detection.

Methods	**Call Center Dataset**			**IEMOCAP-Val**		
	UAR	**RMSE**	**EER**	**UAR**	**RMSE**	**EER**
VGGish	71.4	0.22	23.1	56.5	0.33	33.3
VGGish-CORAL	72.3	0.18	21.8	57.1	0.26	32.2
VGGish-CORAL-TIW	72.6	0.15	21.2	57.3	0.23	31.6

Moreover, the incorporation of the CORAL and task importance weighting (TIW) strategies gains the best performance on all measurements. For example, the obtained results on IEMOCAP achieve 57.3% on UAR, 0.23 on RMSE, and 31.6% on EER, which all significantly (one-tailed z-test, $p < .05$) outperform the baseline results (i.e., 56.5% on UAR, 0.33 on RMSE, and 33.3% on EER), and slightly outperform the VGGish-CORAL method with uniform task importance weights. This implies that our TIW strategy according to label imbalance can help to boost the classification on imbalanced data distribution.

6. CONCLUSION

In this paper, we present a CORAL based end-to-end modelling approach toward the ordinal SER task where the human's emotions are classified into a fixed number of rank-order labels. Specifically, the well-known categorical classifier VGGish is reformed to an ordinal SER classifier VGGish-CORAL with each output unit designed to deal with a binary SER subtask. Then, the final rank of an utterance can be predicted based on the results of a series of binary SER sub-tasks. With CORAL, the training loss is implemented to be able to reflect rank errors, while, in inference, the consistency of binary classification sub-tasks can be guaranteed. Experimental evaluation is conducted on both a real-world call center dataset and the IEMOCAP corpus. Results show that, comparing to VGGish, the VGGish-CORAL approach is always helpful for performance improvement. Moreover, with a complementary of a task importance weighting strategy based on label imbalance, small additional gains can be obtained. Future work will be the exploration of ordinal SER algorithms additionally considering the temporal evolution of the emotions, therefore, to obtain better estimation on the global order for a sequence of utterances from a same session.

7. REFERENCES

[1] Valery Petrushin, "Emotion in speech: Recognition and application to call centers," in *Proc. International Conference on Artificial Neural Networks in Engineering*, 1999, pp. 7–10.

[2] Laurence Vidrascu and Laurence Devillers, "Real-life emotion representation and detection in call centers data," in *Proc. IEEE International Conference on Affective Computing and Intelligent Interaction*, Beijing, China, 2005, pp. 739–746.

[3] Purnima Gupta and Nitendra Rajput, "Two-stream emotion recognition for call center monitoring," in *Proc. INTERSPEECH*, Antwerp, Belgium, 2007, pp. 2241–2244.

[4] Dimitris Pappas, Ion Androutsopoulos, and Haris Papageorgiou, "Anger detection in call center dialogues," in *Proc. IEEE International Conference on Cognitive Infocommunications*, Györ, Hungary, 2015, pp. 139–144.

[5] Bryan Li, Dimitrios Dimitriadis, and Andreas Stolcke, "Acoustic and lexical sentiment analysis for customer service calls," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 5876–5880.

[6] Wenjing Han, Huabin Ruan, Xiaomin Chen, Zhixiang Wang, Haifeng Li, and Björn W Schuller, "Towards temporal modelling of categorical speech emotion recognition," in *Proc. INTERSPEECH*, Hyderabad, India, 2018, pp. 932–936.

[7] Xixin Wu, Songxiang Liu, Yuewen Cao, Xu Li, Jianwei Yu, Dongyang Dai, Xi Ma, Shoukang Hu, Zhiyong Wu, Xunying Liu, and Helen Meng, "Speech emotion recognition using capsule networks," in *Proc. ICASSP*, Brighton, UK, 2019, pp. 6695–6699.

[8] Mohammed Abdelwahab and Carlos Busso, "Study of dense network approaches for speech emotion recognition," in *Proc. ICASSP*, Calgary, Canada, 2018, pp. 5084–5088.

[9] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5200–5204.

[10] Panagiotis Tzirakis, Jiehao Zhang, and Björn Schuller, "End-to-end speech emotion recognition using deep neural networks," in *Proc. ICASSP*, Calgary, Canada,, April 2018, pp. 5089–5093.

[11] Zhaocheng Huang and Julien Epps, "An investigation of partition-based and phonetically-aware acoustic features for continuous emotion prediction from speech," *IEEE Transactions on Affective Computing (Early Access)*, 2018.

[12] Björn Schuller, Bogdan Vlasenko, Florian Eyben, Martin Wöllmer, André Stuhlsatz, Andreas Wendemuth, and Gerhard Rigoll, "Cross-corpus acoustic emotion recognition: Variances and strategies," *IEEE Transactions on Affective Computing*, vol. 1, no. 2, pp. 119–131, 2010.

[13] Jun Deng, Zixing Zhang, Erik Marchi, and Björn Schuller, "Sparse autoencoder-based feature transfer learning for speech emotion recognition," in *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction*, Geneva, Switzerland, 2013, pp. 511–516.

[14] Zixing Zhang, Bingwen Wu, and Björn Schuller, "Attention-augmented end-to-end multi-task learning for emotion prediction from speech," in *Proc. ICASSP*, Brighton, UK, 2019, IEEE, pp. 6705–6709.

[15] Hector P. Martinez, Georgios N. Yannakakis, and John Hallam, "Don't classify ratings of affect; rank them!," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 314–326, 2014.

[16] Srinivas Parthasarathy, Roddy Cowie, and Carlos 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, 2016.

[17] Reza Lotfian and Carlos Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proc. ICASSP*, Shanghai, China, 2016, pp. 5205–5209.

[18] Zhenghao Jin and Houwei Cao, "Development of Emotion Rankers Based on Intended and Perceived Emotion Labels," in *Proc. INTERSPEECH*, Graz, Austria, 2019, pp. 3277–3281.

[19] Srinivas Parthasarathy, Reza Lotfian, and Carlos Busso, "Ranking emotional attributes with deep neural networks," in *Proc. ICASSP*, New Orleans, USA, 2017, pp. 4995–4999.

[20] Eibe Frank and Mark Hall, "A simple approach to ordinal classification," in *Proc. European Conference on Machine Learning*, Freiburg, Germany, 2001, pp. 145–156.

[21] Wei Chu and S. Sathiya Keerthi, "New approaches to support vector ordinal regression," in *Proc. ACM International Conference on Machine learning*, Bonn, Germany, 2005, pp. 145–152.

[22] Ling Li and Hsuan-Tien Lin, "Ordinal regression by extended binary classification," in *Proc. Advances in Neural Information Processing Systems*, Vancouver, Canada, 2007, pp. 865–872.

[23] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, USA, 2016, pp. 4920–4928.

[24] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka, "Rank-consistent ordinal regression for neural networks," *arXiv:1901.07884v4*, 2019.

[25] Shawn Hershey, Sourish Chaudhuri, Daniel P.W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson, "CNN architectures for large-scale audio classification," in *Proc. ICASSP*, New Orleans, USA, 2017, pp. 131–135.

[26] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335, 2008.

[27] Georgios N. Yannakakis, Roddy Cowie, and Carlos Busso, "The ordinal nature of emotions," in *Proc. ICASSP*, San Antonio, USA, 2017, pp. 248–255.

[28] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in Python," in *Proc. Python in Science Conference*, Texas, USA, 2015, vol. 8, pp. 18–24.