Interspeech 2018
2-6 September 2018, Hyderabad

# Preference-Learning with Qualitative Agreement for Sentence Level Emotional Annotations

*Srinivas Parthasarathy and Carlos Busso*

Multimodal Signal Processing(MSP) lab, Department of Electrical and Computer Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
sxp120931@utdallas.edu, busso@utdallas.edu

## Abstract

The perceptual evaluation of emotional attributes is noisy due to inconsistencies between annotators. The low inter-evaluator agreement arises due to the complex nature of emotions. Conventional approaches average scores provided by multiple annotators. While this approach reduces the influence of dissident annotations, previous studies have showed the value of considering individual evaluations to better capture the underlying ground-truth. One of these approaches is the *qualitative agreement* (QA) method, which provides an alternative framework that captures the inherent trends amongst the annotators. While previous studies have focused on using the QA method for time-continuous annotations from a fixed number of annotators, most emotional databases are annotated with attributes at the sentence-level (e.g., one global score per sentence). This study proposes a novel formulation based on the QA framework to estimate reliable sentence-level annotations for preference-learning. The proposed relative labels between pairs of sentences capture consistent trends across evaluators. The experimental evaluation shows that preference-learning methods to rank-order emotional attributes trained with the proposed QA-based labels achieve significantly better performance than the same algorithms trained with relative scores obtained by averaging absolute scores across annotators. These results show the benefits of QA-based labels for preference-learning using sentence-level annotations.

**Index Terms**: speech emotion recognition, preference-learning

## 1. Introduction

Speech is the primary form of communication, conveying not only lexical content, but also our intentions, desires and emotions. Advanced human-computer interfaces that can respond to the users needs should be able to sense paralinguistic information conveyed in speech, including emotion. A vital component of an emotion recognition system is the labels used to train the system. Generally, emotional labels are either nominal - discrete categories such as happiness, sadness, and anger [1–3], or interval - attributes such as arousal (calm versus excited), valence (negative versus positive), and dominance (weak versus strong) [4–6]. Alternatively, there are ordinal labels that describe the preference between sentences/videos with respect to a given emotional descriptor (e.g., sentence one is more aroused than sentence two) [7]. Conventionally, ground-truth labels for emotional descriptors are collected through perceptual evaluations from multiple annotators listening to the stimulus [8, 9]. The emotional perception process is inherently difficult [10]. There often are inconsistencies between annotators due to both the complexity of the underlying emotions and the reliability of the annotators [11]. This inconsistency leads to poor inter-annotator agreement [11–13]. Therefore, it is important to explore frameworks that capture consistent information provided by the annotators, while filtering the noise. This step is important, since the performance of a recognition system depends on the type and quality of the labels used for training [14].

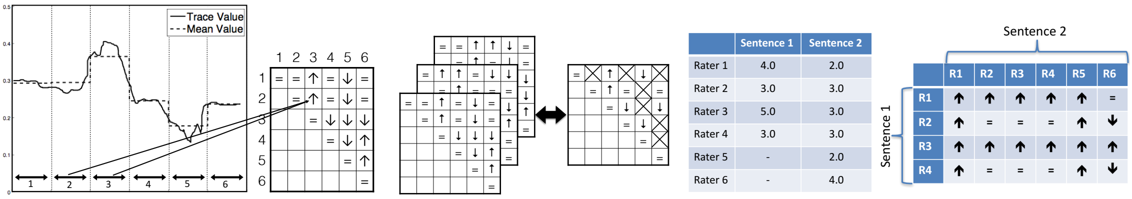
This paper aims to derive reliable ordinal labels from existing interval annotations provided at the sentence level to train preference-learning methods. The formulation for emotion recognition for attribute descriptors is traditionally either a regression (predicting the emotional attribute value) [15, 16], or a classification (recognizing dichotomized classes such as low versus high values for a given attribute) [17, 18] task. An alternative formulation is preference-learning where the task is to rank the test data according to a given criterion (in our case, an ordered emotional attribute). Conventionally, an absolute ground-truth is constructed by averaging annotations from multiple annotators [19]. Previous studies have argued that annotators are more reliable at judging local comparison rather than assigning global scores [7]. McKeown and Cowie introduced the *qualitative agreement* (QA) framework [20] that captures the trends (changes in emotion) on which most annotators agree. This framework is powerful as it detects relative trends from existing time-continuous interval evaluations (e.g., emotional traces). We believe that this framework can be used to provide more robust ground-truth labels compared to absolute labels, as it effectively extracts reliable information from noisy labels. While our previous studies have focused on time-continuous emotional traces [21, 22], most of the current databases contain sentence-level annotations [14, 19, 23], where a global value is assigned after listening to a sentence. This study proposes a QA-based method for sentence-level annotations of emotional attributes, showing its benefits over using absolute scores for preference-learning problems.

The proposed approach compares trends between individual annotations provided by different annotators to two different sentences. The pairwise comparisons create relative labels used for preference-learning to rank-order emotional attribute values. Using the proposed QA-based labels, we systematically compare the performance of two preference-learning methods: RankNet [24] and RankMargin [25]. The experimental evaluation shows that the frameworks trained with the proposed QA-based labels perform significantly better than systems trained with the average scores across evaluators. Furthermore, we show that preference-learning methods consistently perform better than ranking attribute values using the predictions provided by a baseline regression model. The superiority of preference-learning methods further suggests the ordinal nature of emotions [7].

The main contributions of this study are (a) defining a novel QA-based relative label for sentence-level emotional annotations, and (b) training preference-learning with RankNet and RankMargin losses, leveraging the proposed QA-based relative labels. The proposed approach achieves state-of-the-art performance on the MSP-Podcast corpus.

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

10.21437/Interspeech.2018-2478



|  | Sentence 1 | Sentence 2 |
|---|---|---|
| Rater 1 | 4.0 | 2.0 |
| Rater 2 | 3.0 | 3.0 |
| Rater 3 | 5.0 | 3.0 |
| Rater 4 | 3.0 | 3.0 |
| Rater 5 | - | 2.0 |
| Rater 6 | - | 4.0 |

|  | R1 | R2 | R3 | R4 | R5 | R6 |
|---|---|---|---|---|---|---|
| R1 | ↑ | ↑ | ↑ | ↑ | ↑ | = |
| R2 | ↑ | = | = | = | ↑ | ↓ |
| R3 | ↑ | ↑ | ↑ | ↑ | ↑ | ↑ |
| R4 | ↑ | = | = | = | ↑ | ↓ |

(a) Individal matrix for an emtional trace

(b) Concensus matrix

(c) QA-based labels for sentence-level annotations

Figure 1: *Use of QA method to create relative labels. (a) Individual matrix obtained from an emotional trace provided by one rater, (b) consensus matrix obtained by aggregating several individual matrices, (c) proposed approach to construct QA-based labels for sentence-level annotations. In the example, there are 15 preferences for sentence one, 2 preferences for sentence two and 7 draws.*

# 2. Related Work

## 2.1. Preference-Learning for Relative Emotional Labels

Preference-learning is an appealing framework for affective computing. Preference-learning algorithms learn to rank samples according to a given order for an emotional attribute. Few studies have used preference-learning to exploit the relative nature of emotions.

Studies have considered preference-learning for categorical emotions (e.g., happy ranker). Cao et al. [26] trained rankers by establishing preferences from the consensus labels. For a happy ranker, a sentence labeled as happy was always preferred over another sample labeled with a different label. Lotfian and Busso [27] proposed a probabilistic formulation to map individual evaluations of emotional categories into a numerical scale. Relative labels were derived from this metric to train preference-learning algorithms.

Most studies on preference-learning using acoustic features rely on emotional attributes. Lotfian and Busso [28] analyzed practical considerations to train preference-learning on emotional attributes. They evaluated the size of the training set, and the margin needed to consider that one sample is preferred over another. Parthasarathy et al. [29] further extended this framework with a *deep neural network* (DNN) architecture using a RankNet loss function. Martinez et al. [30] further showed the benefits of preference-learning over classification of attribute scores [30]. Our study presents novel training labels for preference-learning using emotional attributes. The key advantage of the proposed approach is that it relies on existing sentence-level annotations of emotional attributes, where many emotional corpora can be used.

## 2.2. Relative Labels with Qualitative Agreement

Studies have showed that annotators are more consistent on local comparisons than giving an absolute score representing a given attribute [7, 31]. Therefore constructing ground-truth labels that capture the relative relationship between samples is more meaningful than conventional labels based on absolute scores for both interval and nominal descriptors. We have shown that the QA method offers an appealing framework to derive ordinal labels from existing individual evaluations [21, 22].

McKeown and Cowie [20] proposed the QA method to obtain reliable trends from existing time-continuous annotations (e.g., emotional traces). The proposed QA method worked on annotations from a fixed set of annotators. Figures 1(a) and 1(b) illustrates the QA framework. First, a trace is discretized into bins, where the mean of the trace, denoted by $b_i$, is computed for bin $i$. The bins are pairwise compared using Equation 1 - 3. A threshold $t_{threshold}$ is used as the margin to establish whether the mean value of one bin is greater than, equal to or less than the mean value of another bin. This process gives an *individual matrix* (IM) for each annotator (Fig.1(a)). The IMs are then combined by considering the trends that X% (e.g., 66%) of the annotators agree on, providing trends across evaluators. Bins without agreement are marked with 'X' (Fig. 1(b)).

$$b_i - b_j \quad \geq t_{threshold} \tag{1}$$

$$b_j - b_i \quad \geq t_{threshold} \tag{2}$$

$$|b_i - b_j| \quad < t_{threshold} \tag{3}$$

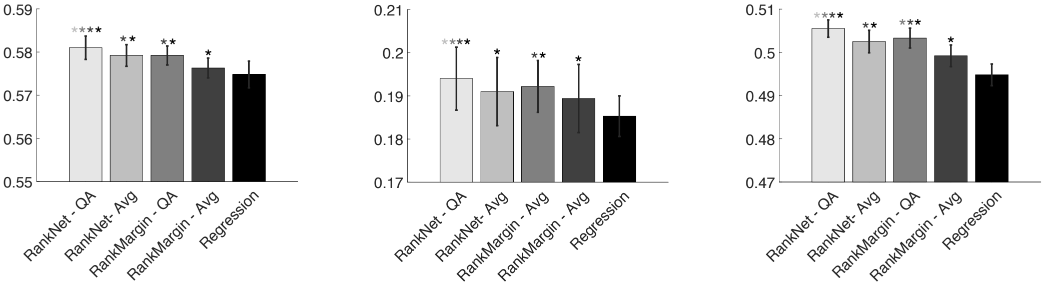
Parthasarathy et al. [21] used rank-based classifiers to show the benefits of the QA framework, using the identified trends as relative labels for time-continuous traces. They studied the effect of the parameter $t_{threshold}$ and the consensus percentage ($X\%$) in the definition of relative labels. Parthasarathy and Busso [22] proposed to use the QA framework to identify emotionally salient regions in time-continuous traces. The approach relied on comparing the bins with the median value of the trace. They showed through perceptual evaluations that hotspot regions identified by the QA method were more appropriated than regions identified by averaging the traces. All of these methods were implemented with time-continuous traces for emotional attributes. This paper extends the framework to sentence-level annotations for attribute descriptors.

# 3. Methodology

## 3.1. QA for Sentence Level Annotations

There are some limitations when using the QA approach with emotional traces. Time-continuous annotation of emotional attributes is a challenging task. Annotators have to continuously judge and annotate the changes in emotion, which requires a high cognitive effort. Therefore, few databases contain time-continuous annotations, which are annotated by few trained annotators. In contrast, many databases are annotated by multiple annotators with global values assigned at the sentence-level. With crowdsourcing, sentences can be annotated by a larger number of annotators (e.g., five annotators per sentence). Since the annotators are not necessarily the same for all the sentences, the QA method has to be adapted for sentence-level annotations.

This paper leverages the concepts behind the QA framework to define ordinal labels from existing sentence-level annotations for emotional attributes. The key concept behind the proposed formulation is to compare the trends rather than the absolute scores. Figure 1(c) describes the approach, which starts with existing sentence-level evaluations of emotional attributes. We assume that evaluators used a Likert scale to score each sentence. For example, for valence the scale may be from 1 (very negative) to 7 (very positive). We do not assume that the sentences are annotated by the same number of evaluators. If $N_1$ and $N_2$ are the numbers of independent annotators for sentence



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

Figure 2: *Kendall's Tau for global ranking of emotional attributes. Bars show the mean value and standard deviations over 10 random initializations. An asterisk over a bar indicates significant differences over the method identified by the asterisk's color.*

one and sentence two, we create a $N_1 \times N_2$ matrix with comparisons between all pairs of annotators (Fig. 1(c)). In each comparison, we evaluate whether the score of one rater for sentence one is greater than, less than or equal to the score provided by another rater for sentence two. The evaluation considers relationships similar to Equations 1-3, where the threshold $t_{threshold}$ is set to one (attribute scores are usually integer numbers).

The matrix records the trends between two sentences. After evaluating all the comparisons, we summarize the trends. In the example in Figure 1(c), we have 15 preferences for sentence one (62.5%), two preferences for sentence two (8.3%) and seven draws (29.2%). Similar to the QA approach, we define a consensus when the preference for one sentence is above $X\%$. If this level of consensus is achieved, the pair of sentences and their preference is stored. Otherwise, we discard this particular pair since the sentences do not reach consensus. In the example in Figure 1(c), we would only consider that sentence one is preferred over sentence two if $X \leq 62.5\%$.

The proposed approach addresses some of the key differences in dealing with sentence-level annotations: (1) we do not need the same annotators to evaluate all the sentences, (2) we do not need the same number of evaluators per sentence, and (3) we do not have to split the sentences into bins of similar duration, creating pairwise comparisons even for sentences of different lengths. The speaking turns are our bins.

### 3.2. Preference-Learning Methods

Preference-learning is a popular framework used for retrieval tasks. The formulation aims to learn the order of the samples $y_1 > y_2 \ldots > y_n$, using pairwise comparisons. We use preference-learning to rank-order sentences according to the emotional attributes arousal, valence and dominance. Given a pair of samples with feature vectors $\mathbf{x_i}$, and $\mathbf{x_j}$ and corresponding labels $y_i$, and $y_j$, the preference-learning framework learns a function $f$ that maps the feature $\mathbf{x}$ into a score $S$. The function $f$ should reflect the preference in the labels (e.g., if $y_i > y_j$, then $S_i > S_j$)). Using deep learning, the function $f$ can be learned using different losses. This study uses two particular losses: the RankMargin [25] and the RankNet [24] losses. The RankMargin loss is a hinge loss on the attribute scores. Given the preference ($y_i > y_j$), the RankMargin loss is given by $L_{RankMargin} = \max(0, \alpha + S_j - S_i)$, where $\alpha = 1$, is a margin to separate the samples. The RankNet loss uses a logistic loss on the attribute scores. Under the previous assumptions, the RankNet loss is given by $L_{RankNet} = \log(1 + \exp^{-(S_i - S_j)})$. The network's parameters are trained to minimize the corresponding losses. This study compares these two losses for rank-ordering emotional attributes.

## 4. Experimental Evaluation

### 4.1. MSP-Podcast Corpus

The experimental evaluation is conducted on the version 1.1 of the MSP-Podcast corpus [23]. The collection of the corpus is an ongoing effort, where this version contains 22,630 sentences collected from audio-sharing websites. The sentences are collected from naturalistic conversations, which are segmented into speaking turns with duration between 2.7s and 11s. The segments from the podcasts are annotated with emotional labels using a crowdsourcing framework [32]. The sentences are annotated by at least five annotators for the emotional attributes: arousal (1 - very calm, 7 - very excited), valence (1 - very negative, 7 - very positive), and dominance (1 - very weak, 7 - very strong). This study uses the individual annotations as well their average values across sentences (consensus labels). We have manually identified speaker information for 18,991 segments. We split the database into train, development and test sets, making our best effort to keep the partitions speaker independent. The test set contains 7,181 sentences from 50 speakers, the development set contains 2,614 sentences from 20 speakers, and the train set contains the rest of the data (12,835 sentences).

### 4.2. Acoustic Features

We use the Interspeech 2013 acoustic feature set used for the *computational paralinguistic challenge* (ComParE) [33]. The feature set contains sentence-level statistics, referred to as *high level functionals* (HLFs), calculated on frame-level features (e.g., mean of the fundamental frequency). Overall, the feature set contains 6,373 acoustic features describing different acoustic properties affected by the externalization of emotion. Further details on the feature set are presented by Schuller et al. [33].

### 4.3. Implementation of Preference-Learning Frameworks

The preference-learning framework employs a DNN to learn the function $f$. To train the network, we define relative labels based on the QA-based approach using $X = 60\%$. As a baseline, we consider relative labels for two sentences by comparing their average scores. We use $t_{threshold}$ = 0 for the baseline. While increasing the threshold increases the reliability on the labels, our preliminary tests showed that varying the values for $t_{threshold}$ produced similar results.

The DNN uses feed-forward connections, with two hidden layers and 256 neurons in each layer. The *rectified linear unit* (ReLU) is used as the activation function of the neurons in the hidden layers. We use a dropout with probability of $p = 0.5$ between hidden layers to regularize the network. We train all our models for 100 epochs with ADAM optimization, using the default initialization and a learning rate of $1e^{-4}$. There are mil-

Table 1: *Kendall's Tau for local ranking of emotional attributes. Columns show ranking results for the top-K%* (Hi*) and bottom-K%* (Lo*) of the retrieved sentences. The rows show preference-learning models (*RN*:RankNet,* RM*: RankMargin), and regression model (*Reg*: Regression). The symbol* $^*$ *indicates that QA-based labels provide significant better* $\tau$ *than the corresponding method trained with average-based (*Avg*) labels. The symbol* $^\dagger$ *indicates that preference-learning method provides significant better* $\tau$ *than the regression model.*

| Emo | Model | Hi-10 | Hi-20 | Lo-10 | Lo-20 |
|---|---|---|---|---|---|
| Arousal | RN-QA | $0.273^\dagger$ | $0.300^{*\dagger}$ | $0.297^*$ | $0.332^{*\dagger}$ |
| | RN-Avg | $0.274^\dagger$ | $0.298^\dagger$ | 0.286 | 0.325 |
| | RMa-QA | $0.272^\dagger$ | $0.298^{*\dagger}$ | $0.295^*$ | $0.329^*$ |
| | RM-Avg | $0.272^\dagger$ | $0.296^\dagger$ | 0.288 | 0.326 |
| | Reg | 0.262 | 0.287 | 0.293 | 0.330 |
| Valence | RN-QA | 0.111 | $0.162^*$ | $0.061^\dagger$ | 0.044 |
| | RN-Avg | 0.109 | 0.158 | 0.058 | $0.051^\dagger$ |
| | RM-QA | $0.117^*$ | 0.159 | 0.060 | $0.045^\dagger$ |
| | RM-Avg | 0.104 | 0.157 | 0.058 | $0.050^\dagger$ |
| | Reg | 0.122 | 0.166 | 0.056 | 0.043 |
| Dominance | RN-QA | $0.159^*$ | $0.200^{*\dagger}$ | 0.241 | 0.287 |
| | RN-Avg | 0.149 | 0.190 | 0.238 | 0.285 |
| | RM-QA | $0.160^*$ | $0.198^{*\dagger}$ | $0.244^*$ | $0.287^*$ |
| | RM-Avg | 0.148 | 0.190 | 0.238 | 0.284 |
| | Reg | 0.156 | 0.194 | 0.249 | 0.291 |

lions of relative labels in the training set. Since training with all preference pairs is expensive, we use 200k random preference pairs per epoch, which gives good results on the validation set.

# 5. Results

We use the Kendall's Tau $\tau$ correlation coefficient to evaluate the performance of the predicted rank-orders. $\tau$ measures the correlation between two lists by considering the number of concordant and discordant pairs in lists. $\tau$ varies between [-1,1], where $\tau = -1$ corresponds to complete disagreement, and $\tau = 1$ corresponds to complete agreement. The ground-truth order on the test set is estimated by sorting the average of the scores (e.g., using the average-based labels).

While this study compares preference-learning methods trained with the QA-based and average-based relative labels, we also use a regression model as a baseline. The regression model is trained to predict the absolute value of the emotional attribute. We use identical architectures as the preference-learning models, trained with the *concordance correlation coefficient* (CCC) as the objective function. For inference, the neural network is used to rank-order the emotional attributes.

## 5.1. Global Ordering of Emotional Attributes

The first evaluation compares the global order obtained by the alternative frameworks for each emotional attribute. We compare five different models. Two models use the RankNet loss with either the QA-based labels (*RankNet-QA*) or the average-based labels (*RankNet-Avg*). Two models use the RankMargin loss with either the QA-based labels (*RankMargin-QA*) or the average-based labels (*RankMargin-Avg*). The fifth model is the baseline using the regression model (*Reg*). We train all the models 10 times with different random initialization, reporting the mean and standard deviation of $\tau$ across trials. All the models are trained on the train set, maximizing performance on the development set. We perform early stopping based on the performance on the development set, evaluating the models on the test set. We test significance using the one-tailed t-test over the 10 trials, testing significance if $p \leq 0.01$.

Figure 2 reports the mean and standard deviation for the different models. The results of the statistical tests are denoted with color-coded asterisk on top of the bars. We observe that models trained with the QA-based labels (i.e., *RankNet-QA* and *RankMargin-QA*) perform significantly better than their corresponding models trained with the average-based labels (i.e., *RankNet-Avg* and *RankMargin-Avg*), suggesting the importance of using relative trends over average values. Amongst the preference-learning methods, the *RankNet-QA* model performs significantly better than all other frameworks. Even though the ground-truth for the rank-order on the test set is determined by the average values, the QA-based methods perform better than the average-based methods. The figure also shows that preference-learning models perform better than the regression model, providing another promising evidence on the use of preference-learning in affective computing.

## 5.2. Local Ordering of Emotional Attributes

In many emotion retrieval scenarios, we are interested in retrieving the most emotional utterances (e.g., the most positive/negative utterances in a set). This task requires preference-learning models to have better precision rate on the extremes of the list. Our second evaluation measures the local ordering of emotional attributes at the top and bottom of the list. We identify the top-K% (High) and bottom-K% (Low) of the utterances in the test set. Then, we evaluate the predicted orders of these utterances using $\tau$. Table 1 lists the mean $\tau$ for different values of $K$. The results correspond to the average values over the 10 trials. We report statistically significant differences with the symbols † and $*$ (one-tailed t-test over 10 trials, asserting significance if $p \leq 0.01$).

In most cases, *RankNet-QA* and *RankMargin-QA* perform better than *RankNet-Avg* and *RankMargin-Avg*, respectively. The differences in many cases are statistically significant. The preference-learning methods almost always perform better than the regression model. While the regression models performs well for some cases, their results are not consistent for ordering sentences in both the top and bottom lists. This inconsistency leads to poorer global performance. With the increase in $K$, we include samples that are separated by a greater margin in the ground-truth rankings, leading to better performance in the ranking. The local order for the top and bottom part of the lists further demonstrates the superiority of the preference-learning frameworks trained with QA-based labels.

# 6. Conclusions

This study proposed a novel framework based on the QA method for constructing relative labels from sentence-level annotations of emotional attributes. The labels capture the relative trends found on individual annotations provided by different raters. We evaluated the proposed QA-based labels using deep learning methods for preference-learning implemented with RankNet and RankMargin losses. Our results show that frameworks trained with the QA-based labels produce significantly better global and local rankings compared to methods trained with average-based labels. Furthermore, the preference-learning methods perform better than a regression baseline, indicating the ordinal nature of emotions. Our future work includes extending the preference-learning framework to identify emotionally salient sentences (hotspots) during long conversations. Furthermore, we would like to extend the framework by (1) considering categorical emotional descriptors, and (2) relying on other modalities such as facial expressions or physiological signals.

# 7. References

[1] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. 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: ACM Press, October 2004, pp. 205–211.

[2] C. Lee, S. Yildirim, M. Bulut, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and 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)*, vol. 4, Honolulu, HI, USA, April 2007, pp. 941–944.

[4] J. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.

[5] ——, “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] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, “The world of emotions is not two-dimensional,” *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, December 2007.

[7] G. Yannakakis, R. Cowie, and C. Busso, “The ordinal nature of emotions,” in *International Conference on Affective Computing and Intelligent Interaction (ACII 2017)*, San Antonio, TX, USA, October 2017, pp. 248–255.

[8] R. Cowie and R. Cornelius, “Describing the emotional states that are expressed in speech,” *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, April 2003.

[9] C. Busso, M. Bulut, and 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. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[10] R. Cowie, “Perceiving emotion: towards a realistic understanding of the task,” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3515–3525, December 2009.

[11] L. Devillers, L. Vidrascu, and L. Lamel, “Challenges in real-life emotion annotation and machine learning based detection,” *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[12] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, “Primitives-based evaluation and estimation of emotions in speech,” *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

[13] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ““Of all things the measure is man” automatic classification of emotions and inter-labeler consistency,” in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 317–320.

[14] 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.

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

[16] S. Parthasarathy and C. Busso, “Jointly predicting arousal, valence and dominance with multi-task learning,” in *Interspeech 2017*, Stockholm, Sweden, August 2017, pp. 1103–1107.

[17] S. Mariooryad and C. Busso, “Correcting time-continuous emotional labels by modeling the reaction lag of evaluators,” *IEEE Transactions on Affective Computing*, vol. 6, no. 2, pp. 97–108, April-June 2015, special Issue Best of ACII.

[18] R. Xia and Y. Liu, “A multi-task learning framework for emotion recognition using 2D continuous space,” *IEEE Transactions on Affective Computing*, vol. 8, no. 1, pp. 3–14, January-March 2017.

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

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

[21] 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.

[22] 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.

[23] 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. To appear, 2018.

[24] L. to rank with nonsmooth cost functions, “C.j. burges and r. ragno and q.v. le,” in *Advances in neural information processing systems (NIPS 2007)*, Vancouver, B.C., Canada, December 2007, pp. 193–200.

[25] H. Martinez, Y. Bengio, and G. Yannakakis, “Learning deep physiological models of affect,” *IEEE Computational Intelligence Magazine*, vol. 8, no. 2, pp. 20–33, May 2013.

[26] 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.

[27] 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.

[28] ——, “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.

[29] 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.

[30] H. Martinez, G. Yannakakis, and J. Hallam, “Don’t classify ratings of affect; rank them!” *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 314–326, July-September 2014.

[31] A. Metallinou and S. Narayanan, “Annotation and processing of continuous emotional attributes: Challenges and opportunities,” in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

[32] 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.

[33] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, “The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism,” in *Interspeech 2013*, Lyon, France, August 2013, pp. 148–152.