INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Development of Emotion Rankers Based on Intended and Perceived Emotion Labels

Zhenghao Jin, Houwei Cao

Department of Computer Science, New York Institute of Technology, New York, 10023, USA
zjin10@nyit.edu, hcao02@nyit.edu

Abstract

In emotion datasets, intended emotion labels and perceived emotion labels both contain valuable information about how human express and perceive emotions, and there is a considerable mismatch between the two. In this paper, we propose a novel method to derive relative labels for preference learning using both the intended labels during emotion expression and the perceived labels given by all raters during perceptual evaluation. Based on analyzing the agreement between the intended and perceived labels, as well as the consistence among all perceptual ratings, we propose three pairwise ranking rules to generate multi-scale relevant scores for preference learning. We further build three sets of rankers for six basic emotions based on the three ranking rules. Through evaluation on the CREMA-D database, we demonstrate that, by considering both intended and perceived labels, our proposed rankers significantly outperform the rankers only relying on the perceptual ratings. We further combine the ranking scores of individual emotions for multi-class classification. Through experiments, we show that the emotion classification systems with ranking information significantly outperform the conventional SVM classifiers.

Index Terms: emotion recognition, preference learning, intended emotion, perceived emotion

1. Introduction

Emotions are essential to human life. They directly influence human perception and behaviors, and have big impacts on our daily tasks, such as learning, social interaction, and rational decision-making. Automatic emotion recognition has found applications in many domains, including multimedia retrieval, human-computer & human-robot interaction, etc [1, 2, 3, 4]. It has also been used in the diagnosis of many neurological, neuropsychiatric diseases and mental health conditions. The traditional paradigm of emotion recognition in speech is to extract acoustic features from the speech signal, then train classifiers on these representations, which can be applied to a new utterance to determine its emotion content. A variety of pattern recognition methods have been explored for automatic emotion recognition, such as Gaussian mixture models [5], Hidden Markov models [6], support vector machines [7], regression [1], and deep neural network [8], etc.

An alternative approach is ranking emotional behaviors, which offer a way of sorting all utterances in a given sample of speech with respect to the degree with which they convey a particular emotion. Although ranking frameworks have been widely used in many information retrieval applications in text, image, video, and music [9, 10, 11, 12], they have been applied to speech emotion recognition since very recently, e.g. [13, 14, 15, 16, 17, 18]. Cao et al. [17] first trained rankers by establishing a binary preference score based on the consensus labels. For example, for a ranker for *Anger* emotion, a sample labeled as *Anger* was always preferred over another sample labeled with other emotions. Lotfian et al. [15] considered the perceptual ratings from all individual evaluators to create a continuous relevance score for preference learning. Parthasarathy et al. [16] first applied qualitative agreement (QA) methods to estimate reliable labels from noisy annotations, then used those labels for preference-learning.

Most of the existing datasets are generated in two-stages: at first, emotion expressions of actors guided by intended emotion labels are recorded; crowd-sourced raters then assign labels to the recorded samples based on their emotion perception. Both stages contain valuable information about how human express and percept emotions, which should be collectively mined to improve the accuracy of emotion recognition. *In this paper, we propose a novel method to derive relative labels for preference learning using both the intended labels during emotion expression and perceived labels given by all raters during perceptual evaluation.* We assume that samples that have the consistent target emotion during the emotion expression and perception are more likely to convey the target emotion. Based on analyzing the agreement between the intended and perceived labels, as well as the consistence among all perceptual ratings, we propose three pairwise ranking rules to generate multi-scale relevant scores for preference learning. We further build three sets of rankers for six basic emotions based on the three ranking rules. Through evaluation on a large dataset, we demonstrate that, by considering both intended and perceived labels, our proposed rankers significantly outperform the rankers only relying on the perceptual ratings. One ranking rule dominates the other two in all experiments. We further combine the ranking scores of individual emotions for multi-class classification. Through experiments, we show that the emotion classification systems with ranking information significantly outperform the conventional SVM classifiers.

2. Dataset & Features

2.1. CREMA-D

We use the emotional speech from the *Crowd-sourced Emotional Multimodal Actors Dataset* (CREMA-D) [19]. The CREMA-D database is an audiovisual corpus collected to explore human emotion expression and perception behaviors in different modalities. It consists of facial and vocal emotional expressions in sentences spoken in a range of basic emotional states (*Anger, Disgust, Fear, Happiness, Neutral*, and *Sadness*). This corpus consists of $7,442$ clips (over 10 hours) of emotional sentences collected from 91 actors with diverse ethnic backgrounds. The task for the actors was to convey that they are experiencing a target emotion while uttering a given sentence. The intended emotion label is the target emotion given to the actors during recording. The categorical emotion labels and real-valued intensity values for the perceived emotion were

http://dx.doi.org/10.21437/Interspeech.2019-1831

also collected through crowd-sourced perceptual evaluations from $2,443$ raters in three modalities: audio, visual, and audio-visual. More than 95 percent of the clips in the database have 8 to 12 perceptual ratings. This study focuses on speech emotion analysis. Therefore we only consider the perceived emotion labels based on audio perceptual evaluations. By considering the agreement between the intended and perceived emotion labels, the dataset can be divided into three subsets of *Matching*, *Non-matching*, and *Ambiguous*. Both *Matching* and *Non-matching* subsets consist of unambiguous clips with a consensus group-perceived emotion identified by the majority of raters. For clips in the *Matching* subset, the group-perceived emotion matches the intended emotion, while for clips in the *Non-matching* subset, the group-perceived emotion differs from the intended emotion. Clips in the *Ambiguous* subset do not have consensus group-perceived emotions using majority vote. Table 1 shows the number of samples per emotion in the three subsets. Further information about the database is provided in Cao et al [19].

Table 1: *Number of utterances from each emotion category in three subsets of Matching: the group-perceived emotion matches the intended emotion, Non-matching: the group-perceived emotion differs from the intended emotion, and Ambiguous: no group-perceived emotions with majority vote. (Anger (A), Disgust (D), Fear (F), Happiness (H), Neutral (N), and Sadness (S).)*

	A	D	F	H	N	S	Total
Matching	770	343	407	330	1040	209	3099
Non-matching	389	793	741	811	26	939	3699
Ambiguous	112	135	123	130	21	123	644
Sum	1271	1271	1271	1271	1087	1271	7442
Matching %	60.6	30	32	26	95.7	16.4	41.6

2.2. Acoustic Features

In this study, we use the feature set provided for the emotion challenge at INTERSPEECH 2009 [20]. This comprehensive set of acoustic features includes 988 High Level Descriptors (HLDs) extracted using OpenSMILE feature extraction library [21]. The set also includes Low Level Descriptors (LLDs) such as prosodic, spectral and voice quality features, from which we estimate High Level Statistical Functionals (HSFs) at the utterance level, such as minimum, maximum, mean, and variance. We use these features for all experiments reported in later section. Detailed description of the features is given in Schuller et al. [20]

3. Methodology

Generally speaking, in machine learning tasks, the learned models can be considerably affected by the training data and the labels assigned to them. Here, we are interested in building rank-based classifiers for emotion recognition. The ranking problem is to sort the utterances with respect to how much they convey a particular emotion. To train a ranker for a target emotion, we need to specify a set of pairs of instances for which one instance conveys the target emotion better than the other. The optimization problem of the ranker is to minimize the number of incorrectly ordered pairs.

3.1. Creation of Relevance Scores for Ranking

The key challenge here is to define the relative labels with pairwise preference. For example, how to establish whether a sample is more anger than another plays an important role in building a reliable anger ranker. Different from many existing methods that only consider either individual consensus target emotion labels or individual perceptual ratings [15, 18], the proposed method considers both the intended emotion labels given to the actors during the recording and the perceived emotion labels given by all individual raters during the perceptual evaluations. The relevance score is calculated based on the agreement between the intended and perceived emotions.

The intended label can be considered as the target emotion during the emotion production/expression, and the perceived label (if majority consensus exists) is the the target emotion during the emotion perception. In Table 1, the matching ratios between the intended and perceived emotion labels vary from 16.4% to 95.7% cross different target emotions in the CREMA-D database. Overall, only 41.6% of the clips have matched intended and perceived emotions, while for the rest of clips the target emotion that the actors intended to convey could not be successfully perceived by the majority of the raters during the acoustic perceptual evaluations. The intuition behind our method is that samples that have the consistent target emotion during the emotion expression and perception are more likely to convey the target emotion. For example, we can assume a sample with [*intended, perceived*] labels of [A, A] is more *Anger* than another sample with labels [A, N], which the majority of raters perceived as *Neutral*. Furthermore, samples that are consistently evaluated with a consensus emotion among raters are more likely to convey that emotion. Therefore, a samples with labels [A, A] is expected to be more *Anger* than a sample with labels [A, Ambiguous], for which there is no consensus group-perceived emotion, and a sample with labels [A, Ambiguous] is expected to be more *Anger* than a sample with labels [A, F], which is expected to be more *Fearful*.

We investigate several alternatives for defining the partial ordering for ranking. More specifically, for any target emotion $x \in \mathcal{E} \triangleq \{A, D, F, H, N, S\}$, the [*intended, perceived*] labels of a clip can fall into one of the following categories:

- matching: $[x, x]$;
- ambiguous: $[x, -]$, where '$-$' stands for *no consensus*;
- non-matching: $[x, y]$, where $y \in \mathcal{E} - \{x\}$;
- perceived matching: $[z, x]$, where $z \in \mathcal{E} - \{x\}$;
- others: $[z, y]$, where $z \in \mathcal{E} - \{x\}, y \in \mathcal{E} - \{x\} + \{-\}$.

We can further calculate a continuous score r_x for the clip as the percentage of the raters who gave label 'x'. Based on the labels and scores, we develop three sets of rankers for a target emotion x based on different pairwise ranking rules:

1. $[x, x] > [x, -] > [x, y] = [z, x] > [z, y]$;
2. $[x, x] > [z, x] > [x, -] > [x, y] > [z, y]$;
3. *rank based on the continuous score* r_x.

Rule 1 and Rule 2 give exactly the same ordering for *Matching*, *Ambiguous* and *Non-matching*: $[x, x] > [x, -] > [x, y]$. The only difference is how they treat *Perceived Matching*, $[z, x]$. Rule 1 treats it the same as *Non-matching* (i.e., *Intended Matching*, $[x, y]$), while Rule 2 assigns it higher preference than *Ambiguous* and *Non-matching*. For example, based on Rule 2, a sample with labels [F, A] is expected to be more *Anger* than a sample with labels [A, F], while they are considered to convey the same level of anger by Rule 1. Rule 3 defines the partial ordering purely based on the perceptual rating, where the continuous score reflects how consistent the utterance can be perceived as the target emotion. Intuitively, a sample with all raters gave label 'A' is more *Anger* than a sample with 90% of the raters labelled it 'A'.

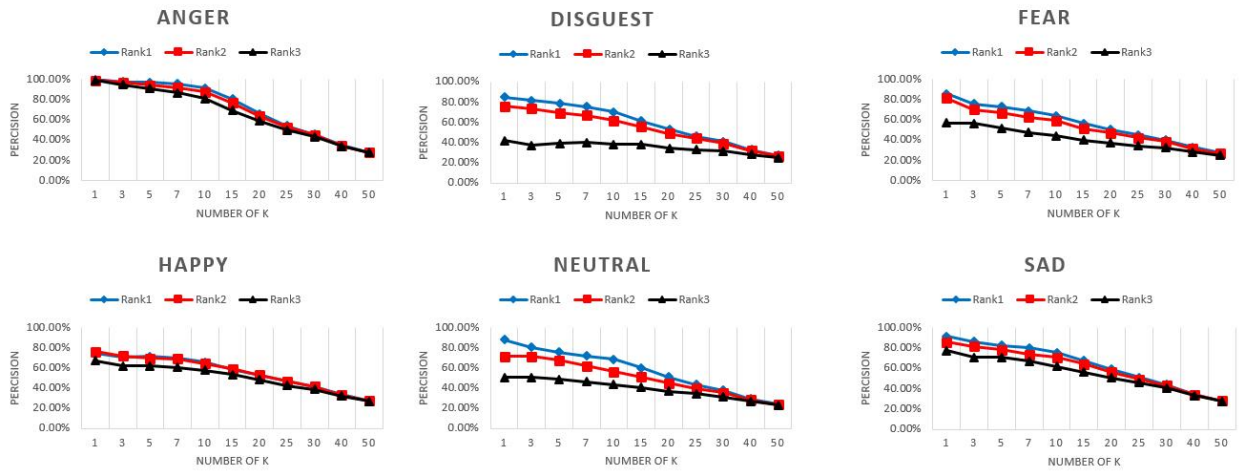

Figure 1: *Precision at k along with K of different sets of emotion rankers trained with different relevant scores on CREMA-D datasets.*

3.2. Development of Emotion Rankers

In this study, we make use of the SVMrank toolkit to train and test our approach [22]. Ranking support vector machines (SVM) is a classical pairwise method for designing ranking models. The basic idea behind them is to formalize learning to rank as a problem of binary classification on pairs that define a partial ordering and then to solve the problem using SVM classification [23]. Similar to [17, 18], we choose to form pairs only from utterances generated by the same speaker and assign relevant scores to all utterances based on how they convey the target emotion as detailed in Section 3.1. The motivation for this approach is the same as that for using ranking SVMs for ranking in information retrieval. There the task is to sort webpages returned by a search engine based on the relevance to the query. In our task, we want to sort all the utterances generated by a speaker based on their relevances for each target emotion $x \in \mathcal{E}$.

We use the three ranking rules developed in in Section 3.1 to generate three sets of relevance scores to train emotional rankers using the CREMA-D database. For each set of relevance scores, six speaker-independent emotional rankers were constructed, one for each basic emotion. In testing, all utterances generated by a speaker whose data were not used in training is given to the ranker for a target emotion. The ranker produces a ranking score for each test utterance, allowing us to sort the utterances by decreasing score. Utterances with higher ranks are considered to express the target emotion more clearly than utterances with lower ranks.

3.3. Ranking-based Multi-class Classification

Given a sample of speech, an emotion ranker developed for a particular emotion x indicate how relevant each utterances is to emotion x. However, the rankers do not directly give a way to decide which particular emotion is expressed by a given utterance. In order to classify the unknown test utterance as expressing one of the basic emotions, we need to combine ranking scores from rankers developed for all emotions. Here, we implemented a simple rule-based approach to combine the ranker scores into a final classification. The emotion of an utterance is decided by directly comparing the ranks assigned by the rankers for the six basic emotions. The utterance is classified as conveying the emotion for which it achieved the highest rank among all the emotions. If an utterance had the same rank assigned by more than one rankers, a decision about which emotion to pick was made randomly.

4. Experimental Results

To evaluate the stability and speaker independence of the developed rankers, we performed all experiments using leave-one-subject-out (LOSO) scheme. In this form of cross-validation, all samples from a given speaker are used as a test set for a model trained on the data from all the other speakers and the process is repeated for all speakers. The overall performance is computed by combining the results from all test folds and computing the overall accuracy for the entire dataset.

4.1. Evaluation of Emotion Rankers

We first analyze the performance of the rankers for each of the six basic emotions. We first look at Precision at k, which is widely used to evaluate the performance of ranking models. It is defined as the percentage of the top ranked k utterances generated by a ranker for a target emotion that were indeed the utterances with the target emotion. Here, we consider the intended emotion labels as the ground-truth. To demonstrate the ranker's performance at various levels, Fig. 1 shows the precision at k for different k and for three sets of rankers trained with different pairwise ordering rules. A perfect ranker will attain 100% precision rate for k smaller than the number of the target emotions in each LOSO fold, then drop steadily.

As illustrated in Fig. 1, the performance of ranker sets 1 and 2 are significantly better than the ranker set 3 for all the six basic emotions. The partial orderings in ranker sets 1 and 2, taking into account the consistence of the intended and perceived labels, tend to produce more accurate rankers than the orderings only based on the agreement of perceptual ratings in ranker set 3. It is also noticed that the best performance is achieved by the ranker set 1 (blue lines) using the labels derived from the pairwise ordering defined in Rule 1. Those labels do not assign any preference ordering between inconsistent labels and perceived matching labels, while ranker set 2 always assign higher prefer-

ence to perceived matching labels than inconsistent labels.

Table 2: *R-precision (%) (R = no. of target emotions) and P@K (precision (%) at K for which we successfully retrieved all target emotions) from different sets of emotional rankers on CREMA-D datasets.*

	R-Precision(%) (R = no. of target emotions)			P@K(%) (all target emotions retrieved)		
	Rank1	Rank2	Rank3	Rank1	Rank2	Rank3
ANG	83.29	79.70	71.29	66.62	57.72	45.48
DIS	63.06	56.78	38.53	35.65	32.09	24.62
FEA	57.72	52.85	40.77	37.24	32.55	24.74
HAP	59.81	59.73	54.23	38.24	37.98	31.63
NEU	65.14	54.11	41.79	43.63	35.15	28.44
SAD	68.99	65.37	56.96	48.88	43.39	36.47
Ave.	66.32	61.42	50.59	45.04	39.81	31.89

In an idealized situation where we know the number of utterances that convey the target emotion for each speaker (n), we can measure the R-precision at different R for different speakers, with $R = n$ for that speaker. In other words if we knew that a speaker said 12 utterances in an angry manner, we can look at the precision at 12 for that speaker and see how many of the top ranked twelve utterances were indeed anger utterances. This would be equivalent to viewing the ranker as a binary classifier, where we assumed that the top n utterances are the ones expressing the target emotion and the corresponding precision at R (R-precision) can be referred to as the one-versus-all classification accuracy with the prior knowledge of the distribution of emotions. Table 2 lists our oracle results in terms of R-precision on different sets of rankers. The average R-precision for ranker set 1, set 2, set 3 are 66.32%, 61.42% and 50.59% respectively. The results reaffirm that ranker sets 1 consistently achieve significant higher R-precision for all the six basic emotions.

On the other hand, in Table 2 we also report the precision (%) at K for which we successfully retrieved all target emotions. As we expected, the ranker set 1 outperforms the others, and achieves the average precision of 45.04%. The CREMA-D dataset contains about 84 emotional utterances (14 for each emotion) for each speaker. This means that the proposed ranking system can always retrieve all target emotions in the top (e.g., top 30) utterances. The result is very promising.

4.2. Multi-class Classification of Emotion

Now we combine the ranking scores for multi-class classification. The accuracy of speaker-independent, multi-class classification of the three ranker sets is shown in Table 3. In order to better understand the results, we report the overall performance on the entire CREMA-D datasets, as well as the performance on the three different subsets of matching, non-matching and ambiguous data. The overall multi-class classification performance for ranker 1, ranker 2 and ranker 3 are 68.66%, 64.05% and 51.82% respectively. This consistently demonstrates the advantages of the pairwise ranking Rule 1. It is also interesting to note that the performance gain are significantly higher in the non-matching and ambiguous data subsets where the intended labels are different from the perceived ones.

Finally, in order to investigate the usefulness of ranking SVM in emotion recognition, we compare the performance of the ranking based multi-class classifiers to the results of the conventional SVM classifiers. We trained baseline SVM classifiers with radial basis kernels constructed with the LIBSVM toolkit [24] with the same acoustic features used in the ranking based classifiers. Three baseline classifiers are trained based on (1) the entire dataset with intended labels, (2) unambiguous clips (matching & non-matching) with the perceived labels, and (3) matching subset only. Similarly, we list the accuracy results for the entire CREMA-D datasets as well as the three subsets of matching, non-matching and ambiguous in Table 4. The overall performance for the standard multi-class classifiers is much lower than that of the proposed ranking-based classifiers. The absolute degradation is as high as 12.38%, from the best ranker 1 classifiers to the best SVM classifiers trained on the entire dataset with intended labels. The same trend can be observed on the three different subsets as well.

Table 3: *Speaker-independent, multi-class emotion classification accuracy for six emotion task on the entire CREMA-D datasets, as well as matching, non-matching and ambiguous subsets when we directly compare ranking SVM scores.*

	Overall	Matching	Non-matching	Ambiguous
Rank1	**68.66%**	**78.76%**	**60.23%**	**68.47%**
Rank2	64.05%	76.28%	54.52%	59.93%
Rank3	51.82%	66.28%	40.74%	45.96%

Table 4: *Speaker-independent, multi-class emotion classification accuracy for six emotion task on the entire CREMA-D datasets, as well as matching, non-matching and ambiguous subsets when we train the conventional SVM multi-class classifier based on (1) entire dataset with the intended labels, (2) unambiguous clips (matching & non-matching) with the perceive labels, and (3) matching subset only.*

Label	Overall	Matching	Non-matching	Ambiguous
Intended	**56.28%**	59.63%	**48.25%**	**53.41%**
Perceived	41.91%	49.37%	35.65%	41.92%
Matching	50.66%	**65.73%**	39.06%	44.72%

5. Conclusion

In emotion datasets, intended emotion labels and perceived emotion labels both contain valuable information about how human express and perceive emotions, and there is a considerable mismatch between the two. In this paper, we introduced a novel ranking framework that simultaneously consider both labels while addressing the mismatch between them. Based on analyzing the agreement between the intended and perceived emotions, as well as the consistence among all perceptual ratings, we developed three different ranking rules to generate relevance scores and use them to train preference learning algorithms, creating three sets of emotional rankers for six basic emotions. The experimental evaluation demonstrated that the precision rates in retrieving target categorical emotions are higher than the ones achieved with an alternative method only based on the consistent level of the perceived labels. We also combined the ranking scores of individual emotions for multi-class classification. We showed that emotion classification systems with ranking information significantly outperform the conventional SVM classifiers. As future work, we will further explore the proposed ranking framework on real spontaneous emotional speech, where the mismatch between the intended and perceived labels is expected to be worse.

6. Acknowledgement

This work is supported by NYIT's Institutional Support for Research and Creativity (ISRC) Grants.

7. References

[1] M. Grimm, K. Kroschel, and S. Narayanan, "Support vector regression for automatic recognition of spontaneous emotions in speech," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2007, Honolulu, Hawaii, USA, April 15-20, 2007*, 2007, pp. 1085–1088.

[2] S. Steidl, "Automatic classification of emotion related user states in spontaneous children's speech," Ph.D. dissertation, University of Erlangen-Nuremberg, 2009.

[3] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, 2009.

[4] M. M. H. E. Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[5] B. Vlasenko, B. W. Schuller, A. Wendemuth, and G. Rigoll, "Frame vs. turn-level: Emotion recognition from speech considering static and dynamic processing," in *Affective Computing and Intelligent Interaction, Second International Conference, ACII 2007, Lisbon, Portugal, September 12-14, 2007, Proceedings*, 2007, pp. 139–147.

[6] H. Meng, J. Pittermann, A. Pittermann, and W. Minker, "Combined speech-emotion recognition for spoken human-computer interfaces," in *2007 IEEE International Conference on Signal Processing and Communications*. IEEE, 2007, pp. 1179–1182.

[7] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[8] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 223–227.

[9] B. Pang, L. Lee *et al.*, "Opinion mining and sentiment analysis," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[10] J. J. M. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo, ICME 2009, June 28 - July 2, 2009, New York City, NY, USA*, 2009, pp. 1436–1439.

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

[12] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis," in *Proceedings of the 2nd ACM Workshop on Multimedia Semantics, MS 2008, Vancouver, British Columbia, Canada, October 31, 2008*, 2008, pp. 32–39.

[13] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Front. ICT*, vol. 2015, 2015.

[14] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016, Shanghai, China, March 20-25, 2016*, 2016, pp. 5205–5209.

[15] ——, "Retrieving categorical emotions using a probabilistic framework to define preference learning samples," in *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 2016, pp. 490–494.

[16] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 252–256.

[17] H. Cao, R. Verma, and A. Nenkova, "Combining ranking and classification to improve emotion recognition in spontaneous speech," in *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, 2012, pp. 358–361.

[18] ——, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, 2015.

[19] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[20] B. W. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 312–315.

[21] F. Eyben, F. Weninger, F. Groß, and B. W. Schuller, "Recent developments in opensmile, the munich open-source multimedia feature extractor," in *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, 2013, pp. 835–838.

[22] T. Joachims, "Training linear svms in linear time," in *Proceedings of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, August 20-23, 2006*, 2006, pp. 217–226.

[23] ——, "Optimizing search engines using clickthrough data," in *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, 2002, pp. 133–142.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, 2011.