RANKING EMOTIONAL ATTRIBUTES WITH DEEP NEURAL NETWORKS

Srinivas Parthasarathy, Reza Lotfian, and Carlos Busso

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

sxp120931@utdallas.edu, rxl099220@utdallas.edu, busso@utdallas.edu

ABSTRACT

Studies have shown that ranking emotional attributes through preference learning methods has significant advantages over conventional emotional classification/regression frameworks. Preference learning is particularly appealing for retrieval tasks, where the goal is to identify speech conveying target emotional behaviors (e.g., positive samples with low arousal). With recent advances in *deep neural networks* (DNNs), this study explores whether a preference learning framework relying on deep learning can outperform conventional ranking algorithms. We use a deep learning ranker implemented with the RankNet algorithm to evaluate preference between emotional sentences in terms of dimensional attributes (arousal, valence and dominance). The results show improved performance over ranking algorithms trained with *support vector machine* (SVM) (i.e., RankSVM). The results are significantly better than performance reported in previous work, demonstrating the potential of RankNet to retrieve speech with target emotional behaviors.

Index Terms— Emotion recognition, preference learning, ranking emotions

1. INTRODUCTION

Recognizing expressive behaviors is one of the key challenges in *human computer interaction* (HCI) with important applications across domains (i.e., healthcare, education, security, entertainment). Conventionally, automatic emotion recognition systems classify affective behavior into emotional categories using speech, video and other physiological cues [1–3]. They can also recognize emotional behaviors described by attributes such as arousal (calm versus active), valence (negative versus positive) and dominance (weak versus strong) [4, 5]. For emotional attributes, the machine learning tasks usually consist of regression problems, where the task is to predict the value of the attribute [6, 7], or binary or multi class problems, where the goal is to detect discrete levels of these emotional attributes (i.e., low versus high arousal) [8]. An appealing alternative formulation that has received less attention is ranking emotional behaviors using preference learning frameworks [9].

Preference learning is a popular framework in retrieval tasks. Instead of recognizing categories or predicting the values of dependent variables, preference learning determines the relative ranking between samples with respect to a given metric. This problem is commonly formulated as pairwise comparisons between pairs of samples. Preference learning is very appealing for applications in affective computing. For example, it can be used to rank-order aggressive behaviors in surveillance applications, to identify emotional hotspots [10] in longitudinal recordings, and to retrieve target behaviors with given emotions [11]. However, there are few studies on preference learning in affective computing [9, 12–16]. A popular framework in these studies is a ranker trained with *support vectors machines* (Rank-SVM) [17]. Recently, *deep neural networks* (DNNs) have become popular in many speech processing tasks for their superior performance over other statistical methods. This study explores the use of deep learning for preference learning in ranking emotional attributes (arousal, valence, and dominance). To the best of our knowledge, this is the first study that systematically evaluates emotional rankers with DNN.

This study trains rankers using RankNet, which has a DNN architecture [18]. We implement the approach creating a ranker for each emotional attribute (e.g., an arousal ranker). Following the work in Lotfian and Busso [14], we define relative labels for pairwise comparisons by selecting samples with attribute scores separated by a margin. We compare the performance of rankers trained with RankNet with those trained using RankSVM, keeping other parameters constant. We show that rankers trained with RankNet perform significantly better than rankers trained with Rank-SVM in retrieving high and low levels of arousal, valence and dominance. The order of the ranking is also better with RankNet.

2. RELATION TO PRIOR WORK

2.1. Preference Learning in Emotion Recognition

There are few studies that have evaluated preference learning in affective computing. Cao et al. [19, 20] proposed rankers for categorical emotions. For a given emotion, they trained the rankers to determine the order of the samples according to a given emotion (e.g., happy rankers). Using the consensus labels, they formed relative labels by pairing one sample from the target emotion, the preferred one, and a second sample from another emotion. They trained a RankSVM per emotion using these relative samples. They used the results of the rankers as a mid-level representation to recognize discrete emotional categories. Lotfian and Busso [15] proposed a probabilistic framework to define preference learning samples, exploiting individual annotations. Instead of relying on the consensus labels, this framework considered inter-evaluator agreement and intra-class confusion between emotions to determine a relevance score for a given emotion. Then, this relevance score is used to define relative labels for preference learning by setting a margin.

Studies have proposed rankers for emotional attributes. Martinez et al. [9] used a preference learning procedure to learn from continuous emotional attributes, showing that rank based transformations of arousal and valence lead to better generalized models than grouping them into classes. Lotfian and Busso [14] showed the practical considerations for implementing preference learning algorithms on emotion, including the margin required to define relative labels and the number of pairs used for training. Parthasarathy et al. [13] proposed an alternative framework to define relative labels by looking at the trends where raters agree in time-continuous emotional traces. None of these studies used deep learning to train the rankers, which is the main contribution of this study.

This work was funded by NSF CAREER award IIS-1453781.

2.2. Preference Learning with Deep Learning

Deep learning has revolutionized the field of speech processing, achieving performance that are significantly superior to the ones achieved by other competitive approaches. The generic quality of DNNs for learning and feature representation makes DNN an ideal framework for preference learning. However, very few studies have considered ranking with DNNs. Severyn and Moschitti [21] used *convolutional neural networks* (CNNs) to rank short text pairs in retrieval tasks. Burges et al. [18] proposed to use gradient descent to learn ranking functions. We follow this approach in our experimental evaluation. While using gradient descent for ranking has been proposed, to the best of our knowledge, this is the first study that uses DNN to rank emotional attributes with acoustic features.

3. DATABASE

This study relies on two emotional databases with conversational speech, where we train the models with the USC-IEMOCAP [22] corpus, and we test the models with the MSP-IMPROV corpus [23]. The corpus were collected from different speakers providing speaker independent partitions. We intend to create models that generalize by using different databases for training and testing

The USC-IEMOCAP database contains 12 hours of recordings from ten actors who participated in dyadic sessions. Multimodal information was collected: speech, videos, and motion capture data for the face, head and hands of the subjects. This study only uses speech. The dyadic sessions consists of both emotional scripts provided to the actors as well as improvised interactions. These methods elicited naturalistic interactions between actors. Each speaking turn was annotated by two raters, using a 5 likert-scale *self-assesment manikins* (SAMs) for arousal (1-calm, 5-excited), valence (1-negative, 5-positive), and dominance (1-weak, 5-strong). The consensus labels is the average of the scores, which range from 1 to 5. Busso et al. [22] gives the details of the corpus.

The MSP-IMPROV corpus includes dyadic interactions from 12 actors (6 male, 6 female). The key feature of this corpus is the control of lexical content over different emotions while preserving the perception of naturalness. This goal was achieved by (1) defining 20 target sentences, and (2) creating emotion-dependent scenarios per target emotion, which lead one of the actors to speak the target sentences. This approach allows actors to express emotions as dictated by the context of the scenarios, creating expressive behaviors that are perceived more natural than read renditions of the target sentences [23]. In addition to the target sentences, the corpus includes all the speaking turns of the improvisation and the natural interaction between actors during the breaks. The corpus has 8438 speaking turns, which are emotionally annotated using a novel framework for crowdsourcing, which tracks the performance of the evaluators in real time, stopping the evaluation when the performance dropped below a predefined threshold [24]. Similar to the USC-IEMOCAP corpus, the MSP-IMPROV is evaluated for primitive attributes by at least five evaluators using 5 likert-scale SAMs for arousal, valence and dominance. We average the scores, which range from 1 to 5. Busso et al. [23] gives the details of the corpus.

4. RANKING EMOTIONAL ATTRIBUTES

4.1. RankNet

While DNNs have been popular for their ability to learn and represent data, we are not aware of any study that has explored their benefits in ranking emotions. Burges et al. [18] introduced RankNet, an algorithm that uses gradient descent to perform preference learning between pairs of samples. The neural network model in RankNet is given by the function f, which maps the feature vector $\mathbf{\Phi}$ into a score $f(\mathbf{\Phi})$. Given two samples U_i and U_j with feature vectors $\mathbf{\Phi_i}$ and $\mathbf{\Phi_j}$, the probability that U_i is preferred over U_j is given by P_{ij}. This probability is mapped by the sigmoid function:

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

where $s_i = f(\mathbf{\Phi_i})$ and $s_j = f(\mathbf{\Phi_j})$. During training, the ideal probability $\bar{P}_{ij}$ is set according to the relative labels between pairs of samples, where $\bar{P}_{ij}$ = 0 implies that U_j is preferred over U_i, and $\bar{P}_{ij}$ = 1 implies that U_i is preferred over U_j. The cross entropy is then applied as a cost function to measure the deviation of the model output P_{ij} from the ideal probability $\bar{P}_{ij}$:

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

Equation 2 simplifies to $C = log(1+e^{-\sigma(s_i-s_j)})$ when $\bar{P}_{ij}$ = 1, and to $C = log(1+e^{-\sigma(s_j-s_i)})$ when $\bar{P}_{ij}$ = 0. Therefore, the cost function is symmetrical on whether U_i is preferred over U_j or vice versa. While the original implementation has an option to include samples that are equivalent (e.g., $\bar{P}_{ij}$=0.5), we force RankNet to provide a preferred sample. We implemented a feed forward DNN architecture, consisting of two hidden layers of size 256 each. We use sigmoidal units, following Equation 1, with σ = 2. A learning rate of 10^{-6} was used for a maximum of 100 epochs. The DNN was implemented using the CNTK toolkit [25].

4.2. RankSVM

In Lotfian and Busso [14], we evaluate a RankSVM framework to recognize emotional attributes. For comparison, we include results from RankSVM as baseline. RankSVM, introduced by Joachims [17], uses SVM for preference learning. Let the set P denote the preference learning training set. The pair $(i,j) \in P$ if the i-th sample is preferred over the j-sample. If $\mathbf{\Phi_i}$ and $\mathbf{\Phi_j}$ are the feature vectors for the i-th and j-th samples, then the optimization problem is formulated a:

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i,j\in P}\xi_{i,j} \\ &s.t. \langle w, (\Phi_i - \Phi_j)\rangle \geq 1-\xi_{i,j}, \xi_{i,j} \geq 0 \qquad \text{for } i,j \in P \end{aligned} \tag{3}$$

where ξ represents the slack variable and C the soft margin variable. The optimum weight vector $\hat{w}$ minimizes Equation 3, which is equivalent to maximize the margin of the support vectors. With $\hat{w}$, the model can be tested by measuring $\langle \hat{w}, (\Phi_i - \Phi_j)\rangle$. If $\langle \hat{w}, (\Phi_i - \Phi_j)\rangle \geq 0$ then i-th is preferred over j-th sample. Therefore, this formulation for pairwise ranking algorithm reduces to a binary classification problem once the features of the two samples are subtracted (i.e., $\Phi_i - \Phi_j$).

4.3. Relative Labels for Emotional Attributes

A major drawback of collecting relative labels from scratch is the computational complexity involved. For N training samples, we would need $N \times (N-1)/2$ comparisons for pairwise labels. Therefore, we rely on the existing annotations to form our relative labels. The perception of emotional behaviors varies across raters, producing labels that are noisy. To increase the reliability of training pairs, we create relative labels by selecting pairs of samples that are separated by at least a margin t. There is a tradeoff between the reliability of the relative labels and the size of the training set. As the margin increases, the size of the corpus decreases since fewer samples satisfy the requirement. The evaluation considers different values for t (Sec. 6.1).

We consider the evaluation in Lotfian and Busso [14] to select the number of pairs for training, which indicated that the performance of rankers saturates after 10,000 training pairs. Therefore, we randomly select 10,000 pairs of samples that satisfy this margin.

5. EXPERIMENTAL EVALUATION AND RESULTS

This section describes the acoustic features used to train the rankers (Sec. 5.1), and the experimental setting (Sec. 5.2).

5.1. Acoustic Features

Emotion affects different aspects of speech production. As a result, it is common to consider multiple acoustic features for emotion recognition. Previous studies have proposed several feature sets that include thousand of features [26]. This high dimension feature vector is usually reduced using feature selection. While this process optimizes performance of the classifiers, the results are hard to replicate by other researchers, since the individual features in the selected feature set are not reported. Recently, Eyben et al. [27] introduced the *Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) as a set of standard features for affective computing. The GeMAPS set consists of a set of minimalistic audio features that were selected based on their performance in previous studies, ease of automatic extractability, and theoretical significance. Having this reduced set of features provides a platform for reproducing research, eliminating the variation caused by using different feature sets.

This study uses the extended parameter set (eGEMAPS). The common extraction procedure involves extracting *low-level descriptors* (LLDs) which are features extracted on a frame-by-frame basis.Then, a set of global functionals (eg. arithmetic mean) are applied on the LLDs to extract *high level features* (HLFs). This set has 88 features, which are extracted with the OpenSMILE toolkit [28]. Eyben et al. [27] gives the details of the feature set.

5.2. Experimental Settings

As mentioned in Section 3, the USC-IEMOCAP database is used for training the models, and the MSP-IMPROV is used for testing the models. An important parameter in defining the relative labels for the rankers is the margin needed to consider that one sample is preferred over other. For RankNet, we evaluate $t \in \{0, 1, 2, 3\}$. Notice that the range of the annotations is from 1 to 5 for the three emotional attributes (we present the results in Section 6.1). For RankSVM, we rely on the parameters selected in Lotfian and Busso [14]. In that study, we evaluated practical consideration for preference learning using RankSVM, including optimal values for this margin. The study considered the SEMAINE database, where the labels range from -1 to 1. The optimal margin was $t = 0.5$ for arousal and $t = 0.45$ for valence (the annotations for the SEMAINE database does not include dominance). Since the range in both corpora considered here is from 1 to 5, we set the margin at $t = 1.0$ for arousal and $t = 0.9$ for valence. For dominance, we set the margin at $t = 1.0$ since arousal and dominance are commonly correlated.

In addition to the two ranking algorithms (RankNet, RankSVM), we train a regression model with DNN, which we refer to as DNNRegression. The model predicts the values of the attributes for each testing sentence, which are then sorted creating a ranked list. For consistency, we also used two hidden layer, feed forward architecture with 256 nodes each. The nodes are activated by a sigmoidal function.

We evaluate the performance with *precision at k* (P@k), which is commonly used in retrieval tasks. P@k measures the precision in retrieving k% of the samples. Let's consider valence as an example. First, we estimate the median of the valence scores in the test set. Samples with scores below the median are considered as "low valence", and samples above the median are considered as "high valence". Second, we rank the testing samples using the ranker. We consider the top $k\%$ of the list and the bottom $k\%$ of the list. Finally, we consider success if the retrieved samples at the top of the list are from the "high valence" group, and the samples at the bottom of the list are from the "low valence" group. For example, P@50 includes all the samples (50% from the top, and 50% from the bottom of the list). The advantage of this approach is that we can easily compare rankers, regression models, and binary classifiers (this study only considers rankers and regression models).

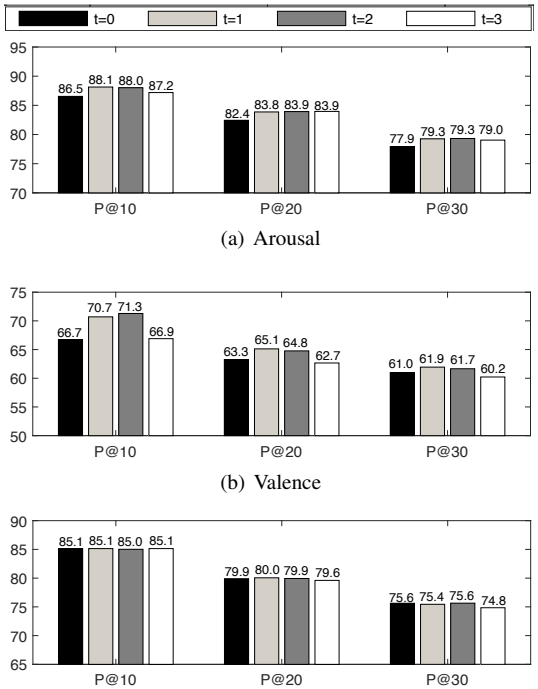


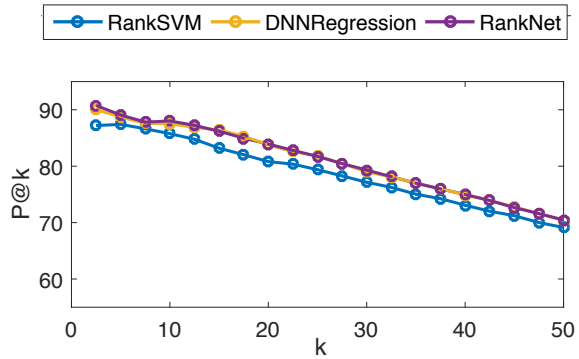
Fig. 1. Evaluation of the margin t for RankNet. P@10, P@20, P@30 for different $t \in \{0, 1, 2, 3\}$.

We also evaluate the results with the Kendall's Tau coefficient, which measures the correlation between two ordered lists. The value for τ varies between [-1,1], where -1 indicates reverse order, and 1 indicates perfect order. Unlike P@k, which assumes a binary split to define success, τ was calculated considering the ordinal ranking of all testing samples.

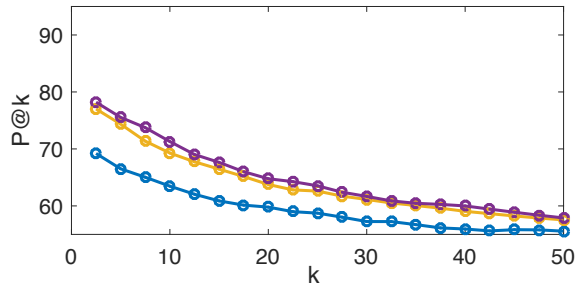
6. RESULTS

6.1. Performance of RankNet for Different Margins

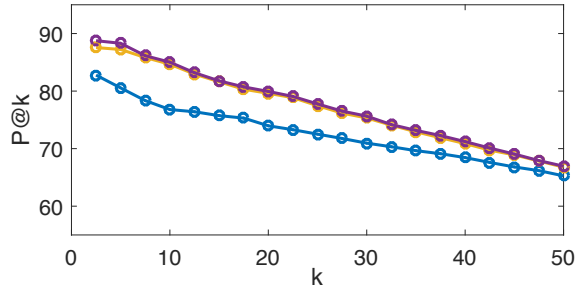
The margin between samples to form the relative labels for training plays an important role in the performance of the ranking algorithm [14]. Figure 1 shows the performance of the rankers for P@10, P@20, P@30 in terms of the margin t. We evaluate $t \in \{0, 1, 2, 3\}$ for arousal (Fig. 1(a)), valence (Fig. 1(b)) and dominance (Fig. 1(c)). Notice that the range in the attributes is from 1 to 5, so the maximum separation between two samples is 4. The figures show similar performance for $t = 1$ and $t = 2$, which indicate that the approach is not sensitive to the margin t within this range. For other values of t, the performance drops. We implement the rest of the evaluation with $t = 2$ for RankNet.



(a) Arousal



(b) Valence



(c) Dominance

Fig. 2. Precision at k for RankNet, RankSVM and DNNRegression.

6.2. Comparison with Other Methods

Figure 2 gives the performance for arousal (Fig. 2(a)), valence (Fig. 2(b)), and dominance (Fig. 2(c)). The figures include the two ranking schemes, RankNet and RankSVM, and the DNN regression model. When we compare the ranking algorithms, the results demonstrate that RankNet outperforms RankSVM by a large margin. This result is clearly observed for arousal, valence and dominance. When we compare regression (DNNRegression) and ranking (RankNet) frameworks built with deep learning, the results are not conclusive. The results are very similar, although the performance for RankNet is always slightly better. The separation between RankNet and DNNRegression increases for valence. In Lotfian and Busso [14], the results for RankSVM was significantly better than *support vector regression* (SVR). These differences are attenuated when ranking and regression frameworks are trained with deep learning.

Table 1 lists the performance of RankSVM, RankNet and DNNRegression at three values of k: P@10, P@20 and P@30. We compare whether the difference between conditions are statistically significant, using the one tailed z-test on the difference in population proportions, asserting significance at p-value=0.01. The

Table 1. Precision@k for k =10%, 20%, 30%.

	RankSVM	RankNet	DNNRegression
Arousal			
P@10	85.77	88.02	87.54
P@20	80.81	83.93	83.72
P@30	77.15	79.32	79.02
Valence			
P@10	63.46	71.29	69.28
P@20	59.79	64.77	63.76
P@30	57.26	61.66	61.13
Dominance			
P@10	76.79	86.15	84.67
P@20	73.97	79.94	79.61
P@30	70.95	75.65	75.33

Table 2. Kendall's Tau coefficient to assess order of the ranked list.

	RankSVM	RankNet	DNNRegression
Arousal	0.36	0.41	0.41
Valence	0.08	0.14	0.13
Dominance	0.28	0.35	0.34

table shows that RankNet and DNNRegression algorithms perform significantly better than RankSVM (p-value < 0.01), for all cases, with the exception of P@10 for arousal (see Fig. 2(a)). Even though rankers trained with RankNet perform better than DNNRegression, the differences are not significant (p-value < 0.01). We obtain an absolute gain of 7.83% for P@10 in valence, when we use RankNet instead of RankSVM. Previous studies have shown that valence is a challenging emotional attribute for machine learning frameworks trained with speech features [29]. We significantly improve our performance for valence when we use RankNet (see Fig. 2(b)). We also observe an increase in performance for dominance, when using RankNet. The improvement over RankSVM is 9.36% (absolute), which is the highest performance gain over RankSVM observed in this evaluation.

Table 2 shows the values for Kendall's Tau for RankNet, RankSVM, and DNNRegression. The table shows that RankNet provides the best performance. We analyze the statistical significance of the differences in the results after applying the Fisher transform to normalize the values to follow a Gaussian distribution. RankNet and DNNRegression algorithms are significantly better than RankSVM (p-value < 0.001). The differences between RankSVM and DNNRegression are not significant, confirming the results shown in Table 1. Notice that the values in Table 2 are significantly better than the results reported in related work [9].

7. CONCLUSIONS

The study showed the benefits of using deep learning in ranking emotional attributes. We utilized RankNet, which uses the gradient descent for pairwise comparisons between samples. With cross-corpora evaluations, the results showed that RankNet algorithm perform significantly better than RankSVM. For valence, the results also showed that RankNet gives better performance than a DNN-based regression algorithm.

In our future work, we will implement DNN rankers for categorical emotions using the relative labels defined in Lotfian and Busso [15] (probabilistic relevant scores). Likewise, we will evaluate temporal evolution of the emotions by using *recurrent neural networks* (RNNs). This approach will allows us to consider contextual information for each of the pairwise comparisons.

8. REFERENCES

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Sixth International Conference on Multimodal Interfaces ICMI 2004*, State College, PA, October 2004, pp. 205–211, ACM Press.

[2] C.M. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S.S. Narayanan, "Emotion recognition based on phoneme classes," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 889–892.

[3] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, HI, USA, April 2007, vol. 4, pp. 941–944.

[4] A. Mehrabian and J. Russell, *An approach to environmental psychology*, The MIT Press, Cambridge, MA, US, June 1974.

[5] J.A. Russell, "Evidence of convergent validity on the dimensions of affect," *Journal of Personality and Social Psychology*, vol. 36, no. 10, pp. 1152–1168, October 1978.

[6] M. Grimm, K. Kroschel, and Shrikanth Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007)*, Honolulu, HI, USA, April 2007, vol. 4, pp. 1085–1088.

[7] M.A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, April-June 2011.

[8] S. Mariooryad and C. Busso, "The cost of dichotomizing continuous labels for binary classification problems: Deriving a Bayesian-optimal classifier," *IEEE Transactions on Affective Computing*, vol. To appear, 2017.

[9] 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.

[10] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 3598–3602.

[11] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[12] Y.-H. Yang and H.H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 762–774, May 2011.

[13] 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.

[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] R. Lotfian and C. Busso, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 490–494.

[16] M. Soleymani, , G. Chanel, J.J.M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *ACM Workshop on Multimedia Semantics*, Vancouver, British Columbia, Canada, October 2008, pp. 32–39.

[17] T. Joachims, "Optimizing search engines using clickthrough data," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2002)*, Edmonton, Alberta, Canada, July 2002, pp. 133–142.

[18] 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.

[19] H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech," in *Interspeech 2012*, Portland, Oregon, USA, September 2012, pp. 358–361.

[20] 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 2014.

[21] A. Severyn and A. Moschitti, "Learning to rank short text pairs with convolutional deep neural networks," in *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2015)*, Santiago, Chile, August 2015, pp. 373–382.

[22] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[23] 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. To appear, 2017.

[24] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 374–388, October-December 2016.

[25] A. Agarwal, E. Akchurin, C. Basoglu, G. Chen, S. Cyphers, J. Droppo, A. Eversole, B. Guenter, M. Hillebrand, X. Huang, Z. Huang, V. Ivanov, A. Kamenev, P. Kranen, O. Kuchaiev, W. Manousek, A. May, B. Mitra, O. Nano, G. Navarro, A. Orlov, M. Padmilac, H. Parthasarathi, B. Peng, A. Reznichenko, F. Seide, M.L. Seltzer, M. Slaney, A. Stolcke, H. Wang, K. Yao, D. Yu, Y. Zhang, and G. Zweig, "An introduction to computational networks and the computational network toolkit," Technical Report MSR-TR-2014-112 (DRAFT v1.0), Microsoft Research, Redmond, WA, USA, February 2016.

[26] C. Busso, M. Bulut, and S.S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds., pp. 110–127. Oxford University Press, New York, NY, USA, November 2013.

[27] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "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.

[28] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[29] 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.