Using Agreement on Direction of Change to Build Rank-Based Emotion Classifiers

Srinivas Parthasarathy, *Student Member, IEEE,* Roddy Cowie, *Member,IEEE,* Carlos Busso, *Senior Member, IEEE,*

***Abstract*—Automatic emotion recognition in realistic domains is a challenging task given the subtle expressive behaviors that occur during human interactions. The challenges start with noisy emotional descriptors provided by multiple evaluators, which are characterized by low inter-evaluator agreement. Studies have suggested that evaluators are more consistent in detecting qualitative relations between episodes (i.e., emotional contrasts), rather than absolutes scores (i.e, the actual emotion). Based on these observations, this study explores the use of relative labels to train machine learning algorithms that can rank expressive behaviors. Instead of deriving relative labels from expensive and time consuming subjective evaluations, the labels are extracted from existing time-continuous evaluations over expressive attributes annotated with FEELTRACE. We rely on the *qualitative agreement* (QA) analysis to estimate relative labels which are used to train rank-based classifiers (rankers). The experimental evaluation on the SEMAINE database demonstrates the benefits of the proposed approach. The ranking performance using the QA-based labels compare favorably against preference learning rankers trained with relative labels obtained by simply aggregating the absolute values of the emotional traces across evaluators, which is the common approach used by other studies.**

***Index Terms*—Rank-based emotion recognition, time-continuous emotional descriptors, relative emotional labels, emotion recognition**

I. Introduction

BUILDING robust emotion recognition systems depends on datasets where there is a reliable relationship between the descriptors that they provide and the emotional significance of the behaviors that they show [1], [2]. Finding appropriate descriptors is a hard problem. The terms may refer to discrete emotional categories such as happiness, sadness, and anger [3]–[5], or attributes such as arousal, valence and dominance [6], [7]. In both cases, though, labels are assigned through perceptual evaluations, where external observers use the chosen terms to characterize the emotional content. Inferring emotion from the ambiguous expressive behaviors observed in daily human interaction is a complex process [8], and it is not realistic to expect perfect agreement across evaluators [9]. In fact, the inter-evaluator agreement for perceptual evaluations is often very low: exact values depend on the number of classes and underlying expressive behaviors [9]–[12]. Metallinou et al. [13] showed low agreement even for evaluations completed by the same rater multiple times. That situation has serious implications for the accuracy of emotional classifiers that are trained with these labels [2]. We should not rely on classifiers trained on unreliable data.

One of the promising leads is that annotators tend to agree more on relative trends (e.g., is the first video more positive than the second video?) than on absolute ratings (e.g., what is the valence of the video?) [13]. There are similar findings for music emotion recognition, where 'ground truth' derived from absolute values was found to be less reliable than relative labels [14]. Soleymani et al. [15] suggested that conducting quantitative evaluations of affective states is a harder perception problem than conducting qualitative comparisons. Those results suggest that it is a priority to explore (1) ways of obtaining reliable relative rankings, and (2) machine learning algorithms that can effectively exploit that kind of data.

This study builds on a method proposed by Cowie and McKeown [16] which is known as *Qualitative Agreement* (QA). The central idea is to define labels based on relative assessments derived from existing time-continuous labels annotated with FEELTRACE [17]. The approach operates by identifying segments in which the evaluators reach a predefined level of agreement on relative levels, regardless of the absolute values of their annotations. Their data indicate that it has statistical advantages over approaches that use absolute values. We show that it can be used to underpin machine learning, and the resulting techniques have advantages over techniques based on absolute values.

We clarify issues in the definition of relative emotional labels, and identify a rank based classifier (ranker) that can learn the relevant pairwise evaluations. We evaluate the approach with the SEMAINE database [18] over arousal, valence, power and expectation. In the light of the clarification, we study how achieved accuracy depends on the thresholds used to define inter-evaluator agreement. We then compare the performance of our technique with an approach used in previous preference learning studies on ranking emotional behaviors [19], which serves as a baseline. It uses relative labels obtained after aggregating the absolute values of the traces across evaluators. The results show that using QA-based labels has numerical as well as conceptual advantages. We obtain absolute improvements up to 5% (8% relative) compared with the performance of rankers trained with the baseline approach.

The rest of the paper is organized as follows. Section II describes the theoretical background, existing methods for describing emotions, and the underlying challenges involved in deriving reliable labels for the purpose of training emotion recognition systems. The section also describes the database used for this study. Section III describes the proposed ap-

S. Parthasarathy, and C. Busso are with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, TX 75080 (e-mail:sxp120931@utdallas.edu, busso@utdallas.edu).

R. Cowie is with the School of Psychology at Queens University, Belfast, UK

Manuscript received November xx, xxxx; revised xxxx xx, xxxx.

proach to extracting consistent relative rankings from multiple continuous emotional annotations. The section also presents the preference learning algorithm used to exploit the relative ranking scores. Section IV presents the implementation of the approach discussing the parameters and acoustic features. Section V describes emotional ranking evaluations, in which we compare the proposed QA-based labels with the labels derived after aggregating the absolute scores of the emotional traces (i.e., baseline labels). Section VI concludes the paper with a summary of the contributions of this study, future directions and final remarks.

II. Background

A. Underlying Psychological Issues

The QA approach was prompted by psychological questions about the task of generating a number that describes an impression of emotion. Two distinct lines of argument indicate that the task is very far from straightforward. The first line of argument goes back to the beginnings of psychology. The question is whether experiences can be mapped onto a numerical scale. Most early psychologists accepted that experiences could be ranked, but thought that they defied truly numerical description. Tichener argued that looking at two rooms, "We can say, by eye, that illumination of the first room is greater than that of the second. How much greater, we cannot possibly say." [20, p.86]: James, Stumpf, Mach, and Wundt agreed [21, p.59]. Fechner [22] believed that numerical descriptions could be derived, but only via complex (and debatable) inferences from pairwise comparisons. His most famous critic, Stevens, argued for approaches where numbers were assigned directly [23]; but later work criticized his approach in turn, and argued for yet another set of methods based on inference from comparisons [24].

Beyond those unresolved conceptual debates are undisputed practical issues. Numerical estimates are extraordinarily sensitive to range, anchor points, and sequential effects [21, p.268-296]. For example, Stevens advocates direct estimation, but reports huge context effects on judgments of relative brightness of the same two stimuli. In one, the ratio of brightest to darkest was judged to be 100 to 1. In the other, it was judged to be about 3 to 1 [21, p.59]. If there is a numerical scale associated with experience, it is extraordinarily elastic. The implications for techniques where raters generate many data points in a session are obvious, and disturbing.

The second line of argument rests on evidence that ambiguity is a feature of everyday emotional displays. It is rare to find expressions which show all the signs of an archetypal emotion [25], and commonplace for signs in different modalities to invite different conclusions [26]. Beyond that, it is always open to question whether signs in any modality are direct reflections of emotional states; part of an honest attempt to portray chosen stances; or deceptive [27]. Hence, unsurprisingly, people use context to disambiguate evidence from face and voice [28]. If the displays are in fact ambiguous, the last thing we should ask for is ratings which give them a single agreed value. An interesting light on that point is that the claim made for the first continuous rating system was that it revealed differences in the way people perceived the same episode [29].

Given that background, it may not be wise to insist on asking how humans can be induced to deliver data that meets pre-specified ideals (such as reliable measures on a numerical scale). It seems sounder to ask how we can disentangle the kinds of information that they can deliver from kinds that they cannot. A proposal that is consistent with the psychology is that an extended display probably contains some extracts (not all) that people can rank reliably (without quantifying the difference). The QA approach followed directly from that reasoning.

B. Related Work on Emotion Recognition

The theoretical issues noted above are reflected in practical work on annotating expressive behaviors. Several researchers have highlighted the inconsistency of labels provided by different raters [30]–[33]. Depending on the task, the Kappa statistics for discrete categorical evaluation is usually lower than κ=0.5 [9]–[12], [34]. For continuous dimensional annotations, it is also a challenge to obtain high correlation between scores provided by different raters [35]. The low inter-evaluator agreement affects the performance of emotion recognition systems since these noisy labels are used to train the classifiers. Unsurprisingly, there is high correlation between emotion recognition performance and inter-evaluator agreement [34], [36].

Reviewing the evidence, Metallinou et al. [13] came to a conclusion similar to the nineteenth century psychophysicists': that evaluators are better at detecting relative emotional changes than absolute values. On one side were the differences when raters are asked to label the emotional content of the stimulus using labels like happiness, anger and sadness, or when they are asked to assign an absolute value for continuous dimensional annotations. On the other side, studies have shown that evaluators are consistent in detecting emotional trends [13]–[15]. After watching two videos, people can reliably determine which one is more positive or active.

The obvious inference is that labelling should be based on comparative judgments of that kind. However, that has a major drawback, which is that the number of evaluations becomes very large. If there are N samples to be annotated, the approach would require $N \times (N-1)/2$ pairwise comparisons. If we annotate multiple emotional attributes (i.e., arousal, valence, power), the cost and the time of conducting perceptual evaluations would be enormous. Also, combining material from separately labelled sources would always require a new labelling exercise. The QA approach proposed by Cowie and McKeown [16] offered a solution to that problem by using continuous emotional traces annotated with FEELTRACE (see Section III-A for the details).

The QA approach transforms raw traces into matrices that capture the meaningful ordinal information in them, and discard noise imposed by the attempt to map information that is functionally ordinal onto a ratio scale. Their form is described in Section III-A. The result of the transformation should be to reveal agreement that is obscured by noise in the raw traces. Cowie and McKeown confirmed that for the well-established global dimensions – intensity, valence, and arousal

– the transformed data were much more likely to show highly significant agreement (p <0.01) than raw traces were to pass a comparably stringent test of agreement (alpha>0.85). The opposite held for the more specific dimension of expectation, suggesting that the issue is, in line with the theory, bound to the particular kinds of judgment involved.

The initial analyses show that the approach is interesting, but they do not show that it can be used in a machine learning context. This paper takes up that challenge.

One of the key requirements is for machine learning algorithms that rank-order samples in terms of a given emotional attribute, rather than assigning a value or a class. Very few studies have addressed that issue in the context of emotion [37]. Suitable algorithms exist, though. They are popular in information retrieval problems [38], [39]. In the context of emotion, Yang et al. [14], [40] used rankers for music emotion recognition. Cao et al. [41], [42] created rankers as a mid-level representation to recognize categorical labels. They trained a ranker for each discrete emotional class (i.e., which label is more happy). The results of the rankers were combined to recognize the given emotion. Lotfian and Busso [19], [43] and Martinez et al. [37] discussed practical implementations for preference learning to rank emotional behaviors. Soleymani et al. [15] presented a method to rank affective behaviors on movies based on regression models. These preference learning algorithms present solutions to practical problems such as emotion retrieval, where the queries are defined by specifying target expressive content, surveillance and security applications, where the preference algorithm ranks large speech repository according to emotional content selecting threatening behaviors to be further analyzed by forensic experts, emotional hotspot detections, where the preference algorithm selects speech segments within a dialog with the strongest emotional content [44], and remote assistant technologies, where health care practitioner can review the most relevant events of patients with emotional disorder. Our study explores preference learning algorithms trained with relative labels that are derived from existing emotional traces.

C. Database

The study relies on the SEMAINE database [18] (SEMAINE stands for Sustained Emotionally colored MAchine-human Interaction using Nonverbal Expression). It is a large audiovisual database that consists of dyadic interactions between a 'user' (always a human) and an 'operator' (either a human or a virtual agent). The operator simulates the role of a *sensitive artificial listener* (SAL) agent and his/her role is to elicit emotional reactions from the user. The SAL agent takes four personalities: Spike, Poppy, Obadiah and Prudence: each aims to elicit a particular kind of emotional response from the user (angry, happy, gloomy and reasonable respectively). While there are different scenarios used to record the corpus, this study relies on the Solid-SAL portion, in which the operator is a human. We consider 40 sessions recorded from ten different subjects (IDs: 2, 3, 5, 7, 8, 10, 11, 13, 14, and 15).

The SEMAINE database has been emotionally annotated on multiple dimensions. This study uses arousal (calm versus active), valence (negative versus positive), power (weak versus strong), and expectation (predictable versus unexpected). The selection of these emotional dimensions was influenced by the study of Fontaine et al. [45] which indicated that these four dimensions are adequate to capture most of the differences between everyday emotion descriptors. While the recordings of the operators and the users were annotated, we only consider the turns from the users. The subjective evaluation was conducted with the FEELTRACE toolkit [17]. Instead of assigning a global score to a segment, FEELTRACE records frame-by-frame the location of the mouse's cursor over a *graphical user interface* (GUI), in which the axes represents the given dimension. As the evaluator watches the video, he/she moves the mouse's cursor to reflect his/her perception, providing time-continuous emotional traces. Each of the emotional attributes that we consider was separately traced by at least six evaluators. Figure 1 shows the emotional traces of six raters for arousal (session 53 in the database).

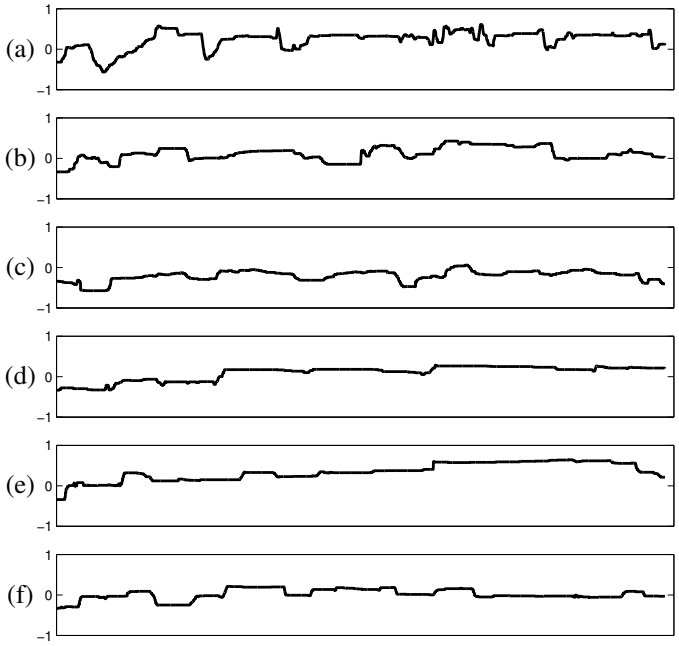

Fig. 1. Six emotional traces for arousal for session 53 in SEMAINE database.

The traces collected with toolkits such as FEELTRACE depend on processes that take time. The rater watches the stimuli, infers the emotional content, and reacts accordingly by moving the mouse's cursor. That implies an intrinsic delay between informative features in the recording and their effects in the traces. Our work [46], [47] and that of other researchers [48] have studied the evaluators' reaction lag in the traces. These studies demonstrated improved performance when the emotional traces are compensated for this reaction lag. This study uses the average delays across sessions observed in Mariooryad and Busso [47] for arousal and valence (this study did not consider power and expectation). For power and expectation, we use the average delays reported in Nicolle et al. [48]. Table I reports the delays used in this study.

TABLE I
AVERAGE REACTION LAG ACROSS ALL SESSIONS FOR DIFFERENT EMOTIONAL DIMENSIONS [47], [48].

	Arousal	Valence	Power	Expectation
Reaction Lag	5.44s	4.08s	5.00s	5.00s

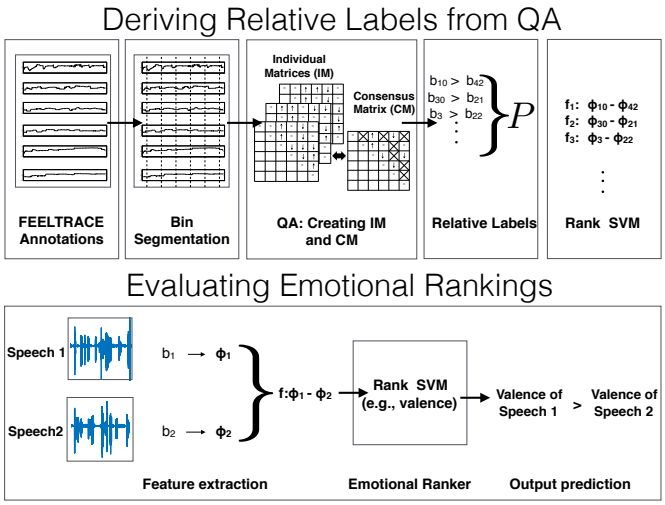

Fig. 2. Diagram illustrating the methodology to derive relative labels using the QA approach (top panel) used for preference learning (bottom panel)

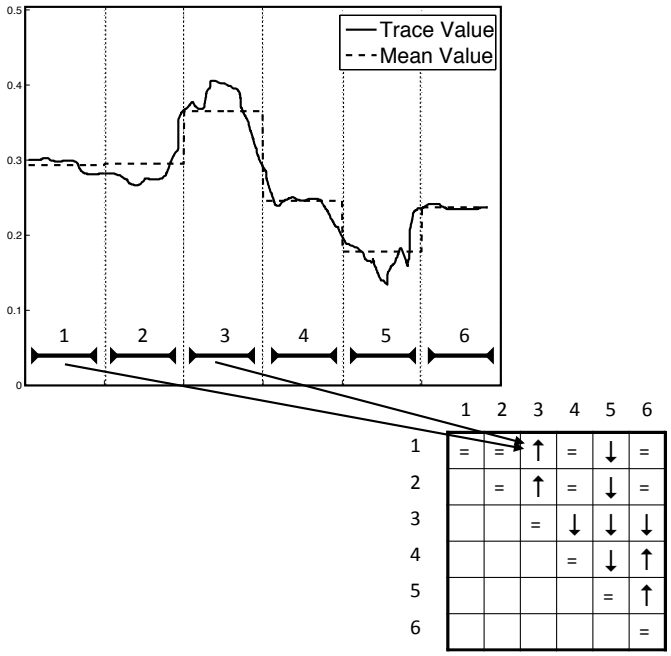

Fig. 3. Qualitative agreement analysis. The figure illustrates the process of creating individual matrices for a given trace.

III. METHODOLOGY

This section describes the methodology used to rank order emotional behaviors. Figure 2 shows a diagram of the proposed framework. First, we present the approach to deriving consistent relative rankings from existing absolute time-continuous emotional annotations (Sec. III-A). Then, we describe the proposed machine learning framework trained with the proposed relative annotations (Section III-B).

A. Relative Labels from Existing Annotations

As we discussed in Section II-B, evaluators tend to agree more on relative changes than on absolutes values. While the six traces in figure 1 present clear differences, some emotional changes are consistently evaluated. Global methods to estimate the reliability of labels, such as the Cronbach's alpha, provide low agreement when there are regions where raters make very different judgments, or where changes are in the same direction, but substantially different in magnitude. What the low agreement does not reflect is the fact that evaluators do agree on some relationships. The *qualitative agreement* (QA) method mentioned earlier was developed to address this drawback of existing methods.

The QA approach divides traces into segments and identifies pairs of segments for which the evaluators agree on the qualitative relationship. Each pair involves two regions in the trace such that a set proportion of the raters agree either which of them is higher; or that they are very close. Initially, the approach builds an *individual matrix* (IM) for each trace, and these are later combined into a *consensus matrix* (CM).

Consider a given emotional dimension (arousal, valence, power or expectation). The first step in the QA approach is to form the *individual matrix* (IM). Figure 3 illustrates the process, where the solid line represent the trace (absolute values) provided by one rater for the given emotional dimension. First, the trace is divided into bins of equal length (the illustration in Fig. 3 has six bins). The length of the bins, L, is a parameter of the approach, which is set to 3s in this study for most of the evaluation (see discussion in Sec. IV-A). Then, we estimate the mean value of the trace during each bin, denoted by b_i. The dotted line in figure 3 shows the mean values for these bins. Then, we compute simple pairwise comparisons between all the mean values of the bins. For this purpose, we define a threshold $t_{threshold}$, which Martinez et al. [37] denoted as *minimum distance*. This parameter defines the margin required to say that one score should be counted as greater than the other. For the bins i and j, where $i < j$, we perform the following comparisons:

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

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

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

With these comparisons, we identify the trends between the ith-bin and jth-bin: decreasing (Eq. 1), increasing (Eq. 2), or similar (Eq. 3). This information is entered into the (i, j) entry of IM, producing a skew-symmetric matrix with the diagonal element satisfying Eq. 3 (Fig. 3 only shows the upper triangular elements in IM). In the example on Figure 3, $b_3 > b_1$. Therefore, $\text{IM}_{(1,3)}$ is set as "increasing". Likewise, $b_5 < b_1$ so $\text{IM}_{(1,5)}$ is set as "decreasing". The rest of the entries are set using this approach.

Once the individual matrix is created for each annotator, we estimate the *consensus matrix* (CM) between raters. This

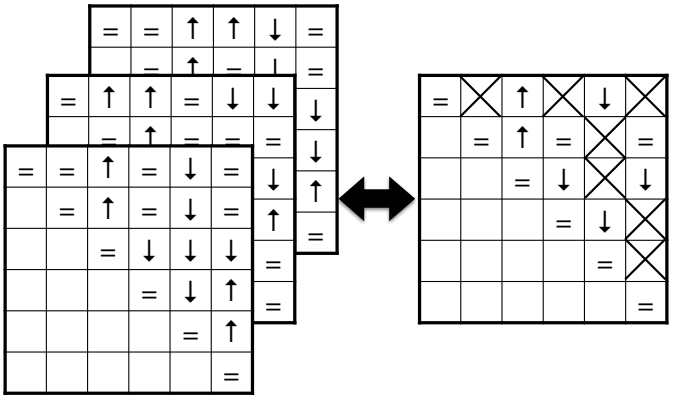

Fig. 4. Consensus matrix formed by combining individual matrices. We cross out the entries without agreement.

matrix summarizes the relative agreement across raters for a given emotional attribute. Figure 4 illustrates the process of constructing the CM. Assume that we are interested in segments in which X% of the raters agree on the trend. The approach compares the corresponding entries across all individual matrices. If X% of them agree on the trend, we set the entry of CM to the given trend (increasing, decreasing or similar). Otherwise, the entry is not considered since the evaluators did not reach agreement on the trend. We refer to this parameter as *agreement tolerance*. For example, an agreement tolerance of 66% would require that at least four out of six raters agree on a particular trend (assuming that there are six independent raters). In Figure 4, the CM is formed assuming 100% agreement between three raters. We cross out the entries without agreement.

An important advantage of the QA approach is that we can study optimum inter-evaluator agreement level. Notice that as the required percentage increases, the number of items in the training set (i.e. pairwise bins with agreement) will decrease; but so will the noise in the set. We would expect the optimum to depend on the distribution of evidence for a given attribute. We explore different agreement tolerances in the experimental evaluation (see Sec. V).

B. Rank-Based Classifiers

Table II presents the algorithm to derive relative QA-based labels and training the proposed rankers. The consensus matrix provides pairwise comparison between bins that can be directly used to train preference learning algorithms. Instead of detecting the underlying classes of the testing samples, as done in conventional machine learning problems, preference learning aims to rank the samples according to their relevance. Given an input query, the machine learning algorithm predicts a ranking order or ranking score for the documents associated with the query. Rank-based approaches have been very popular in areas such as information retrieval, web data mining, and artificial intelligence.

Rank-based approaches are generally formulated as pairwise comparisons of two samples, where the task is to select the preferred one [38]. We collect all the relative labels to create the set P (see Table II). The pair $(i,j) \in P$ if the i-th sample

TABLE II
ALGORITHM FOR TRAINING THE RANKERS WITH QA-BASED LABELS.

Training Algorithm
Given: N bins of training data
% STEP 1: Create individual matrices (IM) per trace
IM=zeros(N,N)
for $i, j < N$ do:
  if $b_i - b_j > t_{threshold}$
    IM(i,j) :"↓"
  if $b_j - b_i > t_{threshold}$
    IM(i,j) :"↑"
  else
    IM(i,j) :"="
end
Repeat for each trace
% STEP 2: Create consensus matrix (CM)
CM=zeros(N,N)
for $i, j < N$ do:
  if X% agreement for IM(i,j) across traces
    CM(i,j) =IM(i,j)
  else
    CM(i,j) ="X"
end
Define set P with all entries from CM other than "X"
% STEP 3: Train rank SVM
Estimate feature $\mathbf{\Phi_i}$ from segment b_i
Train binary SVM with features $f = (\mathbf{\Phi_i} - \mathbf{\Phi_j})$ to solve Eq. 5
return $\mathbf{w}$

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

$$(i,j) \in P \Longleftrightarrow \mathbf{w}^T \mathbf{\Phi_i} > \mathbf{w}^T \mathbf{\Phi_j}, \tag{4}$$

where the machine learning task is to find the optimum weight vector $\mathbf{w}$. The projections of $\mathbf{\Phi_i}$ and $\mathbf{\Phi_j}$ onto the $\mathbf{w}$ define the order of the samples. The weight vector $\mathbf{w}$ can be optimized by minimizing the objective function [38], [39]

$$\min_{\mathbf{w}} L(\mathbf{w^T}(\mathbf{\Phi_i} - \mathbf{\Phi_j}), P) + \frac{1}{2}\|\mathbf{w}\|^2 \tag{5}$$

where $L()$ is a loss function that depends on $(\mathbf{\Phi_i} - \mathbf{\Phi_j})$. Therefore, the ranking problem is equivalent to a binary classification problem where the features correspond to the differences of the feature vectors of the samples. Depending on the loss function, different approaches have been proposed. This study uses Rank-SVM [39], where the complexity parameter C of the SVM is set to 1. Notice that any linear SVM classifier can be used to train Rank-SVM by deriving relative labels.

The problem of ranking expressive behavior using relative labels can be addressed using rankers. We use the labels from the QA analysis to define the set P using the pair of bins rated as 'increasing' or 'decreasing'.

$$S = \{b_i, b_j || i - j| < n\} \tag{6}$$

Table III presents the algorithm for testing preference between two given speech segments. The algorithm projects their features into the optimum weight vector, selecting the segment with the highest value.

TABLE III
ALGORITHM FOR TESTING THE RANKERS

Testing Algorithm
Given: Bins b_1 and b_2
Estimate feature $\mathbf{\Phi_1}$ and $\mathbf{\Phi_2}$ from segments b_1 and b_2
if $\mathbf{w}^T\mathbf{\Phi_1} > \mathbf{w}^T\mathbf{\Phi_2}$.
return $b_1 > b_2$
else
return $b_1 < b_2$

IV. EXPERIMENTAL SETUP

This section presents the experimental setup of the proposed rankers trained with the relative labels. Section IV-A presents the implementation of the approach describing the selection of the parameters of the approach. Section IV-B describes the acoustic features and feature selection approach used for the analysis. Section IV-C describes the baseline labels used to compare the proposed QA-based labels.

A. Implementation of the Approach

An important parameter in the proposed approach is the size of the bins (L). The nature of the traces is such that they are noisy and grainy. One method of controlling the noise is to create bins that are long enough to make the average value reliable. Long bins also facilitate the estimation of reliable features over the bins. However, if the size of the bins is too long, they will not capture the time evolving nature of the trace. Cowie and McKeown [16] studied the effect of the bin size on the correlation between traces for a given session. With short durations, they showed an increase in correlation between traces as the bin size increased. However, this trend changed when the averaging function began to mask the inherent structure of the trace. There was considerable variation in the point at which this happened, but overall, the study suggested setting bin size in the range 1-3s. Here, we set L equal to 3s to have reliable features (see discussion on Sec. IV-B). Notice that assigning emotional descriptors at the segment level is a common approach in speech emotion recognition [2].

Another important parameter is the threshold used in Equations 1, 2 and 3. The minimum distance $t_{threshold}$ defines how different b_i and b_j need to be, to identify an increasing or decreasing trend. The annotations for the SEMAINE database are in the scale [-1,1]. While increasing the minimum distance generates pairs that are clearly different, the number of pairs for training decreases. This study considers two values for this parameter ($t_{threshold} = \{0.1, 0.2\}$).

The last parameter in our approach is the percentage of evaluators who are required to agree on direction, for an entry to be created in the consensus matrix (i.e., agreement tolerance). We evaluate the ranking performance with 66%, 80%, and 100% agreements. We report the results in terms of this parameter.

Each session in the SEMAINE database has an approximate duration of 5 minutes [18]. Each subject participated in multiple sessions. To estimate the QA-based relative labels, we search across all the sessions recorded from a given subject. With this approach, the selected pairs can be derived from different sessions of a given subject. The set of annotators is not consistent across sessions. When we compare annotations of segments from different sessions, they may come from different raters. While we acknowledge this limitation, we argue that the benefit of increasing the number of relative labels outweighs the problem associated with treating each trace independently. The results reported in Section V are significantly higher than the results obtained when the pairs are restricted within a given session (not reported on the paper). Notice that we do not consider segments coming from different speakers in the SEMAINE database to create relative labels – the selected pairs come from the same subject.

Table IV shows the number of relative labels that reach agreement with the QA method for different values of minimum distance ($t_{threshold}$) and agreement tolerance. Increasing the minimum distance ($t_{threshold}$) decreases the number of samples. The table also shows that when the agreement tolerance increases, the number of selected pairs decreases. These are intuitive results as fewer pairs meet the stricter thresholds. The pairs shown in the table define the material used in each part of the experimental evaluation.

B. Acoustic Features

While the SEMAINE database contains audiovisual information, we conduct the ranking experiments using only acoustic features. Emotion plays a key role in speech production affecting prosodic, spectral and voice quality features. Therefore, emotion classifiers are usually trained with multiple features describing various acoustic properties [2], [49]. The common approach consists in deriving *low level descriptors* (LLD) which are frame-by-frame features describing different acoustic properties. Then, global statistics referred to as *high level features* (HLF) are extracted for each sentence or speaking turn [50] (e.g., the mean of the fundamental frequency). Therefore, the feature vector is fixed regardless of the length of the segments. This study uses the popular acoustic feature set introduced by Schuller et al. [51] for the Computational Paralinguistic Challenge at Interspeech 2013. Table V lists the LLD estimated from speech. Table VI gives the HLF estimated from each LLD. The feature set contains 6373 HLF level acoustic features per segment. The features are extracted with the toolkit OpenSMILE [52]. Notice that extracting these features from fixed windows ignoring the underlying linguistic boundaries does not introduce major problems. There is data suggesting that the relevance of some features depends on the length of the natural unit in which they occur [53], but studies on emotion recognition have shown good performance when speech recordings are segmented into turns of fixed lengths [54]. In fact, we have successfully used this approach with the SEMAINE database using 0.5s windows [55]. Since we are using 3s segments, the global statistics will be even more reliable.

Well-known problems arise because the dimension of the feature vector is large relative to the number of samples in the emotion classification problems, and many of the features are likely to be highly correlated. The distinctive character of our

TABLE IV
NUMBER OF RELATIVE LABELS FROM THE CONSENSUS MATRIX. WE USE THE MINIMUM DISTANCE ($t_{threshold}$) TO DEFINE THE PAIRS. THE TABLE INCLUDES DIFFERENT TOLERANCE ON THE INTER-EVALUATOR AGREEMENT (*A*: AROUSAL, *V*: VALENCE, *P*: POWER, AND *E*: EXPECTATION).

$\mathbf{t}_{threshold}$	66% Agreement				80% Agreement				100% Agreement			
	A	V	P	E	A	V	P	E	A	V	P	E
Across-session condition												
0.1	117035	158021	100437	208678	63073	110002	45187	101000	27523	59667	12101	32237
0.2	64936	100423	54587	206192	34815	61373	22013	99397	15712	31522	4840	31595

TABLE V
THE SET OF FRAME-LEVEL ACOUSTIC FEATURES. THIS SET IS REFERRED TO AS *low level descriptors* (LLDS) IN THE COMPUTATIONAL PARALINGUISTIC CHALLENGE AT INTERSPEECH 2013 [51].

Spectral LLDs
RASTA-style filtered auditory spectrum bands 1-26 (0-8kHz)
MFCCs 1-14
Spectral energy 25-650Hz, 1k-4kHz
Spectral roll-off point 0.25, 0.50, 0.75, 0.90
Spectral flux, entropy, variance, skewness, kurtosis, slope
Slope, Psychoacoustic Sharpness, Harmonicity
Energy related LLDs
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-crossing rate
Voice LLDs
F0
Probability of voicing
Logarithmic HNR
Jitter (local, delta)
Shimmer (local)

TABLE VI
THE SET OF SENTENCE-LEVEL FUNCTIONALS EXTRACTED FROM THE LLDS (SEE TABLE V).

Base functionals applied to LLD and $\triangle$ LLD
Quartiles 1-3
3 inter-quartile ranges
1% percentile (≈min), 99% percentile (≈max)
Position of min/max
Percentile range 1%-99%
Arithmetic mean, Root Quadratic Mean
Standard deviation,Skewness, kurtosis
Contour centroid, Flatness
Relative duration signal is above/below 25/ 50/ 75/ 90% range
Relative duration signal is rising/falling
Relative duration LLD has positive/negative curvature
Gain of linear prediction (LP)
LP coefficients 1-5
Base functionals applied to LLD only
Mean of peak distances
Standard deviation of peak distances
Mean value of peaks
Mean value of peaks-arithmetic mean
Mean / Standard Deviation of rising/falling Slopes
Mean/ Standard Deviation of inter maxima distances
Amplitude mean of maxima/minima
Amplitude range of maxima
Linear regression slope, offset and quadratic error
Quadratic regression a ,b, offset and quadratic error
F0 functionals
Percentage of non-zero frames
Mean, max, min, standard deviation of segments length

ranking task suggests that we should avoid using a wrapper-based feature selection method that is optimized to a particular classifier. Instead we implement a two-step feature selection process for each emotional dimension. First, we reduce the set from 6373 to 500 using a gain ratio attribute approach that measures the information gain ratio of each attribute with respect to the class (redundant features may be selected). We select this approach, as implemented by the software WEKA, for its fast computation since it does not require either forward or backward feature selection. Then, we further reduce the set by using the *correlation feature selection* (CFS) method [56], also implemented with WEKA. The approach selects features that are correlated with the label. At the same time, it minimizes the redundancy between the features. The features are selected with the best first search method. It starts with an empty set. Then, it adds an attribute at a time based on the CFS criterion. Enabling the backtracking option allows us to remove selected features so as to prevent local optimum. The final reduced feature set contains 40 features for each preference-learning condition discussed in Section V.

C. Baseline comparison

For comparison, we use the approach that is most often used in preference learning to define relative labels using continuous emotional traces [19]. Instead of considering the trends in each individual trace, as we do with the QA based approach, we average the absolute values of the traces across different raters. Then, we estimate the average score per segment. We consider that one segment is preferred to another if the difference in their scores is above a margin. For consistency, we use the same minimum distance values used in the proposed QA based approach ($t_{threshold} = \{0.1, 0.2\}$). We also use the same segmentation (e.g., 3s segments), compensating the traces with the same evaluators' reaction lag (see Table I). We search for the relative pairs across session conditions, following the same framework presented for the QA based labels. Table VII presents the number of samples defined with this approach.

TABLE VII
NUMBER OF RELATIVE LABELS USING THE BASELINE METHOD, WHERE WE AVERAGE THE ABSOLUTE SCORES OF THE TRACES OVER THE BIN. WE USE THE SAME MINIMUM DISTANCE ($t_{threshold}$) TO DEFINE THE PAIRS (*A*: AROUSAL, *V*: VALENCE, *P*: POWER, AND *E*: EXPECTATION).

$\mathbf{t}_{threshold}$	Baseline Approach			
	A	V	P	E
Across-session condition				
0.1	105592	138656	98228	215835
0.2	42926	84002	41107	213335

This is still a rank-oriented approach. The difference is that it assumes the absolute values are enough to reveal emotional contrast between speech segments. In contrast, our proposed QA based labels discount absolute scores in favor

of ordinal relationships as early as possible – again, reflecting the underlying theory.

V. Results of Rank-Based Classifiers

This section describes the results of the rankers using the QA-based relative labels for arousal, valence, power and expectation.

In emotion recognition problems, it is important that the evaluation includes speaker-independent partitions, in which data from one subject is included either in the training or testing sets, but not in both. This approach is important to validate the generalization of the proposed approach. To maximize the usage of the database, we implement a 10-fold *leave-one-speaker-out* (LOSO) cross validation approach. In each fold, data from nine subjects is used for training and data for the remaining subject is used for testing. The training and testing relative samples are independently derived from these partitions using the corresponding thresholds. The feature selection approach described in Section IV-B is separately implemented in each fold. The final number of features is always 40, although it differs from fold to fold. By selecting the features per fold, we truly rely only on the training data, preserving the speaker independent partitions for training and testing the models. We train the models and select the features without considering the data from the subject used for testing the models. Notice that for real applications, we can use all the data for training, selecting a common feature vector.

We noticed that there are more pairs in which the evaluators agreed on increasing than decreasing trends. Therefore, a ranker that always prefers the second bin over the first bin will have accuracies over 50%. To avoid this bias in the testing set, we randomize the order of the two bins before they are evaluated by the ranker (i.e., chance is 50%).

We systematically train and test the proposed method using various combinations of parameters. We create matched and mismatched conditions between training and testing settings. For example, we expect to observe higher performance when the testing samples are selected with $t_{threshold} = 0.2$, instead of $t_{threshold} = 0.1$, as the distance separating the sample pairs needs to be higher, simplifying the ranking problem. In this case, it is interesting to evaluate the best approach to training the rankers (matched or mismatched conditions). Furthermore, the comparison with the baseline rankers requires to evaluate conditions where the testing set is defined with the baseline labels, instead of the QA labels.

Table VIII shows the ranking results of our experiments. The rows in the table correspond to the conditions we use to define the training labels, and the columns correspond to the conditions we use to define the testing labels. There are four main blocks in the table corresponding to the four emotional attributes considered in this study – arousal, valence, power, expectation. In each block, the columns are divided into four portions indicating the three conditions for agreement tolerance that we place on the testing samples (66%, 80%, and 100%) and the baseline labels. These columns are further divided into two columns indicating the conditions we use for the minimum distance parameter $t_{threshold}$ to define the testing samples (0.1, 0.2). Similarly the rows are divided based on the agreement percentage and minimum threshold parameters used to define the training samples. For example, we achieve 79.9% accuracy for arousal when the training labels are derived with a tolerance agreement of 80% and $t_{threshold} = 0.1$, and the testing labels are derived with a tolerance agreement of 66% and $t_{threshold} = 0.2$. The best performance per column (test condition) is highlighted in bold. Given the high number of training and testing pairs in each condition, testing the difference in proportion between two conditions in Table VIII may give statistical significant results, since most statistical tests are sensitive to the number of samples. Instead, we rely on the test of hypothesis of differences in population means for matched conditions [57], comparing only the difference between corresponding settings. Given that the number of matched conditions is small, ranging from 12 to 48, we use the one-tailed t-test, asserting significance at $p < 0.01$. This section discusses the results reported in this table.

TABLE VIII
Accuracy of pairwise comparisons using the proposed rank-based approach (chances is 50%). The labels for training and testing samples are defined by the minimum distance ($t_{threshold}$) and the agreement tolerance. It also reports results for baseline labels (66% = 66% Agreement, 80% = 80% Agreement, 100% = 100% Agreement, Base = Baseline).

Training condition		Testing condition: 66%		80%		100%		Base	
		0.1	0.2	0.1	0.2	0.1	0.2	0.1	0.2
		Arousal [%]							
66%	0.1	**73.7**	79.5	79.5	84.9	85.8	89.2	**71.9**	81.7
	0.2	73.3	79.5	79.3	85.1	86.0	89.3	71.4	81.8
80%	0.1	73.5	**79.9**	**79.7**	**85.4**	**86.2**	**89.7**	71.4	**82.0**
	0.2	72.1	78.4	78.2	84.4	85.0	88.9	70.3	80.6
100%	0.1	71.5	77.6	77.3	83.1	83.5	87.5	69.6	79.5
	0.2	70.2	76.3	76.1	81.7	82.3	81.2	68.5	78.0
Base	0.1	71.4	76.6	76.7	81.0	81.5	84.7	69.6	78.2
	0.2	68.4	74.0	72.9	78.2	77.9	81.2	67.5	75.6
		Valence [%]							
66%	0.1	60.2	62.0	61.4	63.7	**63.7**	64.7	**59.1**	60.2
	0.2	58.9	60.7	60.1	62.5	61.8	63.7	58.0	59.3
80%	0.1	60.0	61.0	60.4	62.7	62.1	63.9	58.0	59.4
	0.2	58.6	60.3	59.7	62.2	61.4	63.3	57.5	58.7
100%	0.1	59.8	60.5	59.9	62.0	61.5	62.7	57.3	58.6
	0.2	57.0	58.3	57.9	60.2	59.2	61.0	55.8	56.3
Base	0.1	**60.3**	**62.2**	**61.6**	**64.2**	63.4	**65.7**	59.0	**60.9**
	0.2	58.8	60.4	59.5	62.0	61.1	62.5	57.9	59.2
		Power [%]							
66%	0.1	61.8	64.7	65.2	68.1	69.7	71.0	61.2	66.3
	0.2	60.3	62.1	62.6	64.2	65.7	67.5	59.7	64.2
80%	0.1	61.5	64.2	64.6	67.2	68.6	71.0	60.6	65.4
	0.2	60.9	62.6	63.7	65.9	69.0	72.2	59.9	62.8
100%	0.1	**62.9**	**65.8**	**66.3**	**69.1**	**69.8**	72.7	**61.3**	**66.7**
	0.2	61.6	63.7	64.7	66.7	69.0	**73.3**	60.1	63.6
Base	0.1	60.1	62.5	62.8	64.6	66.6	67.7	58.6	61.6
	0.2	58.6	60.4	60.9	63.1	64.1	67.8	57.7	61.9
		Expectation [%]							
66%	0.1	**58.1**	**59.0**	**61.7**	**63.0**	**65.7**	**67.7**	**56.7**	**57.6**
	0.2	57.4	58.3	60.8	62.0	64.6	66.4	56.2	57.0
80%	0.1	57.6	58.3	61.1	62.0	65.0	67.0	56.2	56.9
	0.2	57.0	57.6	60.2	61.1	63.9	65.8	55.1	56.3
100%	0.1	57.3	58.2	60.5	61.7	64.2	66.2	56.0	56.8
	0.2	56.7	57.5	59.6	60.8	63.0	65.0	55.5	56.2
Base	0.1	56.5	57.6	59.2	60.9	62.2	64.5	55.4	56.3
	0.2	55.9	57.0	58.3	60.0	61.1	63.3	54.8	55.6

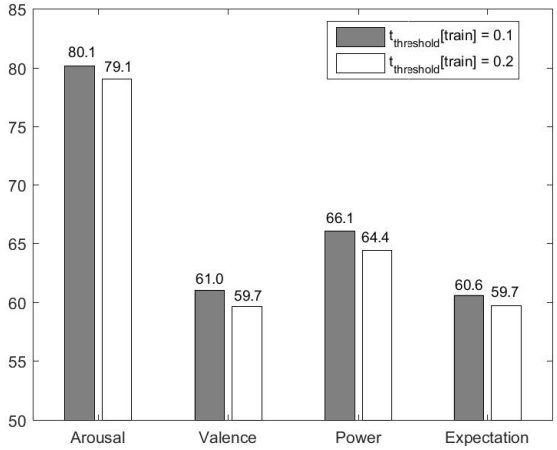

Fig. 5. Average accuracies of rankers trained with minimum distance ($t_{threshold}$) equals to either 0.1 or 0.2 across all testing conditions.

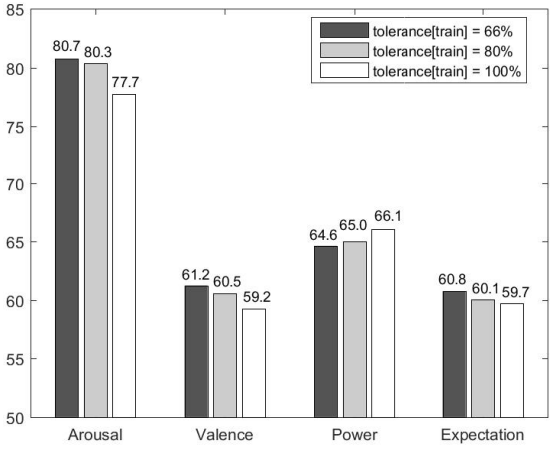

Fig. 6. Average accuracies of rankers trained with different agreement tolerance (66%, 80% and 100%) across across all testing conditions.

A. Comparison of Minimum Distance on the Training Set

The first evaluation compares the performance of the rankers in term of the minimum distance used to define the relative labels on the training sets, $t_{threshold}$. For a given emotional attribute, we estimate the average performance achieved across all the testing conditions when we use either $t_{threshold} = 0.1$ or $t_{threshold} = 0.2$ (i.e., 3 agreement tolerances [train] $\times$ 8 testing conditions = 24 matched conditions). While certain test conditions are easier to rank than others, as discussed in Section V-D, comparing these averages is fair as the same testing conditions are used for both values of $t_{threshold}$.

Figure 5 displays the average performance for $t_{threshold}$=0.1 and $t_{threshold}$=0.2 for arousal, valence, power and expectation. Training the rankers with smaller margin ($t_{threshold}$=0.1) provides higher performance, where the differences are statistically significant across the four emotional attributes. This result is unsurprising, for two reasons. First, we lose potential samples when we impose a stricter threshold. From Table IV, we observe an average loss of 38.7% (relative) of the samples as we increase the threshold from $t_{threshold} = 0.1$ to $t_{threshold} = 0.2$. Second, having a less strict threshold for training also avoids the mismatched problems of training with $t_{threshold} = 0.2$ and testing with $t_{threshold} = 0.1$. In this case, the minimum distance between the samples in the training set is higher than the samples in the testing set. The extra ambiguity in the labels on the testing set makes the ranking problem more challenging. Those effects mean that lower thresholds have an advantage.

B. Comparison of Agreement Tolerance on the Training Set

The second parameter in the proposed QA method is the agreement tolerance to determine the trends (Sec. III-A). For a given emotional dimension, we study the performance achieved by different agreement levels across various conditions (2 minimum distances [train] $\times$ 8 testing conditions = 16 matched conditions). Figure 6 shows the average performance for different agreement tolerances used to derive the training samples. We compare whether the differences are statistically significant between the three conditions: 66%, 80%, and 100% (matched paired t-test, one-tailed, $p < 0.01$). Table IX reports the results of the pairwise comparisons. A value of "1" indicates that the condition on the row produces significantly higher performance than the condition on the column (e.g., for valence, the performance for agreement tolerance of 66% is significantly higher than the performance for agreement tolerance of 80%). A value of "0" indicates that the differences are not statistically significant. A value of "-1" indicates that the condition on the column produces significantly better performance than the condition on the row.

For arousal, valence and expectation, the worst performance was when we defined the training labels with 100% agreement tolerance. The best performance for these three emotional dimensions was achieved with a 66% agreement tolerance, although the difference between 66% and 80% for arousal was not statistically significant. We conclude that imposing a stricter condition for the training labels does not necessarily improve results. Again, that is not surprising. Table IV shows that we lose an average of 49.6% (relative) of the samples when we go from 66% to 80% agreement tolerance. We also lose an average of 61.2% of the samples when we go from 80% to 100% agreement tolerance. Furthermore, defining the training labels with more strict agreement tolerance (e.g., 100%) creates mismatches when the testing labels are defined

TABLE IX
STATISTICAL TEST FOR PAIRWISE COMPARISONS OF RANKERS TRAINED ON DIFFERENT AGREEMENT TOLERANCE (MATCHED PAIRED T-TEST, ONE-TAILED, $p < 0.01$). A "1" INDICATES THAT ROW CONDITIONS IS SIGNIFICANTLY BETTER THAN COLUMN CONDITION. A "-1" INDICATES THAT COLUMN CONDITIONS IS SIGNIFICANTLY BETTER THAN RAW CONDITION. A "0" INDICATES THAT THE DIFFERENCES ARE NOT STATISTICALLY SIGNIFICANT.

Emotion	Agreement	80%	100%
Arousal	66%	0	1
	80%	-	1
Valence	66%	1	1
	80%	-	1
Power	66%	0	-1
	80%	-	-1
Expectation	66%	1	1
	80%	-	1

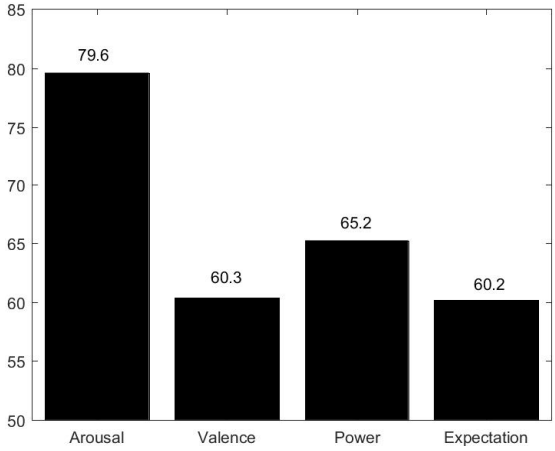

Fig. 7. Average accuracies of rankers for emotional dimensions across all testing conditions.

with less strict tolerance level (e.g., 66%). In this case, the training samples are less ambiguous than the testing samples. Nevertheless, for power, we observe a different pattern, where the best performance is achieved with a 100% agreement tolerance. Table IV indicates that power behaves differently from other dimensions in terms of agreement: the number of samples that meet the 100% tolerance criterion is very low. The obvious interpretation is that for this dimension, there is a sharp difference between a small number of clear cases and the rest; but that needs further investigation.

C. Comparison of Emotional Attributes

We now analyze the performance of the rankers trained using the QA method for different emotional attributes. Figure 7 shows the average performance achieved across all training and testing conditions for each emotional dimension (2 minimum distances $\times$ 3 agreement tolerances $\times$ 8 testing conditions = 48 matched conditions). Table X shows the statistical tests of pairwise comparisons between the four emotional dimensions: arousal, valence, power and expectation (matched paired t-test, one-tailed, $p < 0.01$). This table follows the same convention used for Table IX. Table X shows that the differences in performance between emotional dimensions are all statistically significant, with the exception of valence and expectation. Arousal achieves the best performance with an average accuracy of 79.6%. This value is remarkably high, considering the challenging task of detecting emotional traces in this spontaneous database. Valence and expectation achieve the lowest accuracies. There are well-known challenges in predicting valence just from speech [58], where very few acoustic features correlate with this emotional dimension.

D. Complexity of Test Samples

We next consider the complexity of the sets defined by the different criteria. That can be inferred by comparing the performances of rankers trained in the same way on testing sets defined with different agreement tolerances (Sec. V-B addresses the agreement tolerance on the training set). We find that whatever the training condition, accuracy is higher with test sets defined using a QA-based approach. The implication is that sets defined in that way are lower in complexity – i.e. more coherent.

TABLE X
STATISTICAL TEST FOR PAIRWISE COMPARISONS ACROSS EMOTIONAL DIMENSIONS (MATCHED PAIRED T-TEST, ONE-TAILED, $p < 0.01$). A "1" INDICATES THAT ROW CONDITIONS IS SIGNIFICANTLY BETTER THAN COLUMN CONDITION. A "-1" INDICATES THAT COLUMN CONDITIONS IS SIGNIFICANTLY BETTER THAN RAW CONDITION. A "0" INDICATES THAT THE DIFFERENCES ARE NOT STATISTICALLY SIGNIFICANT.

Emotion	Valence	Power	Expectation
Arousal	1	1	1
Valence	-	-1	0
Power	-	-	1

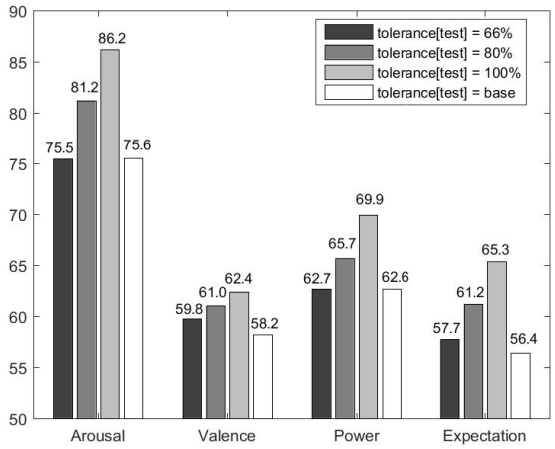

Fig. 8. Average accuracies of QA-based rankers across all training conditions, when the testing labels are defined with the QA-based method (66%, 80% and 100%) and the baseline method.

For each emotional dimension, Figure 8 shows the average performance across all the training conditions using QA-based approach when the testing labels are defined using the QA-based method (66%, 80% and 100%) and the baseline methods (i.e., 2 minimum distances [train] $\times$ 3 agreement tolerances [train] $\times$ 2 minimum distances [test] = 12 matched conditions). Table XI shows the results of the pairwise comparison (matched paired t-test, one-tailed, $p < 0.01$), following the same convention as the one used in Tables IX and X. Again, the result is intuitive. Reducing the agreement tolerance increases the number of ambiguous samples in the testing set, and reduces the performance. However, it is important that the QA parameters do affect complexity as we would expect. The performance of the QA-based rankers tested on the baseline labels achieve similar performance to the QA-based labels with an agreement tolerance of 66%. Table XI shows that testing the rankers using agreement tolerance of 80% or 100% produces significantly better performance than testing on the baseline labels. Samples from the baseline labels are derived using absolute values of the emotional traces, producing noisier labels that are harder to rank. Instead, the QA-based labels capture trends that are consistently annotated across evaluators, creating clearer labels which simplify the ranking problems.

TABLE XI
STATISTICAL TEST FOR PAIRWISE COMPARISONS OF RANKERS TESTED WITH LABELS DEFINED WITH THE QA-BASED METHOD (66%, 80% AND 100%) AND THE BASELINE METHOD (MATCHED PAIRED T-TEST, ONE-TAILED, $p < 0.01$). A "1" INDICATES THAT ROW CONDITIONS IS SIGNIFICANTLY BETTER THAN COLUMN CONDITION. A "-1" INDICATES THAT COLUMN CONDITIONS IS SIGNIFICANTLY BETTER THAN RAW CONDITION. A "0" INDICATES THAT THE DIFFERENCES ARE NOT STATISTICALLY SIGNIFICANT.

Emotion	Test	80%	100%	Base
Arousal	66%	-1	-1	0
	80%	-	-1	1
	100%	-	-	1
Valence	66%	-1	-1	1
	80%	-	-1	1
	100%	-	-	1
Power	66%	-1	-1	0
	80%	-	-1	1
	100%	-	-	1
Expectation	66%	-1	-1	1
	80%	-	-1	1
	100%	-	-	1

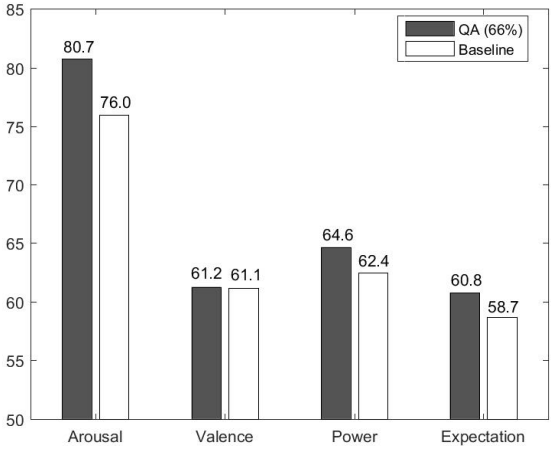

Fig. 9. Comparison between average accuracies of rankers trained with the proposed QA-based labels (66% agreement tolerance) and the baseline labels. With the exception of valence, training the rank-based rankers with the QA-based labels provides statistical significant improvements over rankers trained with baseline labels.

E. QA-based Labels versus Baseline Labels

We compare the performance of the rankers using the QA-based labels with the baseline method. Given the results obtained in Section V-B, we only consider the labels derived with agreement tolerance of 66%, which provides the best performance in most of the testing conditions. The performance of rank-based rankers trained with baseline labels are listed in the "Base" rows of Table VIII. Figure 9 gives the average accuracies across all train and test conditions (2 minimum distances $\times$ 8 testing conditions = 16 matched conditions). The proposed approach is significantly better than the baseline approach for all the emotional dimensions with the exception of valence (matched paired t-test, one-tailed, $p < 0.01$). The average improvement for arousal is over 4% (absolute). By comparing Figures 6 and 9, we observe that even rankers trained with labels derived with 80% and 100% agreement tolerance perform better than the baseline ranker for arousal, expectation and power. For these emotional dimensions, Table VIII shows that the best performance per testing condition is always a ranker trained with the QA-based labels (highlighted values), achieving improvements up to 5% (8% relative) compare to baseline labels. For valence, the differences between the QA approach with 66% agreement tolerance and the baseline method are not statistically significant. These results clearly demonstrate the advantage of learning relative trends for preference learning using the QA approach, instead of deriving the labels from absolute, aggregated values of the emotional traces.

F. Alternative Segmentation of the Bins

The last parameter in the approach that we evaluate is the size of the bins, which is set to 3s in previous evaluation. This section explores two alternative segmentations. The first alternative segmentation is reducing the bin size to 2s. This approach increases the number of segments, but reduces the reliability of the feature vector (see discussion in Section IV-B). The second alternative segmentation considers 3s bins with 2s overlap. Therefore, we have a bin every second. We still average the traces over each 3s bin to estimate its mean value b_j, but the resolution is improved by using overlapped bins. As expected, the number of selected pairs increases 859% when considering overlapping bins. We follow the same experimental procedure explained in Section IV, training the rankers with these alternative bin segmentations. We use matched conditions for the bin size for training and testing. To simplify the analysis, we only evaluated this approach with the testing labels derived using the baseline method, which is the most challenging testing condition (see Fig. 8).

Table XII shows the results, which are aggregated in Figure 10 for illustration, where we have 12 matched conditions (2 minimum distances [train] $\times$ 3 agreement tolerances [train] $\times$ 2 minimum distances [test] = 12 matched conditions). To simplify the comparison, we also report the results with 3s bins without overlap (last two columns in Table VIII). First, the performance drops when we decrease the bin size from 3s to 2s. We believe that using longer bins gives more reliable labels and features. Second, adding overlapped bins slightly improves the performance over 3s bins without overlap for power and expectation, although the differences are not statistically significant (matched pair t-test, one-tailed, asserting significance with $p < 0.01$). For arousal and valence, using overlapped bins reduces the ranker performance. This section shows that adding overlapped bins or reducing the size of the bins do not increases the performance of the proposed system. Using 3s bin provides a good tradeoff between emotion resolution and features reliability.

VI. DISCUSSION AND CONCLUSIONS

Both long-established theory and recent experience raise questions about the kind of evidence on which affective computing relies. It may be unrealistic to ask for the kind of description that is required by the obvious computational techniques. If so, there is a need to look for techniques that make use of the kind of information that people can deliver.

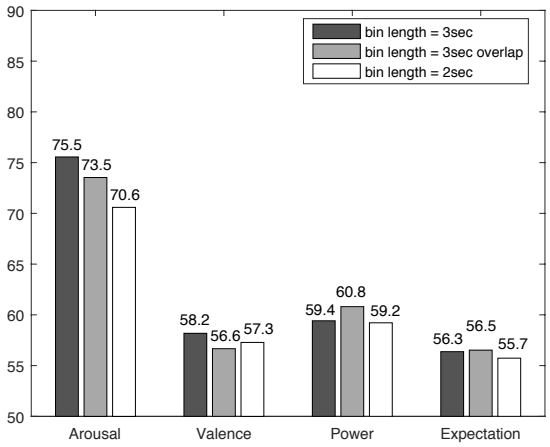

Fig. 10. Analysis of alternative segmentation of the bins. Average accuracies of QA-based rankers across all training conditions, when the testing labels are defined with the baseline method for different bin segmentations.

TABLE XII
PERFORMANCE OF THE RANKERS USING ALTERNATIVE BIN SEGMENTATION. THE TESTING CONDITION USES THE BASELINE LABELS, EMPLYING THE SAME BIN SEGMENTATION APPROACH USED FOR TRAINING (66% = 66% AGREEMENT, 80% = 80% AGREEMENT, 100% = 100% AGREEMENT, BASE = BASELINE).

Training condition		Testing using Baseline Labels					
		3s		3s with overlap		2s	
		0.1	0.2	0.1	0.2	0.1	0.2
		Arousal [%]					
66%	0.1	71.9	81.7	70.8	78.5	67.5	75.5
	0.2	71.4	81.8	70.7	78.4	67.5	76.0
80%	0.1	71.4	82	70.0	77.4	66.7	74.4
	0.2	70.3	80.6	69.6	77.0	66.0	74.5
100%	0.1	69.6	79.5	69.0	76.2	65.8	73.8
	0.2	68.5	78	68.6	76.2	65.8	73.8
Base	0.1	69.6	78.2	70.1	77.6	65.9	74.1
	0.2	67.5	75.6	70.3	77.8	64.2	73.3
		Valence [%]					
66%	0.1	59.1	60.2	57.3	58.3	57.5	58.9
	0.2	58	59.3	56.9	57.8	57.1	58.3
80%	0.1	58	59.4	56.3	57.2	57.1	58.4
	0.2	57.5	58.7	55.6	56.3	56.5	57.7
100%	0.1	57.3	58.6	56.2	57.1	56.3	57.4
	0.2	55.8	56.3	55.2	55.8	55.7	56.5
Base	0.1	59	60.9	57.7	58.7	58.0	59.5
	0.2	57.9	59.2	57.3	58.4	57.6	59.0
		Power [%]					
66%	0.1	61.2	66.3	59.9	62.4	57.9	59.6
	0.2	59.7	64.2	59.4	61.5	57.7	59.2
80%	0.1	58	59.4	59.3	61.4	58.6	60.5
	0.2	57.5	58.7	59.0	61.0	58.3	60.1
100%	0.1	57.3	58.6	60.3	62.9	58.8	61.0
	0.2	55.8	56.3	60.0	62.6	58.4	60.5
Base	0.1	59	60.9	59.4	62.1	57.7	59.4
	0.2	57.9	61.9	59.4	61.8	57.8	59.2
		Expectation [%]					
66%	0.1	56.7	57.6	56.3	56.4	56.0	56.1
	0.2	56.2	57.0	55.9	56	56	56.1
80%	0.1	56.2	56.9	56.2	56.3	55.7	55.8
	0.2	55.1	56.3	56.3	56.3	55.7	55.8
100%	0.1	56	56.8	57.1	57.2	55.4	55.4
	0.2	55.5	56.2	57.1	57.2	55.4	55.4
Base	0.1	55.4	56.3	56.0	56.1	55.6	55.7
	0.2	54.8	55.6	56.0	56.1	55.7	55.7

This study has demonstrated ways of learning from information that we have good grounds to call reliable. It considers a hybrid approach, which makes assumptions about the meaningfulness of averaging, but then treats the results as ordinal; and a thoroughgoingly non-parametric version, based on the QA framework. Both achieve respectable levels of performance, but it is satisfying that the approach with deeper theoretical roots also allows better performance.

A clear practical attraction of the approach, in either version, is that it can exploit methods of data collection that are already established (hence the SEMAINE database could be used), and that are not unrealistic in terms of the time and effort required. If a FEELTRACE-type record is used to construct an individual matrix, the process requires bn seconds of a rater's time, where b is the duration of a bin and n is the number of bins. Constructing it by direct comparison takes $bn(n-1)$ seconds of the rater's time for viewing alone. Expecting large databases to be created in that way would be another kind of unrealism.

More subtly, the QA approach provides a way of taking into account the particular relationship between signals and an attribute. Where signals are conventionalized, one would expect a sharp distinction between cases where there is information for a discrimination and cases where there is not. In that kind of case, lowering the threshold for consensus could only introduce noise.

We do not claim that the QA approach is ideal. However, it is well-motivated theoretically, and we have shown that it can be implemented. The result provides a way of picking out information that people can deliver. That clearly raises the question of how to exploit that information.

There are several avenues for exploration. Two contrasting alternatives illustrate the range.

The more conservative is to use evidence of change to construct traces that reflect consensus. Cowie and McKeown [16] describe a simple way of doing that. A basic 'consensus trace' can be constructed by giving each bin a height equal to the number of bins which, by consensus, are lower than it. The resulting contour can then be scaled to the mean and standard deviation of the raw data (which are considerably more reliable than patterns within the traces [16]).

The more radical takes up arguments that have been made for other reasons [26], to the effect that information about emotion is very unevenly distributed across time and modality. If so, it makes sense to explore a strategy of moving from evidence, as and when it is offered, to an appropriate response. That might either replace or complement efforts to form an integrated description of the person's state. There are some kinds of action that it clearly might be appropriate to take in response to evidence of a shift in emotional tone. Appraisal theory suggests an obvious one: try to establish what might have triggered it. Another is to invite clarification, verbally or by a gesture (e.g. raising an eyebrow). Developing systems that use that kind of strategy effectively is a long term challenge, with rich potential for interactions between psychology and affective computing.

Several more immediate directions invite systematic development. The evaluation included rankers trained with only

acoustic features. Our future work includes evaluations with visual and audiovisual features. We are also planning to evaluate the approach with other similar databases such as the RECOLA corpus [59]. A step further is to consider real applications. Instead of using conventional machine learning methods to recognize emotions, rankers of the kind that we have described may be more successful in rank-ordering expressive behaviors.

Behind all that is a point whose importance is hard to overemphasize. Human impressions of emotion are the root source of information on which affective computing rests. As a result, it is constrained by their structure. Relying on preconceptions about that structure – explicit, or implicit in the computational methods that are used – is not a satisfying strategy. It is also not necessary. Bringing psychology and computing together provides both motives and means to probe that particular aspect of human experience in a depth that has not been undertaken before. The work described here demonstrates that that is a real possibility.

ACKNOWLEDGMENT

This study was funded by the National Science Foundation (NSF) CAREER grant IIS-1453781.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] M. 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] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] Y.-H. Yang and 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.

[15] M. Soleymani, , G. Chanel, J. 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.

[16] 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

[17] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder, "'FEELTRACE': An instrument for recording perceived emotion in real time," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK: ISCA, September 2000, pp. 19–24.

[18] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[19] 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.

[20] R. Herrnstein and E. Boring, *A source book in the history of Psychology*. Cambridge, MA, USA: Harvard University Press, January 1965.

[21] S. Stevens, *Psychophysics*. New York, NY, USA: John Wiley & Sons Inc, February 1975.

[22] G. Fechner, *Elements of Psychophysics*. Holt, Rinehart and Winston, January 1975.

[23] S. Stevens, "To honor Fechner and repeal his law," *Science*, vol. 133, pp. 80–86, January 1961.

[24] N. H. Anderson, *Methods of Information Integration Theory*. New York, NY, USA: Academic Press, August 1982.

[25] K. Scherer and H. Ellgring, "Multimodal expression of emotion: Affect programs or componential appraisal patterns?" *Emotion*, vol. 7, no. 1, pp. 158–171, February 2007.

[26] R. Cowie, G. McKeown, and C. Gibney, "The challenges of dealing with distributed signs of emotion: theory and empirical evidence," in *International Conference on Affective Computing and Intelligent Interaction and Workshops (ACII 2009)*, Amsterdam, The Netherlands, September 2009.

[27] K. R. Scherer and T. Bänziger, "On the use of actor portrayals in research on emotional expression," in *Blueprint for affective computing: A sourcebook*, K. R. Scherer, T. Bänziger, and E. B. Roesch, Eds. Oxford, England: Oxford university Press., November 2010, pp. 166–176.

[28] H. Aviezer, R. Hassin, J. Ryan, C. Grady, J. Susskind, A. Anderson, M. Moscovitch, and S.Bentin, "Angry, disgusted, or afraid? studies on the malleability of emotion perception," *Psychological Science*, vol. 19, no. 7, pp. 724–732, July 2008.

[29] R. Levenson and J. Gottman, "Marital interaction: Physiological linkage and affective exchange," *Journal of Personality and Social Psychology*, vol. 45, no. 3, pp. 587–597, September 1983.

[30] A. Metallinou, A. Katsamanis, Y. Wang, and S. Narayanan, "Tracking changes in continuous emotion states using body language and prosodic cues," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 2288–2291.

[31] L. Devillers, R. Cowie, J. C. Martin, E. Douglas-Cowie, S. Abrilian, and M. McRorie, "Real life emotions in French and English TV video clips: an integrated annotation protocol combining continuous and discrete approaches," in *International conference on Language Resources and Evaluation (LREC 2006)*, Genoa,Italy, May 2006, pp. 1105–1110.

[32] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *IEEE International*

Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), Prague, Czech Republic, May 2011, pp. 2376–2379.
[33] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. To Appear, 2015.
[34] 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, 2015.
[35] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*, Santa Barbara, CA, USA, March 2011, pp. 827–834.
[36] B. Vlasenko and A. Wendemuth, "Annotators' agreement and spontaneous emotion classification performance," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1546–1550.
[37] 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.
[38] O. Chapelle and S. Keerthi, "Efficient algorithms for ranking with SVMs," *Information Retrieval*, vol. 13, no. 3, pp. 201–215, June 2010.
[39] 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.
[40] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 448–457, February 2008.
[41] 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.
[42] ——, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," *Computer Speech & Language*, vol. 29, no. 1, pp. 186–202, January 2014.
[43] 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.
[44] S. Parthasarathy and C. Busso, "Defining emotionally salient regions using qualitative agreement method," in *Interspeech 2016*, San Francisco, CA, USA, September 2016.
[45] 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.
[46] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 85–90.
[47] ——, "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.
[48] J. Nicolle, V. Rapp, K. Bailly, L. Prevost, and M. Chetouani, "Robust continuous prediction of human emotions using multiscale dynamic cues," in *International conference on Multimodal interaction (ICMI 2012)*, Santa Monica, CA, USA, October 2012, pp. 501–508.
[49] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, January 2011.
[50] B. Schuller, A. Batliner, D. Seppi, S. Steidl, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and V. Aharonson, "The relevance of feature type for the automatic classification of emotional user states: Low level descriptors and functionals," in *Interspeech 2007 - Eurospeech*, Antwerp, Belgium, August 2007, pp. 2253–2256.
[51] 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.
[52] 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.
[53] R. Cowie and E. Douglas-Cowie, "Prosodic and related features that signify emotional colouring in conversational speech," in *The Role of Prosody in Affective Speech*, S. Hancil, Ed. Berlin, Germany: Peter Lang Publishing Group, July 2009, pp. 213–240.
[54] J. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4940–4943.
[55] J. Arias, C. Busso, and N. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 278–294, January 2014.
[56] M. A. Hall, "Correlation based feature-selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, April 1999.
[57] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*. Upper Saddle River, NJ, USA: Prentice-Hall, 2006.
[58] 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.
[59] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

Srinivas Parthasarathy received his BS degree in degree in Electronics and Communication Engineering from College of Engineering Guindy, Anna University, Chennai, India (2012) and MS degree in Electrical Engineering from the University of Texas at Dallas - UT Dallas (2014). During the academic year 2011-2012, he attended as an exchange student The Royal Institute of Technology (KTH), Sweden. He is currently pursuing his Ph.D in Electrical Engineering at UT Dallas. At UT Dallas, he received the Ericsson Graduate Fellowship during 2013-2014. He joined the Multimodal Signal Processing (MSP) laboratory in 2012. In summer and fall 2014 he interned at Bosch Research and Training Center working on Audio Summarization. His research interest includes the area of affective computing, human machine interaction, and machine learning.

Roddy Cowie is Emeritus professor of Psychology at Queens University, Belfast. He has used computational methods to study a range of complex perceptual phenomenaperceiving pictures, the experience of deafness, what speech conveys about the speaker, and, in a series of EC projects, the perception of emotion, where he has developed methods of measuring perceived emotion and inducing emotionally colored interactions. Key outputs include special editions of Speech Communication (2003) and Neural Networks (2005), and the HUMAINE Handbook on Emotion-Oriented Systems (2011). He is a member of the IEEE.

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests include digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human inter- action, in-vehicle active safety system, and machine learning methods for multimodal processing. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.