INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Defining Emotionally Salient Regions using Qualitative Agreement Method

Srinivas Parthasarathy and Carlos Busso

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

sxp120931@utdallas.edu, busso@utdallas.edu

Abstract

Conventional emotion classification methods focus on predefined segments such as sentences or speaking turns that are labeled and classified at the segment level. However, the emotional state dynamically fluctuates during human interactions, so not all the segments have the same relevance. We are interested in detecting regions within the interaction where the emotions are particularly salient, which we refer to as *emotional hotspots*. A system with this capability can have real applications in many domains. A key step towards building such a system is to define reliable hotspot labels, which will dictate the performance of machine learning algorithms. Creating ground-truth labels from scratch is both expensive and time consuming. This paper also demonstrates that defining those emotionally salient segments using perceptual evaluation is a hard problem resulting in low inter-evaluator agreement. Instead, we propose to define emotionally salient regions leveraging existing time-continuous emotional labels. The proposed approach relies on the *qualitative agreement* (QA) method, which dynamically captures increasing or decreasing trends across emotional traces provided by multiple evaluators. The proposed method is more reliable than just averaging traces across evaluators, providing the flexibility to define hotspots at various reliability levels without having to recollect new perceptual evaluations.

Index Terms: emotion recognition, emotional hotspot detection, qualitative Agreement.

1. Introduction

Recognition of expressive behaviors using multimodal cues plays an important role in human computer interaction. Most of the current studies have focused on classifying discrete categories such as happiness, sadness, and anger [1, 2, 3] or predicting attribute based descriptors such as arousal (calm versus active) and valence (negative versus positive) [4, 5]. Most of the studies have focused on short, pre-segmented clips (audio,video), which are annotated at the segment level (e.g., IEMOCAP database [6]). However, during human interaction the expressed behaviors are usually neutral, with few segments conveying emotions. There is a growing interest in developing systems that are dynamic in nature, where the emotions are tracked continuously over time detecting salient segments that deviate from neutral behaviors [7]. Some studies have focused on detecting points where the emotional content change during a dialog [8]. These research directions are appealing from an application perspective. We are interested in both defining and detecting emotionally salient regions in longer spontaneous interactions, which we refer to as *emotional hotspots*. A hotspot detection system can play an important role in emotion retrieval tasks [9, 10], providing insights to investigate the triggers and motivations behind expressive behaviors.

One of the major barriers toward detecting affective behaviors is the definition of reliable emotional labels [11, 12]. There are clear differences on how we perceive emotions, which make perceptual evaluations a complex process [13]. Unreliable labels can greatly affect the performance of classifiers. This problem is particularly important for detecting hotspots or changes in emotions, since current emotional annotations are not suitable for this task. The most convenient labels for these tasks correspond to continuous-time evaluations using toolkits such as FEELTRACE (e.g., SEMAINE database [14]), where emotional traces are aggregated across multiple raters. Metallinou et al.[15] showed that inter-evaluator agreement for continuous traces is low and directly aggregating the traces leads to unreliable labels. Creating labels for hotspot detection from scratch is expensive, time consuming, and, as described in this paper, prone to low inter evaluator agreement. There is a need to define labels for emotionally salient segments that are reliable.

The goal of this paper is to define labels describing emotionally salient segments from existing emotional evaluations. We exploit the *qualitative agreement* (QA) method [16] to define labels for emotional hotspots in the SEMAINE database [14]. For a given emotional attribute, the QA approach searches for trends where the evaluators agree. We modify this powerful framework to define segments with high and low levels of arousal and valence. The results show that the emotionally salient segments defined by the proposed QA-based approach are more reliable than defining these segments after aggregating the emotional traces. The framework is easily extended to other databases relying on continuous-time annotations, providing the infrastructure to build reliable emotional hotspot detectors.

2. Database and Emotional Hotspots

2.1. Database

The study uses the SEMAINE database, which consists of dyadic interactions between a user and an operator. In the solid *Sensitive Artificial Listener* (SAL) portion of the database, the operator is a human who plays a particular character (Spike, Poppy, Obadiah and Prudence) to elicit emotions from the user (angry, happy, gloomy and reasonable) [14]. We only consider segments from the users. The conversations are annotated using FEELTRACE [17]. As the evaluators perceive the emotion on the video, they move a joystick over a *graphical user interface* (GUI), where the axes represent the extremes of an emotional attribute. The values are continuously collected forming traces with values between -1 and 1. This study considers arousal (calm versus active), and valence (negative versus positive). The evaluators watch a video and respond by moving the cursor after perceiving its emotional content. This process introduces a delay which has been carefully studied [18, 19, 20]. We compensate the emotional traces with the delays suggested by Mariooryad and Busso [19]. The traces for valence are shifted 4.68s, and the traces for arousal are shifted 3.90s. There are 40 sessions emotionally annotated by six raters. These sessions were collected from 10 users interacting with operators playing the role of four different characters.

This work was funded by a NSF Career Award (IIS-1453781).

http://dx.doi.org/10.21437/Interspeech.2016-429

Table 1: Percentage of ground truth data under each region (low,neutral,high), for different domains.

Dimension	Percentage of Ground-truth data		
	Low	Neutral	High
Arousal	1.69	93.41	3.50
Valence	2.18	95.61	1.63

2.2. Definition and Annotation of Emotional Hotspots

This study defines an emotional hotspot as a segment having either high or low level of a given attribute. For valence, for example, hotspots corresponds to segments with very negative, or very positive emotions. We attempted to define ground-truth labels for emotional hotspots using perceptual evaluations. We asked three evaluators to annotate a subset of 16 sessions (8 for arousal, 8 for valence) recorded by the ten speakers. These sessions were evenly divided between the four characters to cover different emotional content. We independently annotated hotspots in arousal and valence. The task consisted of selecting emotionally salient segments by marking regions that the evaluator perceived having high or low level of a given attribute. The other segments by default were marked as neutral. The evaluation was conducted with OCTAB [21]. Neutral regions dominated the annotation, as expected. This leads to skewed classes. Since each session was annotated by three evaluators, we fused the annotations using simple majority vote (2 out of 3). Segments without agreement are left without labels. Table 1 shows the distribution of annotations, where around 5% of the data was considered as hotspots.

We evaluated the reliability of the evaluations with Fleiss' Kappa (κ). This statistic provides a measure of reliability for tasks with multiple raters and multiple nominal values (low, neutral, and high regions, in our case). The segments where the user spoke are split into fixed windows of 250ms. Then, we compared the annotations from the evaluators estimating Fleiss' Kappa per class, and overall. The overall agreement for arousal is $\kappa_{arousal} = 0.1355$ and for valence is $\kappa_{valence} = 0.1212$, showing the difficulty of this perceptual task. Table 2 shows that there is better consensus amongst raters in evaluating high regions than low regions for both attributes.

3. Methodology

3.1. Qualitative Agreement Framework

It is a common practice to aggregate continuous-time annotations from multiple evaluators by averaging theirs scores. However, previous studies have shown that constructing emotional labels with this approach is challenging and unreliable [15, 22]. The absolute scores provided by raters have low inter evaluator agreements, so adding the absolute scores leads to noisy labels, and as a result, classifiers with poor performance [12].

Cowie and McKeown [16] proposed the *qualitative agreement* (QA) method, which provides an appealing alternative framework to aggregate emotional traces. The intuition behind the approach is that raters tend to agree better on relative trends rather than absolute values [22]. The QA method is a non parametric approach that aims to capture the relative trends that raters agree on. The first step is to create an *individual matrix* (IM) for the trace provided by each evaluator (Fig. 1). A continuous trace is first discretized into *bins*. The average value of the traces within the i^{th} bin is assigned to the bin, denoted b_i. The difference between the values from two bins is estimated and compared to a threshold (Eqs. 1-3).

Table 2: Reliability of the ground truth hotspot labels using Fleiss' Kappa.

Dimension	Region-wise κ			Overall κ
	Low	Neutral	High	
Arousal	0.0651	0.1375	0.1938	0.1355
Valence	0.0778	0.1145	0.2256	0.1212

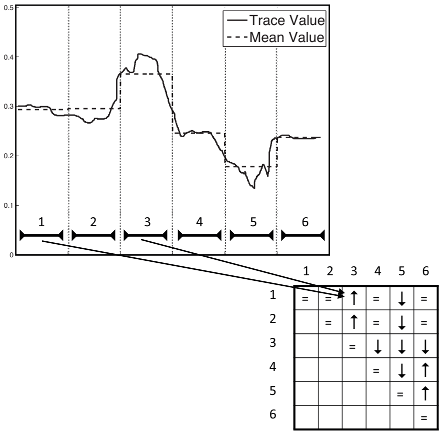


Figure 1: The figure illustrates the process of creating individual matrices under the QA method.

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

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

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

If $i < j$, we assign the "decreasing" symbol when Equation 1 is satisfied, "increasing" symbol when Equation 2 is satisfied and "equal" symbol when Equation 3 is satisfied. Accordingly, we obtain the IM by comparing all possible pairs, as shown in Figure 1. The IM is skew-symmetric, where the figure only shows the elements above the diagonal.

Different individual matrices are then combined using a simple voting scheme to form the *consensus matrix* (CM) (Fig. 2). A second parameter is the agreement tolerance, which defines the confidence level imposed in the process. We add a trend in the CM when X% of the raters agree on a given trend. Otherwise, the cell is left empty. Figure 2 illustrates the formation of the CM matrix from the IMs using a 100% consensus agreement. Entries with no consensus are crossed out.

3.2. Using QA to define Emotional Hotspots

We propose to use the QA approach to create labels for emotional hotspots using existing continuous-time labels. Instead

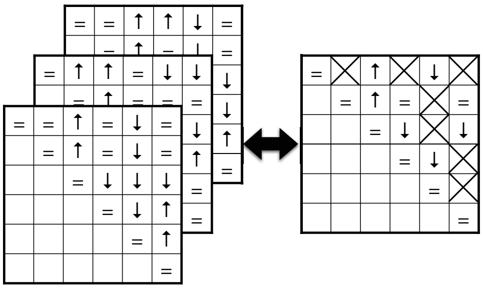


Figure 2: The Consensus matrix, which is formed by combining IMs. We cross out entries without agreement.

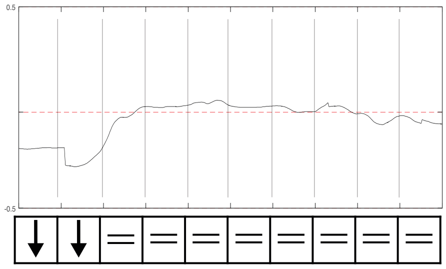


Figure 3: Creating individual vector for each trace by comparing bins to its global median value.

of comparing different bins as done in Section 3.1, we define hotspots by comparing each bin with a reference average value. For a given trace, we estimate the global median across all its bins, denoted b_{median}. The new set of equations are:

$$b_i - b_{median} > t_{threshold} \tag{4}$$

$$b_{median} - b_i > t_{threshold} \tag{5}$$

$$|b_i - b_{median}| < t_{threshold} \tag{6}$$

These comparisons are used to create an *individual vector* (IV), which reflects whether the bins have lower (decreasing – Eq. 4), higher (increasing – Eq. 5), or similar (equal – Eq. 6) value than b_{median}. Figure 3 illustrate this process.

These IVs are then aggregated to form a *consensus vector* (CV), as illustrated in Figure 4. We also use agreement tolerance to define the CV. In the figure, the agreement tolerance is set to 66%. Consecutive increasing trends reflect hotspot with high values. Consecutive decreasing trends reflect hotspot with low values. For those segments, X% of the evaluators agree on the trends. This is an important advantage of these QA-based labels derived from FEELTRACE-type traces. We can control the reliability of the labels by using different parameters.

For defining hotspots using the QA method, we use the following parameters. The length of the bins L is fixed to 3s. To be able to capture the trends, Cowie and McKeowen [16] suggested values for L varying between 1 and 3s. Consecutive bins have a shift of 250 milliseconds to build continuous hotspot labels. This approach produces an overlap of 2.75 seconds but gives us reliable, continuous traces which can later be used for regression tasks. We study the reliability of the QA method at different thresholds, $t_{threshold} = [0.025, 0.05, 0.075, 0.1, 0.125, 0.15, 0.175, 0.2]$. All bins satisfying Eq. 4 are grouped in the *high* category as these bins are greater than the median value. Bins satisfying Eq. 5 are grouped in the *low* category as these bins are lower than the median value. We use an agreement tolerance of 66%. We neglect bins without consensus.

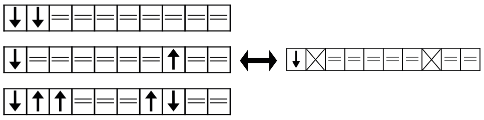


Figure 4: Consensus vector formed by combining individual vectors. matrices. We cross out entries without agreement.

4. Results and Discussion

This section compares our QA-based labels for hotspots with the ground truth evaluations for hotspots(Sec. 2.2). As baseline, we use the averaging method commonly used to define emotional labels. Individual traces from evaluators are averaged across time to form one trace. The resulting trace is discretized into bins and the average score during each bin is compared with the median value, defining hotspots for high and low regions. The bin size and the threshold to define hotspots are exactly the same as the ones used for the proposed QA method. Unlike our method, this baseline approach relies on absolute values across traces instead of their common trends.

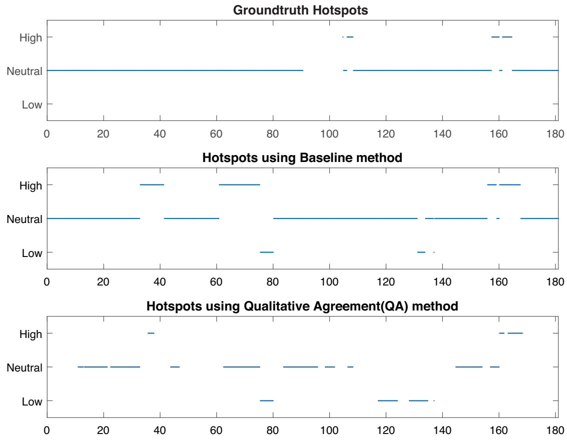


Figure 5: Example of hotspots for arousal for one session. The figures show our ground truth (Sec. 2.2), the baseline method, and the QA-based approach ($t_{threshold}$ set to 0.1).

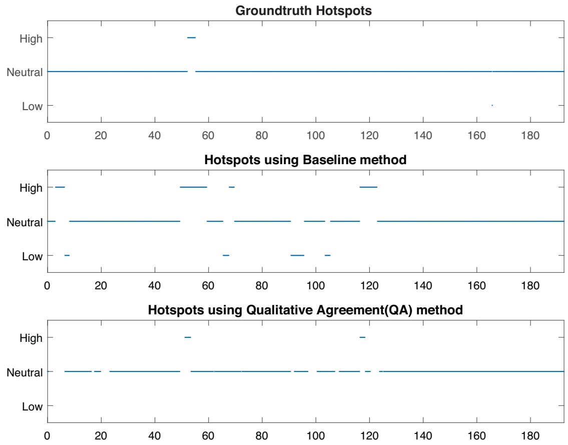


Figure 6: Example of hotspots for valence for one session. The figures show our ground truth (Sec. 2.2), the baseline method, and the QA-based approach ($t_{threshold}$ set to 0.1).

First, we compare the proposed hotspots labels to the ground truth labels. Figures 5 and 6 show examples of the hotspots labels for arousal and valence, respectively, defined by the ground truth, the baseline method, and the QA-based approach for one session. The task of identifying hotspots is similar to the task of *voice activity detection* (VAD) where systems are trained to distinguish between regions of speech and silence. Therefore, we leverage metrics usually used for evaluating VAD systems [23]. In particular, we use the hit rates corresponding to the recall rate of neutral and low-high regions, defined as:

$$H_{h,l} = \frac{N^{pred}_{high,low}}{N^{ref}_{high,low}} \tag{7}$$

$$H_{neu} = \frac{N^{pred}_{neutral}}{N^{ref}_{neutral}} \tag{8}$$

$$H_{ov} = \frac{H_{h,l} + H_{neu}}{2} \tag{9}$$

Table 3: Hit-rate comparison for baseline and QA-based labels. $H_{h,l}$ - Hit-rate for the high and low regions, H_{neu} - Hit-rate for neutral regions, H_{ov} - Overall unweighted hit-rate.

	Hit-rate					
	Baseline			QA		
	$H_{h,l}$	H_{neu}	H_{ov}	$H_{h,l}$	H_{neu}	H_{ov}
$t_{threshold}$	**Arousal**					
0.025	0.65	0.35	0.50	0.89	0.18	0.54
0.050	0.61	0.53	0.57	0.71	0.45	0.58
0.075	0.42	0.65	0.54	0.62	0.64	**0.63**
0.100	0.36	0.75	0.56	0.25	0.82	0.54
0.125	0.34	0.81	0.57	0.16	0.87	0.52
0.150	0.31	0.84	**0.58**	0.14	0.92	0.53
0.175	0.19	0.88	0.54	0.14	0.95	0.55
0.200	0.17	0.92	0.55	0.14	0.97	0.56
	Valence					
0.025	0.81	0.28	0.54	0.75	0.11	0.43
0.050	0.76	0.49	0.63	0.63	0.50	0.56
0.075	0.65	0.63	0.64	0.64	0.74	**0.69**
0.100	0.48	0.76	0.62	0.48	0.84	0.66
0.125	0.47	0.84	**0.66**	0.31	0.91	0.61
0.150	0.38	0.90	0.64	0.30	0.94	0.62
0.175	0.22	0.93	0.57	0.17	0.96	0.56
0.200	0.16	0.95	0.56	0.07	0.97	0.52

where $N^{pred}_{high,low}$ is the number of correctly predicted bins assigned to either high or low regions, $N^{ref}_{high,low}$ is the total number of bins in the ground truth labels assigned to either low or high regions. Identifying high and low regions are as important as identifying neutral regions (Eq.8). For example, false detection of hotspots will affect H_{neu}. A reliable definition of hotspots should simultaneously increase the accuracies for both tasks, which is captured with H_{ov}.

Table 3 presents the hit-rates for different thresholds used in the evaluation (Eqs. 4-6). Figure 7 shows the results for overall hit-rates (H_{ov}). Low thresholds lead to an increased number of hotspots for low or high regions, as the conditions in Equations 4 and 5 are easily met. For these cases, Table 3 shows that $H_{h,l}$ is high, but H_{neu} is low. As we increase the threshold, we set stricter criteria for hotspots, resulting in fewer hotspots regions. The best overall hit-rates for arousal and valence are achieved by the QA method for $t_{threshold} = 0.075$. Figure 7 shows that H_{ov} for valence are better than for arousal, revealing that the task of defining positive or negative salient segments is easier than the task of judging hotspots for arousal.

We have evaluated the proposed QA-based hotspots with ground-truth labels defined in Section 2.2. As we previously discussed, defining ground-truth hotspots is quite difficult for annotators showing low inter-evaluator agreement (Table 2).

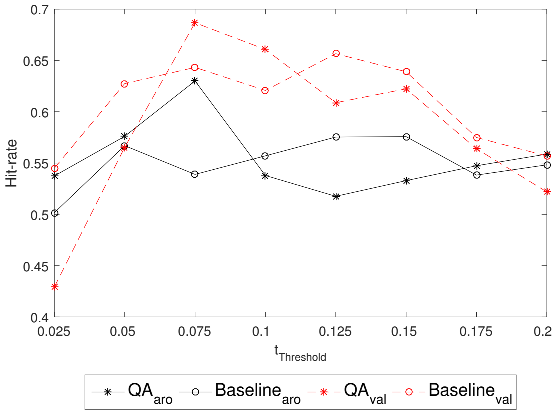
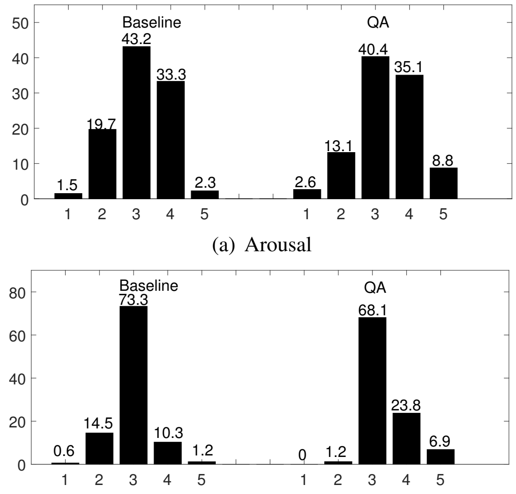


Figure 7: Comparison of overall hit-rates for different thresholds. The figure reports results for arousal and valence for QA-based and baseline approaches.



(a) Arousal

(b) Valence

Figure 8: Perceptual evaluation using a five-point Likert-like scale to assess the appropriateness of the hotspot (1- strongly disagree, 5- strongly agree).

The final evaluation consists of perceptual evaluations, where we ask three raters to determine the appropriateness of the detected hotspots using the best thresholds highlighted in Table 3. We evaluate the selected hotspots using a five-point Likert-like scale (1- strongly disagree, 5 strongly agree). Each evaluator watched the 16 sessions using OCTAB [21], where the detected hotspots where pre-segmented. The evaluators had to rate each hotspot. The perceptual evaluation was conducted twice, one for the QA-based labels and another for the baseline labels. Figures 8(a) and 8(b) show the distribution of the scores assigned to the hotspots for arousal and valence, respectively. For the QA-based method, the percentages of hotspots that the evaluators disagree or strongly disagree are 15.7% for arousal, and 1.2% for valence. These values are lowers than the ones received for the baseline method (21.2% for arousal, 15.1% for valence). For arousal, evaluators agree or strongly agree with the hotspots chosen by the QA-based method 43.9% of the cases, whereas it is only 35.5% for the baseline method. The difference between methods is even higher for valence, where annotators agree or strongly agree in 30.7% of the cases for the QA-based method, but only 11.5% cases for the baseline method. The QA method is better at capturing trends across traces, leading to better hotspots. By setting an agreement tolerance of 66%, we control the reliability of the hotspots, neglecting regions that are ambiguous. This is not possible with the baseline method.

5. Conclusions and Future Work

This paper described a method to define regions of emotionally salient activity (hotspots) using the qualitative agreement method. The flexibility of the QA framework allowed us to define hotspot labels from existing continuous-time emotional traces, controlling their reliability. By extensively studying the performance at various thresholds, we showed that our definitions of hotspots are more reliable than a baseline method consisting of setting thresholds over the averaged traces. The QA framework provides the flexibility to define hotspots labels and changes in emotion. With these labels, we plan to train machine learning algorithms to automatically detect hotspots. Our future work will focus on conducing a comprehensive study on classification and regression tasks which effectively use the proposed QA-based hotspot labels.

6. References

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

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

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

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[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] 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.

[19] ——, "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.

[20] 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.

[21] 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.

[22] 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.

[23] J. Ramirez, J. Górriz, and J. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," in *Speech Recognition and Understanding*, M. Grimm and K. Kroschel, Eds. Vienna, Austria: I-Tech Education and Publishing, June 2007, pp. 1–22.