Predicting Emotionally Salient Regions using Qualitative Agreement of Deep Neural Network Regressors

Srinivas Parthasarathy, *Student Member, IEEE,* and Carlos Busso, *Senior Member, IEEE*

Abstract—Automatic emotion recognition plays a crucial role in various fields such as healthcare, *human-computer interaction* (HCI) and security and defense. While most of previous studies have focused on the recognition of emotion in isolated utterances, a more natural approach is to continuously track emotions during human interaction, identifying regions that are highly emotional. This study proposes a framework to define emotionally salient regions (hotspots), which we then attempt to dynamically detect. Our proposed approach defines hotspots relying on the *qualitative agreement* (QA) method, which searches for trends across continuous-time evaluations provided by different raters for arousal and valence. We illustrate the benefits of the QA method over averaging absolute values of the traces without considering trends across evaluators. After defining hotspot regions, we propose a deep learning framework to automatically detect these emotional hotspots. The proposed method relies on an ensemble of *bidirectional long short term memory* (BLSTM) regressors, trained on individual emotional traces provided by the evaluators, which are combined to automatically detect emotional hotspots. An appealing fusion approach to combine these regressors is to rely again on the QA method, which detects emotional salient regions with F1-scores as high as 60.9% for arousal and 50.4% for valence on the RECOLA dataset.

Index Terms—Emotion recognition, affective computing, emotionally salient regions, regressors of attribute-based descriptors

1 Introduction

An important problem in *human-computer interaction* (HCI) is recognizing affective behavior. We express emotions through modification of cues in multiple modalities including speech, which is a flexible and appealing modality for current interfaces [1]. Automatic systems that can recognize emotional content from speech have enormous potential in various applications. Most research approaches for emotion recognition have considered either identifying emotional categories such as happiness, anger and sadness, or predicting the value of an emotional attribute such as arousal (calm versus excited) and valence (positive versus negative). While performing this task, these approaches have mostly focused on short, pre-segmented utterances. However, naturalistic human interactions are fairly neutral with few segments conveying emotional content, so these approaches are not ideal for identifying and tracking natural emotions in real life applications. There is a need to build systems that are dynamic in nature, can continuously track and predict affective behavior in time, and can identify emotionally salient regions.

Few studies have focused on continuously predicting emotional attribute scores over time [2], [3], [4], [5], [6], [7]. An important aspect of these studies is the availability of time-continuous emotional evaluations such as the ones collected with FEELTRACE [8] or ANNEMO [9]. These evaluations are generally collected from multiple annotators who rate the perceived emotional content from audio-visual clips. The traces allow us to explore new research questions such as changes in emotional content [10], [11], and defining emotional hotspots [12]. The key challenge is that these scores can be noisy due to various factors such as evaluator bias [13], time-variant reaction-lag [14], [15], [16] and low inter-evaluator agreement [1]. We need strategies that derive reliable information from existing time-continuous emotional labels.

This work focuses on both defining and detecting emotionally salient regions (hotspots) from time-continuous annotations. We define hotspots as regions, marked by either very high or very low values for particular emotional attributes (e.g., a possible hotspot for valence would be segments with positive emotions). This study shows that defining hotspots from scratch is a challenging, expensive and time-demanding task, resulting in labels with low inter-evaluator agreements. Instead, we propose to define hotspots using a modified version of the *qualitative agreement* (QA) method proposed by Cowie and McKeown [17]. The proposed method individually quantizes the emotional traces of each evaluator, searching for trends. Then, it aggregates the trends based on the consensus across evaluators. This process naturally captures the trends agreed by evaluators. The analysis shows that hotspots defined with the proposed method are better than hotspots defined by first aggregating the emotional traces of the evaluators and then quantizing the absolute trace into hotspots (i.e., absolute approach). The key additional benefit is the ability of the QA method to detect and avoid unreliable regions in the time-continuous traces, where raters do not reach a consensus on their evaluations.

After defining the hotspots, we present alternative frameworks to automatically estimate these emotionally

- *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:axb124530@utdallas.edu, sxp120931@utdallas.edu, busso@utdallas.edu).*

Manuscript received January 13, 2017; revised xxx.

salient regions using acoustic features. The proposed frameworks rely on *deep neutral networks* (DNN) with *bidirectional long short-term memory* (BLSTM) models, which have been successfully used to capture temporal/contextual information in speech processing, natural language processing and computer vision tasks. The most successful framework consists of training an ensemble of BLSTM-DNN regressors, which are later combined to identify hotspots. The BLSTM-DNN regressors are separately trained for each of the traces provided by the raters. To fuse the ensemble, we also rely on the QA method, identifying trends across the individual estimations. This approach achieves the best performance with F1-scores as high as 60.9% for arousal hotspots, and 50.4% for valence hotspots on the RECOLA database.

The contribution of this paper is the formulation of speech emotion recognition using an alternative and appealing framework. Instead of relying on standard binary or multiclass classification problems, we propose to determine emotional salient regions during human interactions. This is an appealing problem since it uses existing time-continuous evaluations to derive ground truth labels, so further annotations are not needed. This framework also offers flexible solutions for practical applications reducing the analysis to segments that are detected as *emotionally salient*. To achieve this goal, this study (1) defines ground-truth for emotional hotspots using the QA-based framework, and (2) proposes a novel fusion framework of regressors to automatically recognize these hotspots.

The paper is organized as follows. Section 2 describes relevant studies which have focused on dynamic analysis of emotion. The section also presents the QA method, previous studies using DNN, and the RECOLA and SEMAINE databases used in this study. Section 3 gives a formal definition of hotspots. It evaluates the agreement between raters in defining hotspots from scratch, comparing the labels with the ones achieved with the proposed approach based on the QA method. Section 4 describes the alternative frameworks proposed for detecting hotspots. Section 5 describes the experimental evaluations and the results. Section 6 concludes the paper with discussion and future directions in this area.

2 Background

2.1 Related Work: Dynamic Analysis of Emotion

The predominant approach in affective computing is to analyze emotional behaviors at the turn level, without making distinctions on the emotional fluctuation within the segment. Previous work on automatic emotion recognition systems have focused on recognizing discrete emotional categories [18], [19], [20], or detecting emotional attribute values [21]. Studies on categorical emotions tend to use sentence-level annotations, where a single descriptor is assigned regardless of the duration of the turn. The problem is then formulated as a multi-class classification task, where the goal is to identify the target class. Studies relying on emotional attributes have used sentence level descriptors or time-continuous evaluations such as traces provided by toolkits such as FEELTRACE [8] or ANNEMO [9]. Even with time-continuous evaluations, many studies have formulated their problems as a classification problem [22], [23], where the averages of the traces across the segments are used as sentence-level annotations. By analyzing isolated speaking turns, we discard contextual information which plays an important role in perceiving and characterizing emotional cues.

Few studies have focused on dynamically detecting emotions over time. Most of these studies rely on predicting time-continuous emotional traces. The surveys of Gunes and Pantic [24] and Gunes and Schuller [2] discussed some of these studies. Wöllmer et al. [25] provided arguments showing the benefits of designing machine learning solutions for emotional attributes as oppose to categorical emotions. This study evaluated various techniques to estimate the values of arousal and valence such as LSTM, *support vector regression* (SVR) and *conditional random field* (CRF). However, the predictions were estimated at the turn level, after defining the speaking turns.

There are studies that have attempted to continuously predict emotional attributes, departing from sentence based analysis. Nicolaou et al. [3] compared the performance of SVR and BLSTM techniques for training regressors on continuous emotional traces using the SEMAINE database. The study considered visual and acoustic features, evaluating feature and decision level fusion approaches. Metallinou et al. [26] proposed to use *Gaussian Mixture Model* (GMM) to track emotional dimensions at the frame level. They evaluated the proposed approach on recordings from the USC CreativeIT database [13], [27], considering not only speech, but also body motion. Ringeval et al. [4] conducted a thorough evaluation of continuous prediction of emotional dimensions on the RECOLA database using *recurrent neural networks* (RNNs) implemented with LSTM. The study considered the effect of fusing various features, and the size of the window to track emotional dimensions. They concluded that LSTMs were capable of learning the dependencies between continuous ratings from multiple raters and that a decision level fusion of the models led to better performance than feature level fusion. The frameworks proposed in this study rely on regressors for emotional attributes, building upon these studies.

Analyzing temporal evolution of emotions also opens new research questions that are very useful in practical applications. Studies have attempted to detect change of emotions within a dialog [10], [11], [28]. This problem is related to detecting emotional salient segments. Metallinou et al. [28] mapped the observed body language and prosodic cues to the emotional states of the subjects, considered as a hidden variable, using GMMs. They employ *maximum likelihood estimation* (MLE) between the observed and hidden variables at each time instant. To incorporate the time dynamics, the hidden variables and observed variables are augmented with derivatives, capturing information from previous frames. Models are trained and tested on five-second segments where the goal is to continuously capture the changes in emotional attributes. The predicted emotional attributes showed good correlation for arousal and dominance traces on the USC CreativeIT database. The results indicated that the models were better at capturing relative changes than absolute changes of the emotional attributes. Huang et al. [11] and Huang and Epps [10] proposed a martingale framework based on GMMs to detect the instant when the emotion of a speaker change.

Their work focused on changes in categorical emotions. The framework is based on the concept of exchangeability, where a sequence of random variables are considered exchangeable if their joint probability remains unchanged regardless of any permutation of the elements. A martingale process [29] is used to test the hypothesis of exchangeability. A strangeness value is measured and used to detect how different a point is from a defined model, predicting changes in the emotion with this metric. These studies focus on consecutive speaking turns from the IEMOCAP database to construct changes between emotional states.

Studies have also attempted to directly detect emotional salient regions. Lin and Lee [30] hypothesized that emotion perception is thin-sliced in nature, where global affect scores can be accurately predicted by identifying and using regions of high emotional significance. They showed that systems trained using only 20%-30% of the data from the entire session performed well at predicting the global emotion score of the session. They used a mutual information criterion to pick emotionally rich segments, using acoustic features and continuous traces provided by evaluators. Vydana et al. [31] proposed a framework to detect emotionally salient regions for emotion recognition. To identify these significant regions, they modeled the physiological constraints in human speech between neutral and non neutral segments. Their performance was evaluated on discrete categories of emotions. An improvement of 11% was reported while using only emotionally significant regions compared to the entire segments.

2.2 Emotional Labels

Natural human interaction comprises complex and ambiguous emotions [32], [33]. A common approach to describe emotion is to derive emotional labels (ground-truth) collected from perceptual evaluations by multiple annotators. Several studies have noticed poor inter-evaluator agreement due to factors such as difference in perception, emotional bias and the use of contextual information [1], [13], [21], [34], [35]. The lack of agreement inherently affects the performance of an emotion recognition system trained with these labels [1]. While inter-evaluator agreement on absolute scores have been shown to be poor, studies have shown the consistency in detecting relative trends in emotional behavior [17], [36], [37], [38], [39]. Instead of asserting an absolute score, evaluators are more reliable in asserting changes in emotional behavior (i.e., one segment is more positive than another). This observation inspires us to use QA to define hotspots.

2.3 Qualitative Analysis (QA)

The QA approach was proposed by Cowie and McKeown [17] to identify local trends across multiple evaluators in time-continuous evaluations. Their study used traces collected with FEELTRACE [8]. The perceptual evaluation consists of a *graphical user interface* (GUI) where the axes correspond to specific emotional attributes. The extreme values in the axes correspond to the extreme values for those emotional attributes. An evaluator watches the stimulus, perceives the emotional content, and moves the mouse cursor reflecting the perceived level of the emotional attribute. The interface continuously captures the location of the cursor creating emotional traces.

The QA maps time-continuous evaluations into ordinal matrices capturing relative trends across evaluators. The first step captures the relative trends in the traces of an evaluator by forming the *individual matrices* (IMs). Figures 1(a) illustrates the process, where we first segment the traces into N bins of equal lengths (3s in this study). Then, we estimate the average value of the trace within each bin, denoting this value as b_i with $i \in \{1 \dots N\}$. The IM is created with relative comparisons between the values of the bins. If $i < j$, we define a *fall* when $b_i - b_j$ is greater than a threshold (Equation 1), and a *rise* when $b_j - b_i$ is greater than a threshold (Equation 2). Otherwise, we consider that the bins are *similar* (Eq. 3). $t_{threshold}$ is a parameter of the QA method.

$$b_i - b_j \quad > t_{threshold} \tag{1}$$
$$b_j - b_i \quad > t_{threshold} \tag{2}$$
$$|b_j - b_i| \quad < t_{threshold} \tag{3}$$

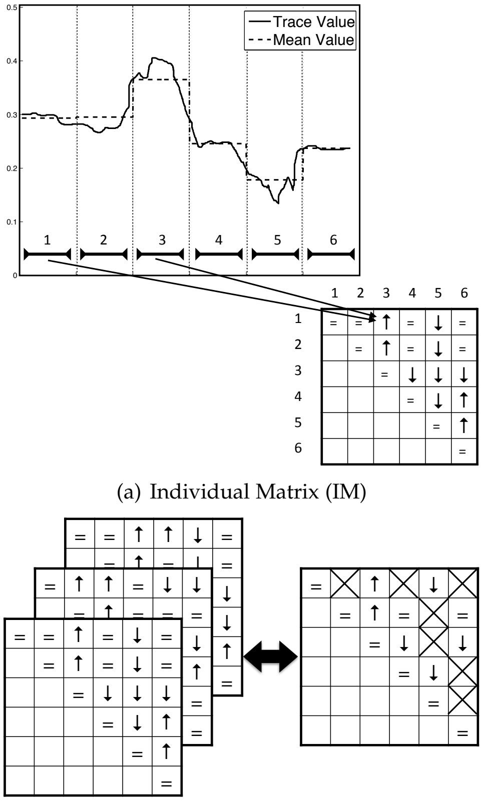
The next step aggregates the IMs to form a *consensus matrix* (CM). Figure 1(b) illustrates the procedure, which aims to capture the agreement between evaluators. If "X%" of the evaluators agree on a trend (entries from different IMs), the corresponding entry in the CM is set with the trend (rise, fall or similar). The variable X is referred to as the *tolerance agreement* and is another parameter of the QA method. Entries that fail to reach an agreement are labeled as segments without consensus ("X" in figure 1(b)).

Parthasarathy et al. [38] utilized the QA method to build preference learning algorithms to rank emotions. This study proposes to use a modified version of the QA method to define hotspots (Section 3.3).

2.4 Regressor Based on BLSTM

Recently, solution based on *deep neural networks* (DNNs) have achieved groundbreaking performance in many fields, including emotion recognition [7]. RNNs have shown to be beneficial in capturing the temporal information required to track time-continuous traces. RNNs establish recursive connections between units modeling dynamic temporal patterns. However, RNNs suffer from the vanishing gradient problem where over time the error in the gradients used to train the RNNs either explode or exponentially vanish. To address this problem, a new class of RNNs was introduced by Hochreiter and Schmidhuber [40] named *long short term memory* (LSTM). Unlike traditional neural networks, which contain sigmoidal activation nodes, LSTMs contain a memory cell to store information as well as three multiplicative gates: the input, forget and output gates. While input and output gates have the normal functions, the extra forget gate controls whether to retain or forget previous cell state memory. Therefore, RNN employing LSTM architecture can capture the temporal information in long sequences.

Regressors trained with LSTMs have shown superior performance in predicting emotional attributes [25], [41], [42]. Wöllmer et al. [25] showed that LSTMs performed better at regression tasks compared to other methods such



(a) Individual Matrix (IM)

(b) Consensus Matrix (CM)

Fig. 1. Formation of the *individual matrix* (IM) from time-continuous traces, and the formation of *consensus matrix* (CM).

as *support vector regressors* (SVR). Wöllmer et al. [23], [41] demonstrated that the long term context modeling of LSTM units along with the temporal information contribute to overall performance. Metallinou et al. [28] used LSTMs to continuously track emotional trends using multiple modalities. While LSTMs use context from one direction, generally the past values, *Bi-directional LSTMs* (BLSTMs) can be trained to learn context from both past and future values. Given the success of this framework, we implement our classifiers and regressors with BLSTM to detect hotspots over time-continuous attributes.

2.5 Database

This study uses the RECOLA database [9], a French corpus that captures affective behavior during dyadic interactions. The participants were asked to solve a collaborative task, communicating through video conference. Multimodal data was continuously recorded from the dyadic pair including video, audio, *Electrocardiogram* (ECG) and *Electrodermal Activity* (EDA). The dataset contains data from 46 participants, from which data for 23 people were released to the research community. We consider these 23 sessions for this study (12 female, 11 male).

The sessions were emotionally annotated using arousal and valence using the ANNEMO toolkit, a website based interface to collect time-continuous annotations. The interfaces has a slider with a scale between -1 and +1, capturing values at a fixed frame rate of 40ms. The traces are post processed to remove annotator bias (local zero-mean normalization) and the reaction delay between evaluators (synchronization) [9]. There is also a reaction lag in the evaluators. It takes some time for the rater to listen the emotional content, judge the stimulus, and react by moving the slider. Previous studies have proposed methods to compensate for the reaction lag. Studies have identified reactions lag that are specific to each annotator as well as constant delay across the annotators [15], [16], [43]. For this work, we consider constant delay across annotator. Valstar et al. [44] studied this reaction lag for the annotations of the RECOLA database. The reaction lags were identified as 2.8s for arousal and 3.6s for valence. For consistency with other studies, we use these delays to compensate for the reaction lag. The consistency between annotations is measured using the Cronbach's alpha, obtaining α =0.8 for arousal and α =0.75 for valence, which show acceptable consistency across the annotations.

To validate the methodology proposed in this study, we use the SEMAINE corpus [45] to replicate the evaluation. The SEMAINE database consists of dyadic conversations between an operator and a user, where the operator assumes a given personality trying to induce emotions from the user. The dialogs are emotionally annotated using FEELTRACE by different evaluators. The details can be found in McKeown et al. [45]. We use all the sessions with at least 6 raters (48 sessions from 11 speakers). Following previous work on the reaction lag of raters in the SEMAINE database [15], we use a delay of 2.84s for arousal, and 3.68s for valence.

3 Hotspot Definition

During daily interactions, we expect to observe mostly neutral behaviors. At certain times, speakers will externalize emotional reactions as they respond to contextual information. We have studied these segments in Parthasarathy and Busso [12] which we refer to as hotspots. We define these regions as having either low or high values for a given emotional attribute (e.g., arousal, valence). This section provides a framework to define hotspots from time-continuous traces which builds upon the method proposed in Parthasarathy and Busso [12].

After motivating the use of emotional hotspots (Sec. 3.1), we present the analysis in three phases. First, we perform experiments to define hotspots from scratch, without relying on existing evaluations. We elaborate on the complexity of the task (Sec. 3.2). Second, we build upon our previous study to define hotspots using time-continuous traces using QA-based labels (Sec. 3.3). For comparison, we also define hotspot by averaging the absolute scores of the traces selecting regions with either low or high values for the emotion attribute. Third, we compare the QA-based and absolute-based methods, illustrating their key differences.

3.1 Modeling Emotions with Hotspots

Studies on speech emotion recognition have mostly focused on the analysis of pre-segmented speech segments. If we are interested in addressing more realistic scenarios, we should expect long, unsegmented recordings. During human interaction, most interactions are emotionally neutral. Therefore, it is important to identify salient regions with emotional behaviors, focusing the analysis (and resources) on the regions that are more relevant. Modeling emotional hotspots offers a principled way of recognizing deviations from natural neutral behavior without the need of predefined class-specific models. This is especially beneficial in naturalistic scenarios where the emotional behaviors are ambiguous [46], and the emotions are not constrained to predefined classes. Another advantage of modeling emotions with hotspots is the capability to continuously track emotions removing the need of pre-segmenting the recordings into speaking turns. Likewise, the emotionally salient regions can provide important cues to capture the global emotional state during an interaction. This is beneficial in predicting the emotional state for long conversations.

From a modeling perspective, we have argued that ordinal descriptors are more reliable than nominal or interval annotations [39]. Studies have shown that we are more consistent in making relative comparisons than absolute assessments (e.g., this sentence is happier than the previous one). The framework is consistent with this argument, where our goal is to predict salient regions that deviate from neutral behaviors (e.g., tracking changes of behaviors instead of characterizing the behavior of isolated speaking turns). Therefore, we argue that using emotional hotspots is also more meaningful than common approaches using pre-segmented speaking turns.

From an application perspective, using emotional hotspots have several advantages with clear implications in many domains including healthcare, call centers, and surveillance. In healthcare, emotional hotspot detection can be used in longitudinal recordings to continuously track or predict mental state of patients with mood disorder. The framework offers a principled approach for spoken summarization using emotional speech [47]. In call-center applications, the use of emotional hotspot can be used to retrieve highly aroused conversations for quality control. In surveillance, emotional hotspot detection is an ideal framework to retrieve emotional speech, reducing and prioritizing the recordings to be examined by experts.

3.2 Defining Hotspots from Scratch

An intuitive method to define hotspots would be to start from scratch, asking evaluators to identify emotionally salient regions. We define initial hotspots on the RECOLA database through perceptive evaluations to motivate the advantages of the proposed approach relying on existing evaluations. Three evaluators watched a subset of the audiovisual recordings using the OCTAB toolkit [48], which provides an online framework allowing the evaluators to select segments in a video. After watching the sessions, they were asked to mark and select emotional segments for arousal and valence. These segments correspond to regions where the emotional content significantly differs from neutral behavior. The unmarked segments are then labeled as neutral. The evaluations were independently conducted on 10 sessions for arousal and 10 sessions for valence. The evaluations were conducted by non French speaking evaluators who relied exclusively on the acoustic and visual cues, ignoring lexical information.

To study the complexity of the task and the reliability of the hotspots identified by the raters, we determine the Fleiss' Kappa on the three evaluators. The Fleiss kappa provides a measure of reliability between multiple annotators for multiple nominal values. For the purpose of this experiment, we have three evaluators and three nominal values: high, low and neutral. Table 1 provides the kappa values for low, neutral, high regions and the overall kappa for the annotations. The kappa values range from κ =-0.10 to κ =0.34, indicating the complexity of defining hotspots from scratch. Using these inconsistent labels to train machine learning algorithms is not an adequate alternative. Moreover, this procedure to define hotspots is time-consuming as evaluators have to watch the entire interaction multiple times to identify hotspots (hotspot is a relative concepts so contextual information is important). All these drawbacks motivate us to look for alternative methods to determine ground-truth hotspots relying on existing annotations.

TABLE 1
Reliability of the ground truth hotspot labels using Fleiss' Kappa.

Dimension	Region-wise κ			Overall κ
	Low	Neutral	High	
Arousal	-0.10	-0.10	0.18	-0.04
Valence	0.00	0.25	0.34	0.2536

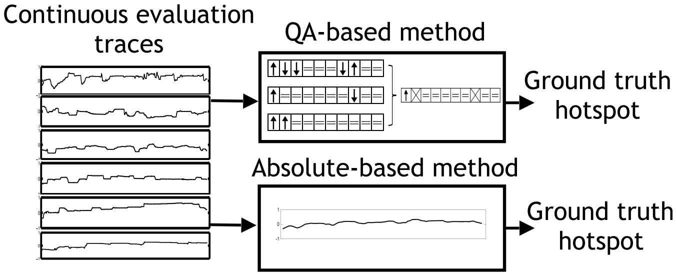
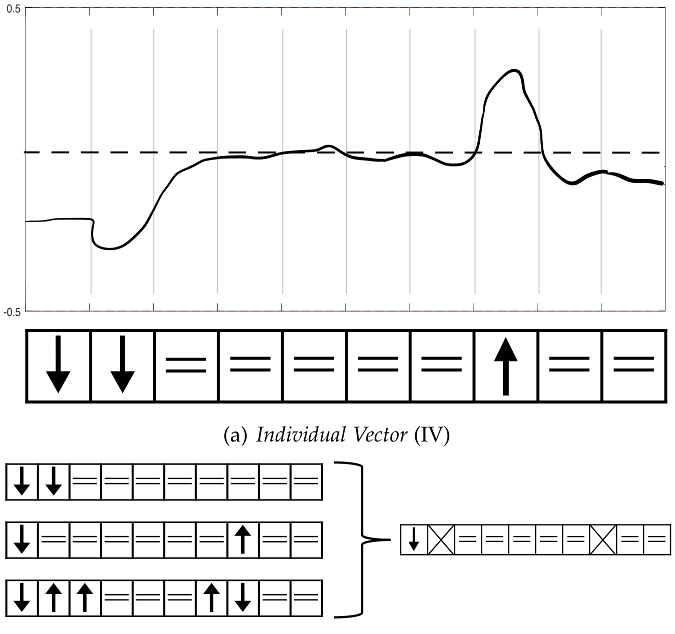


Fig. 2. The figure illustrates the framework for defining hotspots.

3.3 Hotspots with QA-based and Absolute Methods

Since defining hotspots from scratch is both time consuming and inconsistent, we leverage the existing time-continuous evaluations to define hotspots. As discussed in Section 2.5, the RECOLA database was annotated by six annotators for arousal and valence dimensions using the ANNEMO toolkit. This study presents two frameworks to define hotspots illustrated in Figure 2. The first approach uses relative trends relying on a modified version of the QA method. The second approach uses absolute values by setting thresholds over the average trace values across the evaluators. Before estimating the hotspots for these approaches, we normalize individual traces. For each of the six raters, we estimate the mean and standard deviation of all his/her evaluations. Then, we use these values to normalize the traces subtracting the rater's mean and dividing by the rater's standard deviation. The normalization makes the different traces comparable before defining hotspots. Notice that this section analyzes the validity and reliability of the hotspots and it does not require cross-validation. For the hotspot detection framework (Sec. 4), the traces will be normalized using only the labels from the training set.

QA-based method: The original QA approach was explained in Section 2.3, which captures the relative trends



(a) *Individual Vector* (IV)

(b) *Consensus Vector* (CV)

Fig. 3. The figure illustrates defining hotspots through the QA method. It shows the formation of the individual vector (IV) from continuous evaluation traces and the formation of consensus vector (CV) from IV.

between segments in the traces. We propose a modified version of the QA approach to define emotional hotspots. Figure 3 illustrates the proposed procedure. The normalized traces are segmented into bins as usual, estimating the average value for each bin. This procedure is applied to all six traces. Since hotspots are defined as segments with strong deviation from neutral behavior, bins must be compared to a common reference rather than with each other. Therefore, we create an *Individual Vector* (IV) for each individual trace, whose entries are labeled as either *high* (Eq. 4), *low* (Eq. 5) or *neutral* (Eq. 6) based on their deviation from the median value of the trace, b_{median}. Figure 3(a) shows this process.

$$b_i - b_{median} \quad > t_{threshold} \tag{4}$$
$$b_{median} - b_i \quad > t_{threshold} \tag{5}$$
$$|b_i - b_{median}| \quad < t_{threshold} \tag{6}$$

Similar to the original implementation of the QA approach, we define a threshold $t_{threshold}$ to create these classes. We can view the IVs as the individual hotspots for each evaluator. Finally, IVs are aggregated to form the *Consensus Vector* (CV) as shown in Figure 3(b). Entries of the CV are marked based on the consensus amongst the six individual IV entries, following the same procedure used to create the CM in Section 2.3.

While the reliability of hotspots would increase with a stricter agreement, the percentage of hotspots would significantly decrease. Therefore, we fix the tolerance agreement equal to 66% (at least four out of six raters have to agree on a trend). Most importantly, places where we fail to achieve a consensus are marked as no consensus regions and no decision is made on the hotspots for these regions.

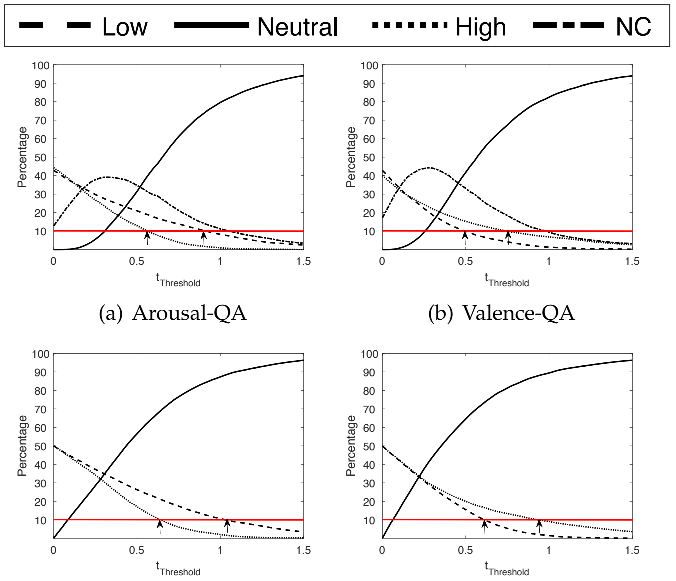
Absolute-based method: To compare the hotspots defined with the QA-based method, we define hotspots through the averaging of the absolute values of individual traces (Method II in Fig. 2). Instead of considering trends across the individual evaluations, we combine them by averaging the normalized traces to form one absolute mean trace. This trace is then segmented into bins and each bin is compared against the median value using a distance, $t_{threshold}$, to form our hotspots. To summarize, the difference between the QA-based and absolute-based methods is the order in which we quantize and aggregate the traces. In the QA-based method, the individual evaluations are first quantized into a discrete space (based on distance from the median) and aggregated based on the agreement between evaluators. In the absolute-based method, the individual evaluations are first aggregated by averaging the absolute scores, followed by quantization into discrete levels. Naturally, the relative trends across evaluators is lost when averaging the emotional traces in the absolute-based method. The QA-based method exploits these trends to obtain the emotional hotspots.

For both methods, we create the bins b_i by averaging the traces over a window of 3s. The window is shifted by 250ms to introduce smoothness for the regression tasks introduced in Section 4. The parameter $t_{threshold}$ defines the hotspots. When the threshold is small, most regions are labeled as hotspots with few neutral regions (see Eqs 4-6). An increase in the threshold will increase the neutral regions, since the deviations from the median values should be higher to consider a bin b_i as a hotspot. An effective way to formalize hotspots, and define $t_{threshold}$, is to fix a percentage of the regions to be labeled as hotspots. For this study, we consider approximately 10% of total samples to be hotspot segments corresponding to high regions, and 10% corresponding to low regions. We use this criterion for both methods (QA-based labels and absolute-based labels). Notice that this criterion depends on the type of interaction on the data. For highly emotional interactions, we may expect to have more than 10% of the segments as hotspots.

We increase the threshold by steps of 0.025 ranging from 0 to 1.5. The percentage of hotspots under different regions with respect to the threshold are shown in Figures 4(a) and 4(b) for the QA method, and Figures 4(c) and 4(d) for the absolute-based method. The thresholds that achieve 10% of hotspot regions are highlighted in the figure with an arrow. We use these values in the rest of the evaluation. We avoid any post processing, such as smoothing the hotspots or using a median filter. We aim to capture even the spikes in the hotspot definition using machine learning frameworks. Table 2 shows the final amount of data in each region (high, low, neutral, no consensus) with the selected thresholds. A considerable portion of the data, falls under the no consensus region for the QA-based method (24.1% for arousal and 25.4% for valence). These regions receive a label when we use the absolute-based method.

3.4 Reliability and Validity of Emotional Hotspots

There are two main differences between the absolute-based and QA-based methods used for defining the hotspots. The first difference is the detection of ambiguous segments. While the absolute-based method reaches a decision for every bin, the QA-based method excludes bins without consensus, where the evaluators do not agree. Figure 5 shows



(a) Arousal-QA

(b) Valence-QA

(c) Arousal-Average

(d) Valence-Average

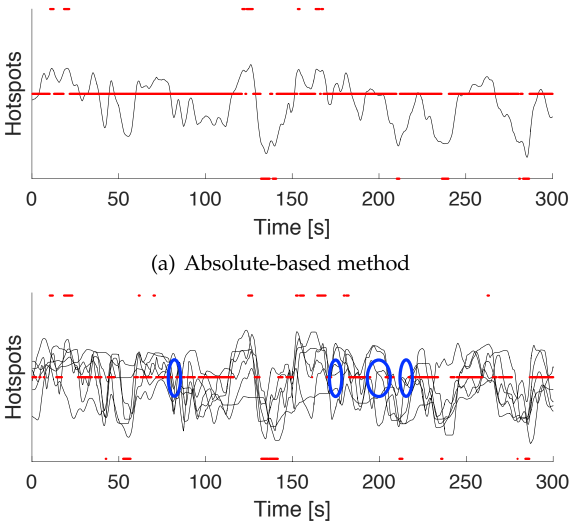
Fig. 4. Percentage of hotspot segments under different thresholds for the QA-based and absolute-based approaches on the RECOLA dataset. QA-Based approach: Arousal-Low: 0.925; Arousal-High: 0.575; Valence-Low: 0.5; Valence-High: 0.75. Avg-Based approach: Arousal-Low: 1.025; Arousal-High: 0.65; Valence-Low: 0.6; Valence-High: 0.925.

TABLE 2
Final percentage of data under different categories for the selected thresholds in Figure 4. The results are obtained on the RECOLA dataset (NC: no consensus).

Attribute	Hotspot Definition	Low [%]	Neutral [%]	High [%]	NC [%]
Arousal	Absolute-based	10.1	80.3	9.6	0.0
	QA-based	9.8	56.6	9.5	24.1
Valence	Absolute-based	10.5	79.3	10.1	0.0
	QA-based	9.9	54.7	10.0	25.4

the hotspots and neutral regions defined using both approaches for the arousal traces of session 48 of the RECOLA database. The segments on top of the traces are high hotspots, the segments below the trace are low hotspots, and segments at the middle of the traces are neutral regions. Figure 5(a) shows the average trace where each segment is split into three classes: high, low or neutral. Figure 5(b) shows all the individual traces from which we estimate the trends to detect hotspots. The figure shows segment without a decision, since the evaluators were not consistent in their assessment (see circles in Fig. 5(b) as examples).

The second difference is that the QA-based method captures trends agreed by majority of the raters and, therefore, it is expected to be more reliable. Scenarios where the traces of one or two raters differ from the others impact the absolute-based method, but they do not affect the QA based method. Table 3 shows the Fleiss Kappa value of hotspots for each of the six raters with respect to the absolute-based and QA-based methods. This analysis compares the salient segments annotated by each rater with the hotspot labels derived from all the traces. We establish these individual hotspots for each rater using the corresponding thresholds used for their definition. For the QA-based method, this approach is



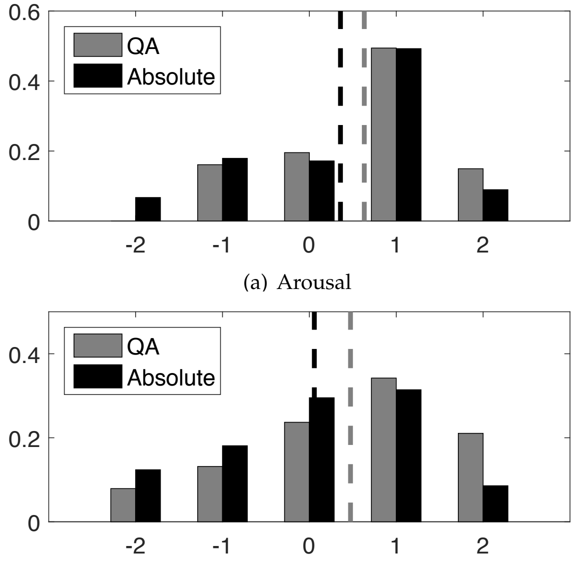
(a) Absolute-based method

(b) QA-based method

Fig. 5. Example of hotspots defined on arousal traces using the absolute-based and QA-based methods on the RECOLA dataset. The bars on top of the traces represents high hotspots, the bars below the traces represents low hotspots and the bars in the middle of the traces represents neutral segments. Circles in Figure 5(b) illustrate segments with no consensus.

equivalent to compare individual vectors with the consensus vector. For the absolute-based method, this approach compares the segments found by using the thresholds over each trace and over the average trace. Table 3 shows that raters have higher agreement with the QA-based hotspots ($\kappa_{Abs.} = 0.33$ versus $\kappa_{QA} = 0.52$ for arousal; $\kappa_{Abs.} = 0.35$ versus $\kappa_{QA} = 0.52$ for valence). The improvements in reliability are due to the ability not only to ignore regions of no consensus, but also to identify trends agreed by most raters. To verify this claim, we remove the segments without consensus in the evaluation of the absolute-based hotspots. The results are given in the columns labeled as "Abs.+QA" (note that information from the QA-based method is used to identify these ambiguous segments). While the values for κ are higher after removing ambiguous segments, the results are still lower than using the trends in the QA-based approach. The last row of Table 3 shows the mean agreement between hotspots in the SEMAINE database. The results are similar to the ones obtained with the RECOLA database, with higher agreement for the QA-based framework.

We also address the validity of the annotation of emotion hotspots (i.e., whether the label measures what it is expected to measure). For this purpose, we conducted perceptual evaluations to validate the use of QA-based method to select emotional hotspots. Nine raters were asked to evaluate the hotspots on a five-point Likert-like scale [(-2) strongly disagree, (+2) strongly agree]. Hotspots from all 23 sessions of the RECOLA database were evaluated with three evaluations per session. Since we are interested in the relative difference between the absolute, and QA-based methods for defining hotspots, we only evaluated mutually exclusive hotspots (i.e., common hotspots selected by both definition methods were not evaluated). Furthermore, hotspots with short durations are hard to evaluate as they lack contextual information. Therefore, we selected only those hotspots with



(a) Arousal

(b) Valence

Fig. 6. Perceptual evaluations of hotspots defined with QA-based and absolute methods using a five-point Likert-like scale (-2 strongly disagree, +2 strongly agree) on the RECOLA dataset. Bars show the distribution of the results. Dashed lines correspond to the means.

TABLE 3
Agreement of hotspots chosen by each rater with respect to the overall hotspots. Agreement is measured using Fleiss Kappa and Pearson Correlation coefficient for QA-based approach (QA), Absolute-based approach (Abs.) and Absolute-based approach without the segments identified by the QA-based approach without consensus (Abs. + QA).

	Fleiss Kappa					
	Arousal			Valence		
	RECOLA					
	Abs.	QA	Abs.+ QA	Abs.	QA	Abs. + QA
Rater1	0.33	0.51	0.43	0.44	0.64	0.55
Rater2	0.41	0.62	0.54	0.39	0.59	0.51
Rater3	0.35	0.52	0.45	0.33	0.47	0.42
Rater4	0.32	0.53	0.42	0.39	0.53	0.48
Rater5	0.32	0.48	0.42	0.21	0.36	0.31
Rater6	0.27	0.43	0.38	0.41	0.56	0.51
Mean	**0.33**	**0.52**	**0.44**	**0.35**	**0.52**	**0.45**
	SEMAINE					
Mean	**0.31**	**0.54**	**0.43**	**0.38**	**0.55**	**0.48**

duration longer than two seconds. Figure 6 shows the distribution of the scores. For arousal, the QA-based hotspots and absolute hotspots achieve mean score of 0.63 and 0.36, respectively. For valence, the QA-based hotspots and absolute hotspots achieve mean score of 0.47 and 0.06, respectively. For both arousal and valence, hotspots defined with the QA-based method are significantly better than hotspots defined with the absolute method (z-test, p-value<0.05). This evaluation shows the advantages of defining hotspots using the QA-based method over the absolute method.

4 Hotspot Detection Frameworks

Having defined a framework to label hotspots, our next research goal is to create machine learning methods to automatically predict them. This study considers three approaches to detect hotspots, displayed in Figure 7. Framework I formulates the problem as a classification problem with three classes corresponding to high hotspot, low hotspot and neutral regions. Framework II presents an alternative approach, where we first perform regression on time-continuous traces obtained by averaging the traces across evaluators. We predict hotspots by adding thresholds to the predicted traces. Framework III builds an ensemble of regressors by separately training our models with individual traces in the annotation. The results of the predicted traces are combined, defining the hotspots. Since the ground truth labels for hotspots are independently defined using QA-based method and absolute-based method, we separate train the classifiers and regressors using these two set of labels. This section describes the details for these frameworks.

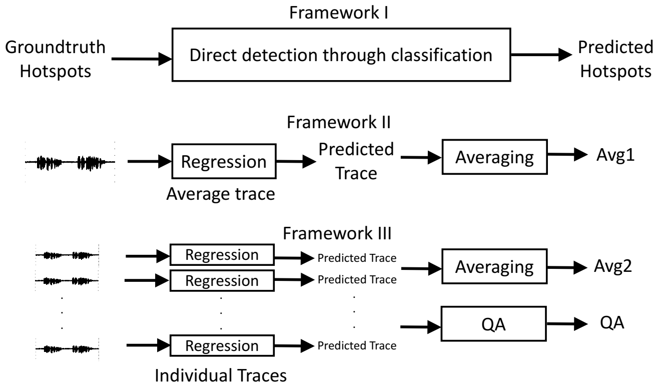


Fig. 7. The figure illustrates the frameworks for predicting hotspots.

4.1 Framework I (Baseline): Direct Prediction of Hotspots

Framework I directly detects hotspots formulating the task as a classification problem with three classes corresponding to high, low and neutral regions. When the classifier is not confident in its output, the segment is labeled as no consensus. The objective is to predict the hotspot or neutral regions given the acoustic features. We use hard labels when using hotspot labels defined with the absolute-based method. However, for the labels derived with the QA-based method, we have regions without consensus. Therefore, we use soft labels with the probability of a segment to belong to each class, avoiding adding an extra class. The probabilities are derived from how many evaluators agree on the trend. For example, if four out of six raters agree on a high hotspot and the other two agree on a neutral region, the label for this segment is 0.66-high, 0.33-neutral, and 0-low.

We train a *deep neural network* (DNN) with two hidden BLSTM layers, each of which has 256 nodes. A dropout with probability 0.5 is used on the activations of the hidden layers to prevent overfitting the model. A softmax layer is used at the output of the DNN to find the probability of the sample belonging to each class, where we use the cross entropy as the loss function to back propagate the error. During the evaluation of the models, we compare the results with the labels defined by the absolute-based (i.e., low, high, and neutral) and QA-based (i.e., low, high, neutral and no consensus) approaches. In the case of hotspots defined with the QA-based method, a segment can have no consensus (i.e.,, ambiguous cases). When the selected class has probability below $p < 0.66$, the classifier is not confident, so we label these regions as having no consensus. This approach is consistent with the tolerance agreement used in the definition of the hotspots.

4.2 Framework II: Regression on average traces

Learning to directly discriminate hotspots into discrete categories is a hard task, especially for values near the threshold. We propose an alternative framework to detect hotspots. Framework II trains a regressor to predict the actual value of traces, formulating the problem as a regression problem. We use the same approach described in Section 3.3 to normalize the traces, where the parameters of the normalization are estimated for each of the six raters across the entire recordings in the training set. Then, we average the traces for a given session providing the baseline labels to train the regressors. Previous studies such as Ringeval et al. [4], Trigeorgis et al. [5], and He et al. [49], have reported good performance of regressors trained on time-continuous traces on the RECOLA database, so this is an appealing approach.

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{7}$$

The regressors are implemented with BLSTM architecture (see Sec. 2.4). This study uses a simple network where each BLSTM block contains a single memory cell. The input, output and forget gates are activated by sigmoidal units. We use a two layered neural network, where each layer has 256 BLSTM nodes. Neural networks are trained for 100 epochs with a learning rate of $1e^{-}6$ and a momentum of 0.95. A dropout rate of 0.5 is used for the hidden layers to avoid overfitting the model. BLSTM networks are trained with sequences of frames (in our case, bins). We train the regressors with the entire sequence in each session of the RECOLA database, which correspond to sequences of 1,200 frames (300s). We use a linear unit at the output with *concordance correlation coefficient* (CCC) (Eq 7) as the criterion for back propagation of the regressors. We select CCC over *mean square error* (MSE) since previous regression studies have shown better performance when estimating emotional attributes [5], [50]. The use of MSE as the objective function often leads to predictions around the mean of the true distribution, which reduces the range of the predicted values. The CCC objective maximizes the Pearsons correlation (ρ) between the predicted and true values while minimizing the difference between their means ($(\mu_x - \mu_y)^2$). By achieving both goals, the range of the predicted values increases, as we have observed in other regression problems [51]. The network is trained with the Keras toolkit with a TensorFlow backend. Once the traces are predicted, we estimate the mean and standard deviation of all the predictions of the traces in the training set. These parameters are used to normalize predictions in the development and testing sets by subtracting the mean and dividing by the standard deviation.

After estimating the normalized traces, we identify hotspots regions when the distance from the median value of the predicted trace is greater than a threshold $pred_{threshold}$. This threshold does not need to be the same as the one used to define the hotspots. It can be set to maximize performance over the development set. Notice that this approach is similar to the absolute-based approach to define hotspots, where the difference is that we use the normalized predicted traces instead of the actual traces.

4.3 Framework III: Ensemble of Regressors

An alternative approach is to train an ensemble of regressors which are later combined using similar approaches used in the definition of hotspots (absolute-based and QA-based approach). We train each regressor in the ensemble with the individual traces provided by the raters, creating one regressor per trace (i.e., six regressors). We use the same RNN-BLSTM structure used for the regressors in Framework II (Sec. 4.2). Since each regressor captures the bias of a given evaluator, the predicted traces are separately normalized for each regressor by estimating its mean and standard deviation across the sentences in the training set. These regressor-dependent parameters are used in the development and testing sets.

We consider two approaches to combine the predicted traces to detect hotspots. First, we use the absolute-based method where we average the predicted traces, selecting segments above or below a given threshold ($pred_{threshold}$). Second, we use the QA-based method using the predicted traces, selecting segments based on the trends in the predictions. We set the tolerance agreement to 66% using the margin $pred_{threshold}$. Again, the key difference in Framework III with the approach to define hotspot is that the fusion is conducted on the predicted traces, instead of the actual traces. This fusion approach using the QA framework is novel, providing a suitable solution for hotspot detection.

5 Experimental Results

This section describes the evaluation conducted to detect hotspots with the proposed methods.

5.1 Features

While the RECOLA and SEMAINE corpora includes multiple modalities collected during the recordings, we only consider speech features to predicting hotspots. We use the *Geneva Minimalistic Acoustic Parameter Set* (GeMAPS) [52], as our acoustic feature. GeMAPS contains hand picked features for affective recognition problems. While most feature sets are large and are computed by brute force methods, the GeMAPS feature set is minimalistic in nature. The features are selected based on their value in previous studies on affect recognition, their theoretical significance and their potential to capture affective changes in speech. Due to the reduced dimension of this set, feature selection is not needed, facilitating reproduction of this study by other researchers. We use the extended parameter set eGeMAPS [52]. The common procedure involves extracting 25 *low-level descriptors* (LLD) which are features extracted on a frame-by-frame basis. Once extracted, a set of global functionals (e.g., arithmetic mean) are estimated on the LLDs to extract *high level features* (HLFs). All features are extracted with the

TABLE 4
Predicted thresholds, $pred_{threshold}$, on the RECOLA dataset, optimized on the development set for different frameworks that utilize the threshold. The hotspot definition method determines the ground-truth for the evaluations and it is shown in the first column.

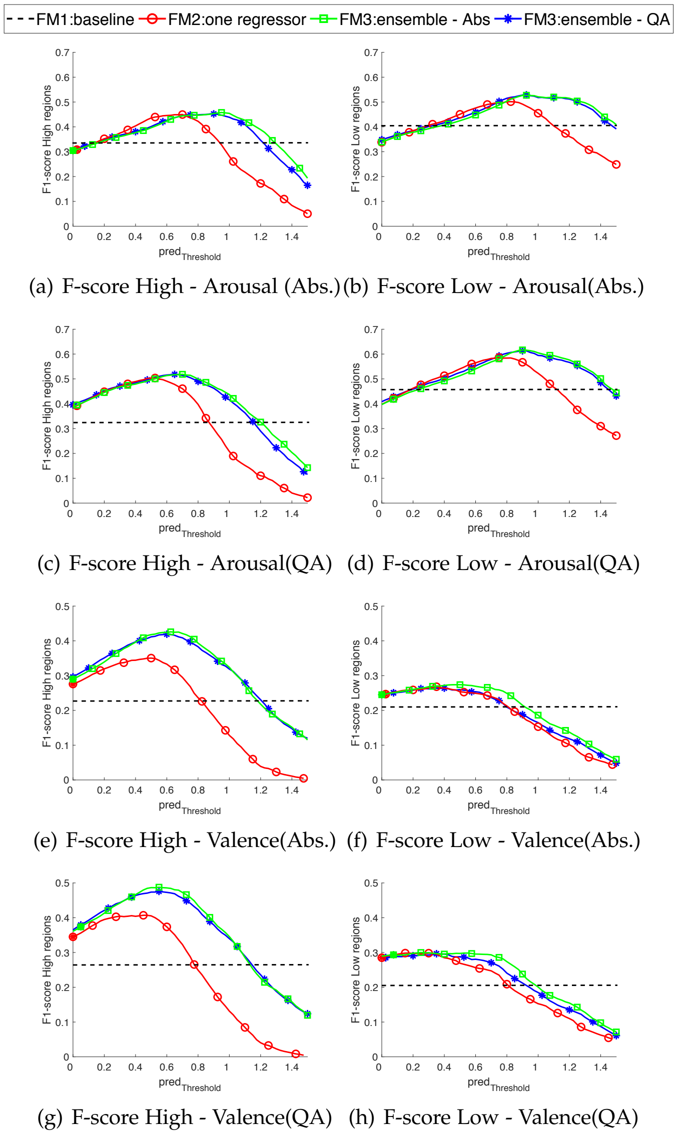
Definition Method	Detection Method	F1-Score	
		L	H
Arousal			
QA-based	Framework II: one regressor	0.85	0.475
	Framework III: ensemble – Abs.	0.925	0.7
	Framework III: ensemble – QA	0.825	0.625
Absolute-based	Framework II: one regressor	0.85	0.65
	Framework III: ensemble – Abs.	0.925	0.75
	Framework III: ensemble – QA	0.825	0.875
Valence			
QA-based	Framework II: one regressor	0.1	0.375
	Framework III: ensemble – Abs.	0.275	0.575
	Framework III: ensemble – QA	0.25	0.525
Absolute-based	Framework II: one regressor	0.375	0.375
	Framework III: ensemble – Abs.	0.375	0.65
	Framework III: ensemble – QA	0.375	0.625

OpenSMILE toolkit [53]. Overall, the feature set contains 88 features.

We use a frame size of 20ms to estimate LLDs. To keep the predicted hotspot traces consistent with the bins used in Section 3.3, we use a 3s window to extract functionals followed by a shift of 250 ms. Our previous study showed that using 3s bins provides segments that are long enough to estimate robust and stable features, but short enough to prevent dynamic changes of emotions within the segments [38].

5.2 Results on the RECOLA Database

We implement the experimental evaluation with a 11 fold cross-validation approach with speaker independent partitions. The 10 partitions have data from two speakers (female and male) and one partition has data from three speakers (two female and one male). With these 11 partitions, we define the training (eight partitions), development (two partitions) and testing (one partition) sets. Notice that with this approach, all the recordings from one of the speakers are exclusively in the training, development or testing sets, ensuring the generalization of the models. Since we are interested in recovering hotspot regions while at the same time avoiding classifying neutral regions as hotspots, we measure the performance of our classifiers using the F1-score. This metric combines precision and recall rate in predicting hotspots. We separately estimate this metric for high and low hotspot segments.



(a) F-score High - Arousal (Abs.) (b) F-score Low - Arousal(Abs.)

(c) F-score High - Arousal(QA) (d) F-score Low - Arousal(QA)

(e) F-score High - Valence(Abs.) (f) F-score Low - Valence(Abs.)

(g) F-score High - Valence(QA) (h) F-score Low - Valence(QA)

Fig. 8. F1-scores achieved by the proposed methods in detecting high and low hotspots for arousal and valence as a function of $pred_{threshold}$. This evaluation is conducted on the development set on the RECOLA dataset.

Frameworks II and III rely on the margin $pred_{threshold}$, which is set on the development set, varying its value between 0 and 2 in steps of 0.025. Figure 8 reports the F1-scores in the development set as a function of $pred_{threshold}$. The figure also shows the results for framework I (baseline), which does not depend on this margin (see solid horizontal lines). The figure illustrates a common trend across the different types of hotspots. For small values of $pred_{threshold}$, most regions are detected as hotspots. For large values of $pred_{threshold}$, most regions are detected as neutral regions. Both of these extreme cases affect the F1-scores, where the best result is achieved somewhere in between. The figure shows that for most of the values for $pred_{threshold}$, the F1-scores for framework II and III are higher than the one for the baseline. This result shows the sensibility of the approach as we vary $pred_{threshold}$. We observe a small difference in the F1-scores between the methods in Framework II and Framework III. The main difference is observed for valence, where framework III reaches higher F1-scores for high-hotspots. We select the best thresholds for each framework to assess the results on the test set. Table 4 lists the values for $pred_{threshold}$ that give the best F1-score for different types of emotional hotspots.

Table 5 shows the performance of the proposed systems in detecting hotspots on the test set. The first part of the table gives the results for arousal, and the second part gives the results for valence. The results consider the cases where the hotspots labels are derived with the QA-based and absolute-based methods. In addition to the F1-scores, the table shows the distribution of the ground truth labels and the predicted

TABLE 5
Results on the RECOLA database. F1-scores of different methods of hotspot detection. The hotspot definition method determines the ground-truth for the evaluations and it is shown in the first column. The table shows the percentage of ground truth and predicted data under hotspot regions.

Definition Method	Detection Method	F1-Score		% of Ground truth data				% of Predicted data			
		L	H	L	N	H	NC	L	N	H	NC
	Arousal										
QA-based	Framework I: Baseline	48.6	30.6	9.8	56.6	9.5	24.1	19.2	31.7	17.4	31.7
	Framework II: one regressor	59.1	51.2	9.8	56.6	9.5	24.1	10.2	50.2	15.5	24.1
	Framework III: ensemble – Abs.	**59.8**	**51.7**	9.8	56.6	9.5	24.1	13.0	49.2	13.7	24.1
	Framework III: ensemble – QA	58.7	51.5	9.8	56.6	9.5	24.1	14.0	42.4	14.9	28.7
Absolute-based	Framework I: Baseline	41.8	30.4	10.1	80.3	9.6	0.0	24.4	52.0	23.6	0.0
	Framework II: one regressor	**49.9**	**47.4**	10.1	80.3	9.6	0.0	13.8	72.0	14.2	0.0
	Framework III: ensemble – Abs.	**49.9**	46.4	10.1	80.3	9.6	0.0	17.7	64.0	18.3	0.0
	Framework III: ensemble – QA	48.9	46.8	10.1	80.3	9.6	0.0	18.9	62.3	12.8	6.0
	Valence										
QA-based	Framework I: Baseline	21.9	24.5	9.9	54.7	10.0	25.4	24.5	33.4	12.9	29.2
	Framework II: one regressor	31.2	43.5	9.9	54.7	10.0	25.4	30.5	29.8	14.4	25.3
	Framework III: ensemble – Abs.	31.9	47.4	9.9	54.7	10.0	25.4	22.9	40.1	11.7	25.3
	Framework III: ensemble – QA	30.9	47.5	9.9	54.7	10.0	25.4	22.0	30.2	12.1	35.7
Absolute-based	Framework I: Baseline	22.4	25.2	10.5	79.3	10.1	0.0	25.8	45.1	29.1	0.0
	Framework II: one regressor	27.6	37.2	10.5	79.3	10.1	0.0	22.9	57.7	19.3	0.0
	Framework III: ensemble – Abs.	**28.7**	42.2	10.5	79.3	10.1	0.0	25.3	62.2	12.5	0.0
	Framework III: ensemble – QA	27.0	41.9	10.5	79.3	10.1	0.0	22.7	53.4	12.3	11.6

labels for low (L), high (H), neutral (N) and non consensus (NC) regions. To facilitate the analysis, we have averaged the performance across conditions showing the average F1-scores in terms of frameworks, approach to define the ground truth, emotional attributes, and type of hotspots. Figure 9 shows the results which are estimated directly from Table 5. These figures have color-coded asterisks which denote the results of statistical tests. An asterisk on top of the bar indicates that the result is significantly higher than the bar with the asterisk's color. We analyze the statistical significance of the results using one-tailed matched pair t-tests, asserting significance at p-value=0.05.

5.2.1 Comparison of Approaches to Define Ground Truth

Table 5 shows an important result. We observe better performance when we use the QA-based method to define the hotspot labels. Figure 9(a) summarizes the results for QA-based and absolute-based labels. On average, the results are 4.7% higher with the QA-based labels across the 16 matched conditions (i.e., four frameworks × two hotspot types × two emotional attributes). This difference is statistical significant (matched pair t-test, one-tailed, p-value<0.01), with improvements ranging from 0.2% to 9.9% (see Table 5). While the performance of the detection method by itself cannot be used to validate the definition of the hotspots, these results show that QA-based labels lead to more reliable prediction of hotspots. We hypothesize that the main reasons for better performance are omitting regions of no consensus and relying on trends across evaluators.

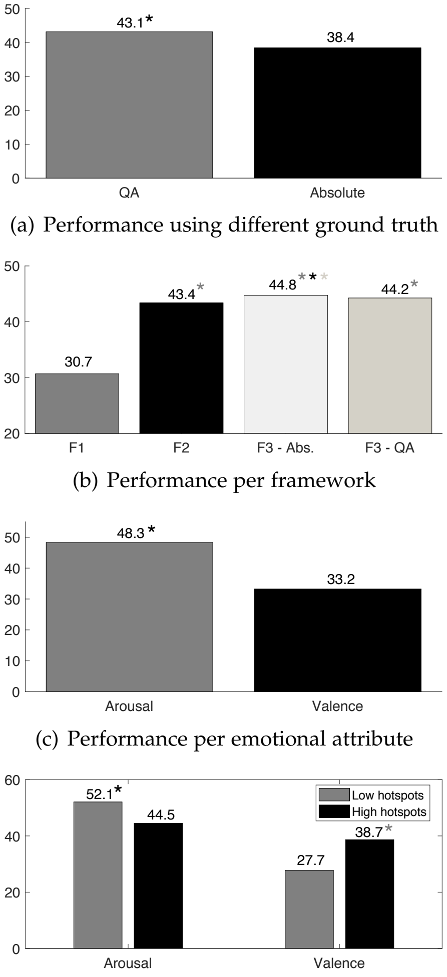
5.2.2 Comparison of Proposed Frameworks

The table shows that Framework II and III are consistently better than Framework I across conditions. This result is clearly observed in Figure 9(b), which provides the average F1-scores in terms of frameworks across the eight matched conditions (two approaches to define ground truth labels × two emotional attributes × two hotspot types). The matched pair t-test indicates that the differences are significant. This result suggests that the approaches relying on regressors are better than the one relying on multi-class classifiers. The concept of hotspot is relative, since its presence depends on the emotional content around a given segment. Framework II and III estimate the emotional traces for the session and then estimate the hotspots, so relative variations are considered. This is done by setting thresholds on the predicted traces which is a more flexible approach for this task. When we compare frameworks II and III, we observe that the average F1-scores are close with slightly better performance for Framework III. Framework III using absolute-based approach to fuse the predicted traces (F3-Abs.) is better than Framework II (matched pair t-test, one-tailed, p-value=0.04). The difference between Framework II and Framework III fusing the predicted traces with the QA-based approach (F3-QA) is not statistically significant. The results suggest that combining the prediction of individual traces to form hotspots (Framework III) is more appealing than learning hotspots by thresholding the prediction of a single regressor. The matched pair t-test shows that fusing the predictions with the absolute method is slightly better than fusing them with the QA-based method (p-value=0.03).

5.2.3 Detection per Emotional Attribute

Figure 9(c) gives the average F1-scores per emotional attribute across the 16 matched conditions (i.e., two approaches to define ground truth labels × four frameworks × two hotspot types). The figure shows that the F1-scores for valence are lower than that the ones for arousal (matched pair t-test, one-tailed, p-value<0.01). This result agrees with previous studies that show the difficulty in predicting valence with acoustic features [54], [55].

We also compared the performance in detecting either high or low hotspots. Figure 9(d) shows the average F1-scores across the eight matched conditions (i.e., two approaches to define ground truth labels × four frameworks). For arousal, the F1-scores for predicting low hotspots is significantly better than the F1-scores for predicting high hotspots. For valence, we observe opposite results, where it



(a) Performance using different ground truth

(b) Performance per framework

(c) Performance per emotional attribute

(d) Performance per type of hotspot

Fig. 9. Summary of results presented in Table 5 on the RECOLA dataset. Statistical significant results are indicated with color-coded asterisks (the bar with the asterisk is significantly higher than the bar with the color indicated by the asterisk).

is more accurate to predict high hotspots than low hotspots. This result suggests that highly positive emotions might be easier to predict or detect than highly negative emotions.

5.2.4 Distribution of the Hotspot Predictions

Table 5 also presents the percentage of data assigned to hotspots, neutral regions and no consensus segments for both the ground truth data and the predicted data. Notice that the ground truth percentage is constant for the different estimation methods, since the thresholds were set on the training set such that 10% of the data were high hotspots and 10% were low hotspot regions (Fig. 4). For Frameworks II and III, the threshold is set to maximize performance on the development set (Fig. 8), so it is not guaranteed that the same distributions are preserved. Achieving percentage of hotspots close to 10% will indicate that our predictions are not biased. Table 5 shows that the percentage of hotspot regions using the QA-based method to define the labels and to fuse the individual regressors (F3-QA) is between 12.1% and 22.7%. The percentage of predicted hotspot regions generally increases when we use the absolute-based method to define ground truth. This result provides another advantage of using trends instead of absolute scores. We also observe that the percentage of the regions predicted as hotspots is higher for valence, which was the emotional attribute with the worse hotspot detection performance. We expect that having better regression performance would improve the distribution of the predicted hotspots.

TABLE 6
Results on the RECOLA database. F1-scores when we remove ambiguous areas without consensus detected by the QA-based approach on the predicted traces. The hotspot definition method determines the ground-truth for the evaluations and it is shown in the first column.

Definition Method	Detection Method	F1-Score L	F1-Score H
	Arousal		
QA-based	Framework I: Baseline	49.7	31.2
	Framework II: one regressor	59.7	52.7
	Framework III: ensemble – Abs.	**60.9**	**53.4**
	Framework III: ensemble – QA	59.8	52.9
Absolute-based	Framework I: Baseline	42.7	31.0
	Framework II: one regressor	50.3	48.5
	Framework III: ensemble – Abs.	50.5	47.3
	Framework III: ensemble – QA	49.6	48.6
	Valence		
QA-based	Framework I: Baseline	22.3	24.8
	Framework II: one regressor	31.5	45.3
	Framework III: ensemble – Abs.	32.7	50.4
	Framework III: ensemble – QA	32.3	50.1
Absolute-based	Framework I: Baseline	22.6	25.7
	Framework II: one regressor	28.3	38.1
	Framework III: ensemble – Abs.	28.8	44.5
	Framework III: ensemble – QA	28	44.1

5.2.5 Performance Ignoring Segments without Consensus

As noted in Section 3.4, one of the advantages of the QA-based method is to avoid regions with no consensus amongst raters. Framework III with the QA-based approach also predicts regions with no consensus across the predicted traces. In previous results, we consider an error when the ground truth has a hotspot and this framework predicts a segment without consensus. This section considers the performance of the proposed methods when we discard segments identified by the Framework III with the QA-based approach as regions with no consensus across the predicted traces. The evaluation in this section can be interpreted as fusing Framework II and Framework III (with absolute-based fusion approach) with the results from the QA-based analysis.

An important observation that validates the results in this section is that most of the bins predicted as no consensus are bins defined as neutral by the ground truth labels. In fact, the average percentage of hotspots in the ground truth labels only decreases 1% when we remove the segments predicted without consensus.

The thresholds for $pred_{threshold}$ are re-estimated using the development set using the same approach as before in Table 4. The best thresholds per framework are used in the test set. Table 6 reports the results. Matched pair t-test shows that knowing regions without consensus improves results for all methods (two approaches to define ground truth labels × four frameworks × two hotspot types × two emotional attributes = 32 matched conditions). Framework III with the QA-based approach improves its performance up

TABLE 7
Results on the SEMAINE database. F1-scores of different methods of hotspot detection. The hotspot definition method determines the ground-truth for the evaluations and it is shown in the first column.

Definition Method	Detection Method	F1-Score	
		L	H
	Arousal		
QA-based	Framework I: Baseline	24.5	22.9
	Framework II: one regressor	38.5	39.8
	Framework III: ensemble – Abs.	**38.6**	**39.6**
	Framework III: ensemble – QA	38.9	38.3
Absolute-based	Framework I: Baseline	17.1	16.2
	Framework II: one regressor	28.1	29.3
	Framework III: ensemble – Abs.	**28.3**	**29.6**
	Framework III: ensemble – QA	27.6	28.6
	Valence		
QA-based	Framework I: Baseline	20.1	21.0
	Framework II: one regressor	25.7	**24.1**
	Framework III: ensemble – Abs.	**25.7**	25.6
	Framework III: ensemble – QA	25.5	25.8
Absolute-based	Framework I: Baseline	16.1	15.8
	Framework II: one regressor	20.5	19.3
	Framework III: ensemble – Abs.	19.7	**20.5**
	Framework III: ensemble – QA	**20.0**	20.2

to 2.6% by removing bins without consensus. The difference between Framework II and Framework III with both fusion approaches is statistical significant, indicating the advantage of using individual predictions for this task.

5.3 Detection on SEMAINE database

We validate the proposed methodology to define and predict emotional hotspots by repeating the evaluation on a different corpus. As mentioned in Section 2.5, we rely on the SEMAINE database [45] for sessions with at least six traces. For the detection of hotspots, we implement the same frameworks used for the experiments on the RECOLA database (Sec. 4), including the selection of the thresholds. While the values are not the same as the ones used for the RECOLA database, they are estimated using the development set following the same approach described Section 5.2. The sessions are divided into 11 speaker independent partitions for the detection task.

Table 7 shows the results. We observe trends similar to the results obtained in the RECOLA database. The QA-based method provides ground truth for hotspots that are better predicted than the ground truth obtained with the absolute method. The results for Framework II and Framework III show similar trends than with the results on the RECOLA database. While the trends are similar, the F1-scores are lower than the ones reported for the RECOLA database (Table 5). The emotional content in the SEMAINE database is broader than the one in the RECOLA database. Furthermore, the traces are annotated by different raters across different sessions, which challenges the training of the individual regressors (e.g., evaluators introduce different bias which creates diversity across regressors).

5.4 Comparison to Related Work

This paper proposes a new formulation in affective computing, so there is no study that can be directly compared with our method. However, there are related problems that some studies have considered. One example is detecting changes of emotions. Huang and Epps [10] reported the *miss detection* (MD) probability, and *false alarm* (FA) probability of recognizing changes of emotions for arousal and valence. Their methods achieved a 13% MD and 11% FA for arousal, and 23% MD, 16% FA for valence.

Several studies have proposed to identify salient regions within a sentence as an intermediate step to recognize the emotion of a sentence [30], [31], [56]. While these methods increase classification performance at the sentence level, they do not report performance metrics on detecting emotional regions. The selected hotspot regions are commonly modeled as hidden variables. Therefore, these results cannot be directly compared with our findings.

6 Conclusions

This study presented an effective approach to define and identify emotionally salient regions, or hotspots, during continuous speech recordings. An important advantage of the proposed approach is that it relies on existing time-continuous annotations, so no further annotations are needed. Instead, the hotspots are defined with a modified version of the QA approach. The approach identifies trends across individual traces, creating hotspot segments when their values are consistently above (high hotspots) or below (low hotspots) their median value. This approach results in consistent labels of hotspots that are more reliable than alternative methods relying on absolute values of the emotional traces. The approach also defines regions without consensus, where trends across emotional traces are not consistent. The paper discussed the advantages of knowing these ambiguous regions, which directly impacts the performance in detecting hotspots.

After defining hotspots, this paper proposed machine-learning frameworks to detect the hotspots. The most promising alternative is to train separate regressors using each of the individual traces available for the sessions. The predicted traces are then fused by averaging their values, or by relying again on the QA approach. The results demonstrated that predicting hotspots is feasible achieving F1-scores as high as 60.9% for arousal and 50.4% for valence on the RECOLA database. The proposed formulation departs from traditional approaches in affective computing aiming to classify emotions or estimate continuous scores at the utterance level. It provides a more practical framework, where speech segments with relevant emotional information during a dialog are automatically detected, without the need to pre-segment the recordings.

The results indicate that improving the performance of the regressors of emotional attributes should also increase the performance of the predicted hotspots. We are collecting a large speech corpus with naturalistic behaviors [57], which should help us to train more robust and accurate regressors. Furthermore, the machine learning formulation can be easily extended to include other modalities such as facial features. As long as the features signal changes from neutral behaviors, we expect that they will be useful to recognize hotspots. Another research direction is to extend this framework when time-continuous emotional traces are not available. Many corpora annotate emotional attributes at the segment level (e.g., one score per sentence regardless of its duration). Adapting the framework for these cases will increase its usability. We can also adapt the framework to model dyadic interactions, capturing dependencies between speakers.

Acknowledgments

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

References

[1] 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.

[2] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, February 2013.

[3] 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.

[4] F. Ringeval, F. Eyben, E. Kroupi, A. Yuce, J.-P. Thiran, T. Ebrahimi, D. Lalanne, and B. Schuller, "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data," *Pattern Recognition Letters*, vol. 66, no. 15, pp. 22–30, November 2015.

[5] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5200–5204.

[6] 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.

[7] H. Meng, N. Bianchi-Berthouze, Y. Deng, J. Cheng, and J. P. Cosmas, "Time-delay neural network for continuous emotional dimension prediction from facial expression sequences," *IEEE Transactions on Cybernetics*, vol. 46, no. 4, pp. 916–929, April 2016.

[8] 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.

[9] 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.

[10] Z. Huang and J. Epps, "Detecting the instant of emotion change from speech using a martingale framework," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5195–5199.

[11] Z. Huang, J. Epps, and E. Ambikairajah, "An investigation of emotion change detection from speech," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1329–1333.

[12] 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.

[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] 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.

[15] ——, "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.

[16] M. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic cca for analysis of affective behavior and fusion of continuous annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1299–1311, July 2014.

[17] 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

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] M. Abdelwahab and C. Busso, "Supervised domain adaptation for emotion recognition from speech," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2015)*, Brisbane, Australia, April 2015, pp. 5058–5062.

[23] M. Wöllmer, A. Metallinou, N. Katsamanis, B. Schuller, and S. Narayanan, "Analyzing the memory of BLSTM neural networks for enhanced emotion classification in dyadic spoken interactions," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 4157–4160.

[24] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *International Journal of Synthetic Emotions (IJSE)*, vol. 1, no. 1, pp. 68–99, January-June 2010.

[25] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, September 2008, pp. 597–600.

[26] A. Metallinou, A. Katsamanis, and S. Narayanan, "Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information," *Image and Vision Computing*, vol. 31, no. 2, pp. 137–152, February 2013.

[27] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Journal of Language Resources and Evaluation*, vol. 50, no. 3, pp. 497–521, September 2016.

[28] 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.

[29] M. Basseville and I. Nikiforov, *Detection of abrupt changes: theory and application*. Prentice Hall, April 1993.

[30] W. Lin and C. Lee, "A thin-slice perception of emotion? an information theoretic-based framework to identify locally emotion-rich behavior segments for global affect recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 5790–5794.

[31] H. K. Vydana, P. Vikash, T. Vamsi, K. P. Kumar, and A. K. Vuppala, "Detection of emotionally significant regions of speech for emotion recognition," in *Annual IEEE India Conference (INDICON 2015)*, New Delhi, India, December 2015, pp. 1–6.

[32] 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.

[33] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009, pp. 1–8.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

[41] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll, "LSTM-modeling of continuous emotions in an audiovisual affect recognition framework," *Image and Vision Computing*, vol. 31, no. 2, pp. 153–163, February 2013.

[42] M. Wöllmer, A. Metallinou, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive multimodal emotion recognition from speech and facial expression using bidirectional LSTM modeling," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 2362–2365.

[43] R. Gupta, K. Audhkhasi, Z. Jacokes, A. Rozga, and S. Narayanan, "Modeling multiple time series annotations based on ground truth inference and distortion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 76–89, January-March 2018.

[44] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "AVEC 2016: Depression, mood, and emotion recognition workshop and challenge," in *International Workshop on Audio/Visual Emotion Challenge*, Amsterdam, The Netherlands, October 2016, pp. 3–10.

[45] 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.

[46] E. Mower, S. Lee, M. Matarić, and S. Narayanan, "Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 961–964.

[47] T. Hasan, M. Abdelwahab, S. Parthasarathy, C. Busso, and Y. Liu, "Automatic composition of broadcast news summaries using rank classifiers trained with acoustic and lexical features," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2016)*, Shanghai, China, March 2016, pp. 6080–6084.

[48] S. Park, G. Mohammadi, R. Artstein, and L. P. Morency, "Crowdsourcing micro-level multimedia annotations: The challenges of evaluation and interface," in *ACM Multimedia 2012 workshop on Crowdsourcing for multimedia (CrowdMM)*, Nara, Japan, October 2012, pp. 29–34.

[49] L. He, D. Jiang, L. Yang, E. Pei, P. Wu, and H. Sahli, "Multimodal affective dimension prediction using deep bidirectional long short-term memory recurrent neural networks," in *International Workshop on Audio/Visual Emotion Challenge (AVEC 2015)*, Brisbane, Australia, October 2015, pp. 73–80.

[50] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Interspeech 2018*, Hyderabad, India, September 2018.

[51] N. Sadoughi and C. Busso, "Expressive speech-driven lip movements with multitask learning," in *IEEE Conference on Automatic Face and Gesture Recognition (FG 2018)*, Xi'an, China, May 2018, pp. 409–415.

[52] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. André, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, April-June 2016.

[53] 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.

[54] 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.

[55] K. Sridhar, S. Parthasarathy, and C. Busso, "Role of regularization in the prediction of valence from speech," in *Interspeech 2018*, Hyderabad, India, September 2018.

[56] J. Lee and I. Tashev, "High-level feature representation using recurrent neural network for speech emotion recognition," in *Interspeech 2015*, Dresden, Germany, September 2015, pp. 1537–1540.

[57] 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.

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.

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. In 2015, his student received the third prize IEEE ITSS Best Dissertation Award (N. Li). He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain), and the Best Paper Award at the AAAC ACII 2017 (with Yannakakis and Cowie). 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 interaction, in-vehicle active safety system, and machine-learning methods for multimodal processing. He was the general chair of ACII 2017. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.