Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators

Soroosh Mariooryad, *Student Member, IEEE,* Carlos Busso, *Senior Member, IEEE,*

Abstract—An appealing scheme to characterize expressive behaviors is the use of emotional dimensions such as activation (calm versus active) and valence (negative versus positive). These descriptors offer many advantages to describe the wide spectrum of emotions. Due to the continuous nature of fast-changing expressive vocal and gestural behaviors, it is desirable to continuously track these emotional traces, capturing subtle and localized events (e.g., with FEELTRACE). However, time-continuous annotations introduce challenges that affect the reliability of the labels. In particular, an important issue is the evaluators' reaction lag caused by observing, appraising, and responding to the expressive behaviors. An empirical analysis demonstrates that this delay varies from one to six seconds, depending on the annotator, expressive dimension, and actual behaviors. Our experiments show accuracy improvements even with fixed delays (1-3 seconds). This paper proposes to compensate for this reaction lag by finding the time-shift that maximizes the mutual information between the expressive behaviors and the continuous-time annotations. The approach is implemented by making different assumptions about the evaluators' reaction lag. The benefits of compensating for the delay is demonstrated with emotion classification experiments. On average, the classifiers trained with facial and speech features show more than 7% relative improvements over baseline classifiers trained and tested without shifting the time-continuous annotations.

Index Terms—time-continuous emotion annotation, emotion recognition, emotional descriptors, maximum mutual information

1 Introduction

A key aspect in the field of affective computing is defining reliable emotional labels that capture the underlying processes in the production of expressive behaviors [1], [2], [3], [4]. An alternative to discrete emotional categories such as happiness and anger is to describe expressive behaviors with emotional descriptors such as activation/arousal, valence and dominance/power [5], [6], [7], [8], [9]. When these descriptors are assigned at sentence or turn level (i.e., one label per segment), they do not capture the localized emotional information in the stimulus. Human affective states are influenced by the situational context including intentions, environment, and behaviors of other interlocutors. These expressive manifestations tend to evolve during the interaction, changing the nature and intensity of the emotions across time. While defining shorter units can be used to assign specific labels to local segments (e.g., words [10]), it is questionable whether annotators can provide reliable emotional labels after perceiving short stimuli. To address this problem, studies have proposed to continuously track the emotional dimensions to capture the localized, time-variant emotional information [11], [4], [12]. The judgments of the evaluators are continuously recorded as they observe the stimulus and judge its emotional content (i.e., many values per second). Although these time-continuous annotations provide precise representations of the emotional profile across time, they introduce challenges that need to be addressed [3], [13]. This paper addresses one of these challenges: the delay between the emotional annotations and the actual expressive behaviors caused by the evaluators' reaction lag.

Scherer [14] suggested an adapted version of the Brunswik's lens model to illustrate the process of emotion communication. According to this model, the speaker's communicative goals (trait or state) affect the facial expression, gestures, phonation and articulation resulting in specific visual and acoustic patterns referred to as distal cues (i.e., distal from the perspective of the listener). These cues are encoded and transmitted resulting in proximal cues, which are perceived by the listener. The observer fuses the information and makes judgments about the affective state of the speaker. During perceptual evaluations, the observer has to watch the stimuli, perceive the underlying emotional behavior, and make judgments before moving the cursor to annotate the perceived emotion. These processes, happening in real-time, introduce a delay between the expressed distal cues and the collected annotations. Figure 1 illustrates this delay during continuous annotation of videos, which we refer to as the evaluators' *reaction lag* [13]. Understanding this delay is very important, since these emotional descriptors are used as ground truth in the analysis, recognition and synthesis of emotions. For example, emotion recognition systems will be trained and tested with wrong labels, when annotations are

- *S. Mariooryad and C. Busso are with the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, TX 75080 (e-mail: sxm096221@utdallas.edu, busso@utdallas.edu).*

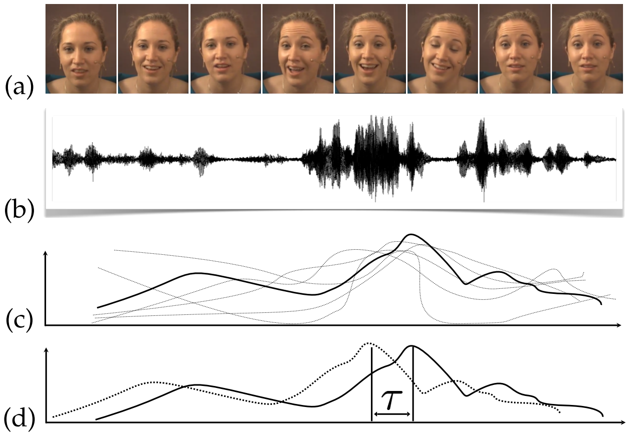


Fig. 1. Evaluators' reaction lag in time-continuous annotations. (a) facial expression, (b) expressive speech, (c) multiple, nonaligned evaluations (dashed lines) and average annotation (solid line), and (d) evaluators' reaction lag between average annotation (solid line) and underlying affective content (dashed line).

not aligned with the actual behaviors.

This work proposes schemes to compensate for the evaluators' reaction lag using an approach based on *maximum mutual information* (MMI) criterion. The study relies on the continuous-time annotations of the SEMAINE database [15], which were evaluated with the FEELTRACE toolkit [11]. First, we analyze the reaction time of the evaluators who annotated selected clips with clear expressive behaviors. We observe average reaction times longer than two seconds. We demonstrate that even fixed delays of 1, 2 and 3 sec. improve classification performance. To systematically address this challenge, we propose to approximate the evaluators' reaction lag by finding the time-shift that maximizes the mutual information between the expressive behaviors (parametrized with facial features) and the annotations. The results on emotion recognition experiments demonstrate the benefits of considering this delay, achieving relative improvement on accuracies up to 7%, compared with the baseline (without compensation). Furthermore, the findings of this study have a direct impact on related fields relying on time-continuous annotations such as emotional music analysis (e.g., *MoodSwings* [16] and *EmuJoy* [17]).

The rest of the paper is organized as follows. Section 2 presents related work addressing the limitations introduced by time-continuous annotations. Section 3 introduces the SEMAINE database and the audiovisual features used in this study. Section 4 presents our empirical analysis and the proposed method to automatically estimate the evaluators' reaction lag. Section 5 validates the benefits of correcting the annotations with the automatic delay estimation through emotion classification experiments. Section 6 concludes the paper, providing discussion and future directions of this study.

2 Related Work

Defining reliable emotional labels is very important in the study of expressive behaviors [1], [2], [3], [4]. Conventional emotion recognition systems are built to classify human emotions into discrete prototypical categories that are universally recognized [18], [19]. These distinct categories account for prototypical or fully-blown emotional states, but they are not rich enough to characterize complex and ambiguous behaviors observed in real human interactions, such as embarrassment, anxiety, boredom, and other subtle emotions [20], [21]. An alternative emotional representation consists in dimensional labels describing affective states [22], [5]. Several studies have investigated the minimum set of dimensions required to effectively characterize expressive behaviors [6], [7], [8], [9]. The commonly accepted emotion dimensions include activation/arousal (calm versus active) and valence (negative versus positive). Other dimensions include expectation (predictable versus unexpected) and power (weak versus strong), which are needed to discriminate between certain emotional categories with similar activation and valence properties [8] (e.g., fear vs. anger).

While these labels are generally assigned at sentence level (i.e., one label per turn) [23], [24], new toolkits such as FEELTRACE, [11], *ANNotating EMOtions* (ANNEMO) [12], and Gtrace [4] provide powerful tools to continuously track these emotional dimensions across time (many values per sec).

Time-continuous annotation presents an appealing approach to capture detailed localized behaviors. The annotation captures differences in intensity within emotional classes (i.e., cold anger versus hot anger) and offers the flexibility to study emotional behaviors at different time-granularity (i.e., phoneme, syllable, word, phrase or sentence levels). Studies have used time-continuous labels to annotate audiovisual emotional recordings [4], [12], [25], TV programs [26], movie clips [27] and music [17], [16]. Although the approach improves the characterization of emotions, it brings open challenges that need to be addressed [3].

The evaluators' reaction lag in observing, appraising and annotating the expressive behaviors introduces a delay between the underlying expressive behaviors and the emotional annotations (Fig. 1(d)). Nicolle et al. [28] proposed a measure based on correlation to estimate the probability distribution for the delay in the annotations. The approach assumed a linear relationship between the annotations and facial features of the subjects. Their study on the SEMAINE database identified delays between three to six seconds depending on the emotional dimension (i.e., activation, valence, expectation and power). The common approach to estimate discrete labels from time-continuous annotations consists in averaging the emotional evaluations across time (i.e., shifting the problem to a multi-class recognition problem) [29], [30], [31], [32]. Since the median duration of the turns in the SEMAINE corpus is 2.76 secs, the delay is significant and the resulting labels do not represent the actual expressive behaviors. We hypothesize that this is one of the reasons of the low emotion recognition performance reported in classification studies on this database [33], [34], [35].

A second problem of time-continuous annotations is the inter-evaluator delays (Fig. 1(c)). The reaction time in annotating the expressive behaviors varies across evaluators [3]. Nicolaou et al. [36] assumed that the inter-evaluator delays are constants. They compensated for this inter-evaluator delay by minimizing the mean square distance between the evaluations. They extended the approach by considering time-variant, localized delays across the evaluations [37]. This approach was based on *dynamic probabilistic canonical correlation with time warping* (DPCTW) and it is referred to as an unsupervised method in the sense that the stimuli is not incorporated in estimating the delay. Notice that all these studies only considered inter-evaluators delays. They did not compensate for the evaluators' reaction lag between the annotations and the expressive behaviors (Fig. 1(d)). Nicolaou et al. [38] extended their approach with two supervised variants of the original scheme that utilize the facial cues in aligning the annotations from multiple evaluators. The approaches provide a framework to compensate for the reaction lag between annotations and the expressive behaviors.

This study extends our previous work on estimating the evaluators' reaction lag using mutual information between time-continuous annotations and facial features [31], [13]. In contrast to previous work, we consider both the evaluators' reaction lag and the inter-evaluator delays. We motivate the proposed work by manually analyzing annotations from key segments expressing clear emotional behaviors. We extend the delay estimation approach by using parametric distributions to derive the mutual information values, which improves the robustness of the estimation. We validate the benefits of compensating for the evaluators' reaction lag with multi-class classification experiments over the segmented chunks.

3 Database, Annotation and Features

This work relies on the *sustained emotionally coloured machine-human interaction using nonverbal expressions* (SEMAINE) database to study the evaluators' reaction lag captured by continuous annotations of this corpus [15]. This audiovisual database contains recordings from users interacting with an operator. Each operator acts as a *sensitive artificial listener* (SAL) agent [39], in which he/she adopts the role of four characters to induce different emotional reactions on the user. The characters have the following personalities: happy, gloomy, angry and pragmatic. The recordings of the corpus used multiple SAL implementations: i) solid SAL: operator is a human, ii) semi-automated SAL: operator is a virtual character controlled by a human, and iii) automated-SAL: operator is a virtual character controlled by a dialog management system. Only 94 sessions from the solid SAL portion of the corpus are currently available. In the solid SAL recordings, the user and operator sit in separate rooms, communicating through teleprompter screens. The voices and videos are simultaneously recorded.

TABLE 1
Set of *action units* (AUs) extracted using CERT [40].

AU	description	AU	description
AU 1	Inner Brow Raise	AU 15	Lip Corner Depressor
AU 2	Outer Brow Raise	AU 17	Chin Raise
AU 4	Brow Lower	AU 18	Lip Pucker
AU 5	Eye Widen	AU 20	Lip stretch
AU 6	Cheek Raise	AU 23	Lip Tightener
AU 7	Lids Tight	AU 24	Lip Presser
AU 9	Nose Wrinkle	AU 25	Lips Part
AU 10	Lip Raise	AU 26	Jaw Drop
AU 12	Lip Corner Pull	AU 28	Lips Suck
AU 14	Dimpler	AU 45	Blink/Eye Closure

In 52 out of the 94 sessions, the users' videos are annotated by 2 to 8 evaluators in terms of activation (i.e., calm versus active), and valence (i.e., negative versus positive), expectation (predictable versus unexpected) and power (weak versus strong). They also evaluated continuous descriptors capturing the intensity of fear, anger, happiness, amusement, certainty and others. The annotations are collected using the FEELTRACE toolkit [11], in which the evaluators continuously track the emotional behaviors by moving the mouse cursor over a space defined by the emotional dimension (one label at a time). The values are mapped into the interval [-1, +1] for each dimension.

3.1 Facial Features

To represent facial expressions, we use the *facial action coding system* (FACS) [41], which defines a set of *action units* (AUs) corresponding to contraction or relaxation of one or multiple facial muscles. We use the *computer expression recognition toolbox* (CERT) [40] to estimate the AUs. This toolkit uses appearance-based features (i.e., Gabor filter banks) to extract the AUs frame-by-frame. The algorithm provides robust performance against changes in illumination, and reasonable head rotations. We extract 23 features consisting of 20 AUs, listed in Table 1, and three head rotation angles (i.e., pitch, yaw, roll). In 8 out of the 52 emotionally annotated sessions, the users' face was not detected by the CERT face detector module (sessions 82, 88-91, 95-97). Therefore, our study only includes the remaining 44 sessions, recorded from nine unique participants.

TABLE 2
The set of frame-level acoustic features used in this study. This set is referred to as *low level descriptors* (LLDs) in the Interspeech 2011 speaker state challenge [42].

Spectral related features
RASTA-style filt. auditory spectrum, bands 1-26 (0-8kHz)
MFCC 1-12
Spectral energy 25-650Hz, 1k-4kHZ
Spectral roll off point 0.25, 0.50, 0.75, 0.90
Spectral flux, entropy, variance, skewnewss, kurtosis
Zero-crossing rate
Prosodic features
L1 norm of auditory spectrum components (loudness)
L1 norm of RASTA-style filtered auditory spectrum
RMS energy
F0
Probability of voicing
Voice Quality
Jitter (local, delta)
Shimmer

TABLE 3
The set of sentence-level functionals extracted from the LLDs (see Table 2).

33 base functionals
Quartiles 1-3
3 inter-quartile ranges
1% percentile (≈min), 99% percentile (≈max)
Percentile range 1%-99%
Arithmetic mean, standard deviation
Skewness, kurtosis
Mean of peak distances
Standard deviation of peak distances
Mean value of peaks
Mean value of peaks-arithmetic mean
Linear regression slope and quadratic error
Quadratic regression a and b and quadratic error
Contour centroid
Duration signal is below 25% range
Duration signal is above 90% range
Duration signal is rising/falling
Gain of linear prediction (LP)
LP coefficients 1-5
6 F0 functionals
Percentage of non-zero frames
Mean, max, min, standard deviation of segments length
Input duration in seconds

3.2 Acoustic Features

The acoustic features are extracted at segment level using the openSMILE toolkit [43]. We use the feature set introduced as baseline for the Interspeech 2011 speaker state challenge [42]. First, a set of frame-level *low level descriptors* (LLDs), listed in Table 2, are extracted. For each LLD, we estimate a set of global statistics referred to as *high level descriptors* (HLDs), given in Table 3. This procedure yields a 4368D feature vector for each segment. The feature set includes prosodic (e.g., energy and fundamental frequency), spectral (e.g., *Mel-frequency cepstral coefficients* (MFCCs), RASTA) and voice quality (e.g., jitter, shimmer) features. The details about these features are given in Schuller et al. [42].

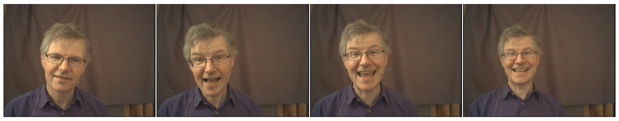


Fig. 2. Four frames displaying one of the expressive events used for the analysis (Session 47 at 3’54”).

4 Evaluators’ Reaction Lag

This section provides empirical evidences of the evaluators’ reaction lag by manually measuring the delay between the annotators’ reactions and clear emotional events in the database (Sec. 4.1). We describe the proposed data-driven approach to estimate the evaluators’ reaction lag (Sec. 4.2), and the assumptions made on the analysis (Sec. 4.3).

4.1 Empirical Study of Evaluators’ Reaction Lag

To motivate our analysis on the evaluators’ reaction lag, we carefully identified five different clear expressive events on the SEMAINE database from three unique users (see Table 4). For consistency, we include only sessions annotated by the first six evaluators (R1-R6). Figure 2 illustrates frames corresponding to one of these events. We manually identified the exact time at which the evaluators responded to the expressive behaviors, by observing the annotations. Using these values, we estimate the reaction lag for these events per evaluator. We separately conduct the analysis for activation and valence dimensions. Table 4 reports the empirical results. We left blank spaces in the table for cases in which the evaluators did not respond to the expressive events.

Table 4 shows that the average delay across all samples and evaluators is 3.18 secs for activation and 3.45 secs for valence. Most of the delays are between one and six secs, which agrees with the delays reported by Nicolle et al. [28]. The average standard deviation of the delays across evaluators (1.30 for activation, 1.58 for valence) is slightly lower than the average standard deviation across expressive events (1.73 for activation, 1.93 for valence). The results reveal a clear time-variant reaction lag that depends on the evaluator, emotional dimension and the actual expressive behavior. Hence, the correction method of the annotations should consider inter-evaluator differences, and variations due to the nature of the stimuli. Given the complexity of this task, we explore different simplifications to model the evaluators’ reaction lag.

4.2 Automatic Estimation of the Reaction Lag

The analysis in this section considers the entire data (44 sessions of the 9 unique users evaluated by 2 to 8 evaluators). We propose a data-driven approach

TABLE 4
Empirical analysis of the evaluators' reaction lag. Five clear expressive behaviors in the SEMAINE database are selected and the delay [sec] per evaluator is manually estimated for activation and valence.

affective event			activation								valence							
			evaluator ID						statistics		evaluator ID						statistics	
Session	User	Time	R1	R2	R3	R4	R5	R6	mean	std	R1	R2	R3	R4	R5	R6	mean	std
21	3	0′20″	2.72	2.80	6.72	6.36	3.84	4.00	4.41	1.74	3.28	2.88	-	6.84	3.44	4.28	4.14	1.59
35	7	1′08″	2.88	1.52	0.96	2.20	2.76	3.08	2.23	0.84	1.96	1.08	1.48	-	2.72	3.52	2.15	0.98
36	7	1′51″	1.28	1.24	2.00	1.84	-	1.68	1.61	0.34	1.76	1.48	0.84	3.56	2.88	2.00	2.09	0.98
37	7	2′01″	2.40	1.60	2.10	6.00	3.00	6.60	3.62	2.14	4.40	2.10	2.50	4.90	3.00	4.30	3.53	1.15
47	2	3′54″	4.60	4.72	1.92	-	4.96	-	4.05	1.43	1.40	5.88	6.92	5.92	9.92	1.92	5.33	3.20
mean			2.78	2.38	2.74	4.10	3.64	3.84			2.56	2.68	2.94	5.31	4.39	3.20		
std			1.20	1.44	2.27	2.41	0.99	2.07			1.25	1.91	2.74	1.41	3.10	1.18		

based on the *maximum mutual information* (MMI) criterion to systematically estimate the evaluators' reaction lag. Mutual information measures the dependency between two random variables X and Y. Equation 1 gives the mutual information for multivariate continuous random variables with *probability density functions* (pdfs) $f_X(x)$ and $f_Y(y)$, and joint pdf $f_{XY}(x,y)$. The approach consists in measuring the mutual information between the underlying emotional behaviors (EMO) and a τ-sec shifted version of the time-continuous emotion annotations (ANN^τ). The optimal delay $\hat{\tau}$ is defined as the time shift that maximizes this mutual information $I[EMO; ANN^\tau]$ (Eq. 2).

$$\begin{aligned} I(X;Y) = & \int_y \int_x f_{XY}(x,y) \log\left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)}\right) dx dy \quad (1) \\ \hat{\tau} = & \arg_\tau \max I[EMO; ANN^\tau] \quad (2) \end{aligned}$$

In our previous work [13], we proposed to estimate the mutual information using non-parametric *probability mass functions* (PMFs) obtained from k-means codebooks. A requirement in this approach was to discretize the continuous features and annotations. In contrast, this study directly estimates Equation 1 using a parametric model for $f_{EMO}(x)$, $f_{ANN^\tau}(y)$, and $f_{EMO,ANN^\tau}(x,y)$. In particular, the model assumes that the pdfs and joint pdf of these continuous random variables follow a multivariate Gaussian distribution. Under this assumption, the mutual information between joint Gaussian random variables X and Y takes the closed-form:

$$I(X;Y) = \frac{1}{2}\log\left(\frac{det(\Sigma_{XX})det(\Sigma_{YY})}{det(\Sigma)}\right) \quad (3)$$

$$\Sigma = \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \quad (4)$$

where Σ_{XX}, and Σ_{YY} are the covariance matrices of X, and Y, respectively, and Σ is their joint covariance matrix (Eq. 4). While this parametric approach yields similar results as the ones reported in our previous work [13], it simplifies the estimation of the mutual information by eliminating the discretization process. Furthermore, it gives smoother mutual information profiles with respect to the time shifts, providing more robust estimations of the reaction lag.

Although the underlying emotional content, EMO, is not explicitly available, it can be approximated by measuring the deviation from neutral behaviors of the audiovisual features. We estimate this continuous emotional profile with facial features (EMO^F). The emotional profile is captured by the values of the AUs, which give the deviation of facial muscles from neutral facial poses (i.e., a neutral face has all AUs set to zero). Hence, it yields a perfect representation to characterize emotional behaviors. The AUs describe the facial appearance conveying the subjects' emotional state at the frame level. Therefore, we expect to obtain reliable delay estimations by capturing these deviations from neutral facial poses. During our preliminary analysis, we also tried to estimate this continuous emotional profile with acoustic features (notice that acoustic features are only available when the users is speaking). The profile was extracted with our recently developed frame-by-frame shape-based representation of energy and F0 contours using *functional principal component analysis* (FPCA) [29], [44]. This technique provides discriminative features to detect emotional speech. Unfortunately, the mutual information values in Equation 2 were significantly lower than the ones achieved with facial features, which indicates that the speech-based emotional profiles do not accurately represent the underlying emotional content of the frames. Furthermore, the curves describing the mutual information as function of the shift exhibit rather noisy behaviors, so the estimation of the speech-based delays are not consistent. Therefore, the study only includes delays estimated with facial features (incorporating acoustic cues to estimate the delay remains an open problem).

4.3 Assumptions Made to Model the Reaction Lag

As discussed in Section 4.1, the reaction lag depends on the evaluator, emotional dimension, and the actual expressive behavior. Are all these dependencies important? This study estimates the reaction lag using three different sets of assumptions: (a) τ is constant across evaluators and sessions - it only depends on the emotional dimension; (b) τ is constant, but it only depends on the evaluators, and the emotional dimension; and (c) τ is constant across all sessions,

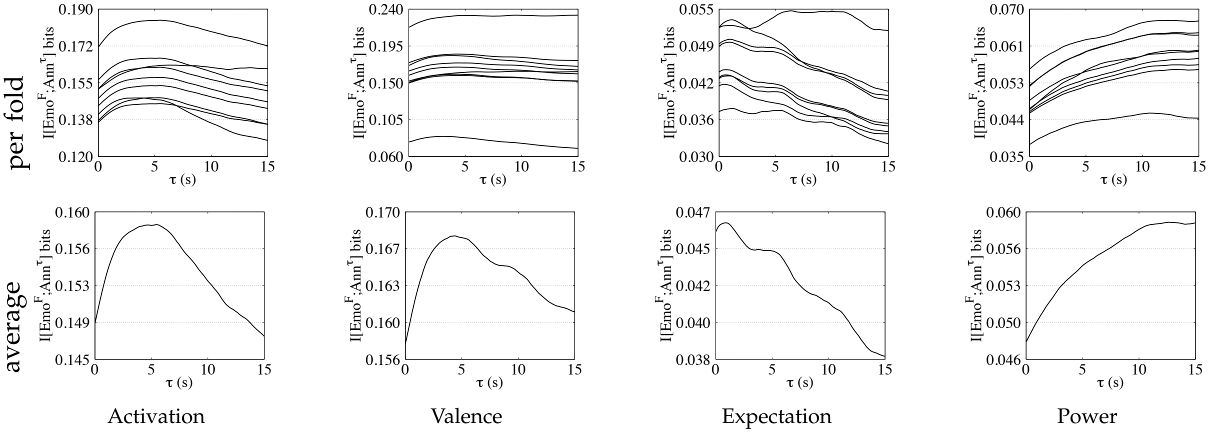


Fig. 3. Analysis of the evaluators' reaction lag. The figures show the mutual information between facial features and the τ-sec-delayed emotion annotations for activation, valence, expectation, and power. The first row gives the mutual information per cross-validation fold and the second row gives the average across folds.

but it is estimated after compensating for session-dependent, inter-evaluators delays (pre-alignment) - it depends on the evaluators, sessions and emotional dimensions.

4.3.1 Constant Evaluator-Independent Reaction Lag Across all Sessions

In this setting, we assume that the reaction lag does not vary significantly across the evaluators (i.e., we neglect the inter-evaluator delays). For each session, we take the average across the time-continuous evaluations to aggregate their annotations yielding ANN. The estimated delays are used in the classification experiment presented in Section 5. The training and testing partitions of the classification experiments are defined with a speaker-independent, cross-validation approach. To avoid using information from the testing set, we estimate the evaluators' reaction lag for each fold considering only the training set. Then, the emotional annotations of both training and testing data are corrected by the estimated delay $\hat{\tau}$. Therefore, the value for $\hat{\tau}$ varies across folds.

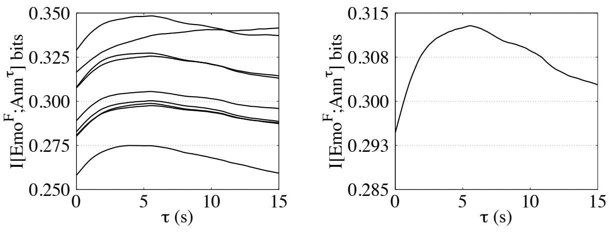
We estimate the mutual information values, $I[EMO; ANN^{\tau}]$, as a function of delay (τ), by increasing the delay from 0 to 15 secs using 40 ms steps. Figure 3 depicts the results for activation, valence, expectation and power (emotional dimensions in the SEMAINE database). The first row gives the mutual informations for each fold, and the second row gives the average profile across folds. The figures clearly demonstrate the effect of the reaction lag. The mutual information profiles show an initial rising pattern, followed by a falling pattern. This trend is also observed in the probability distributions of the delay estimated by Nicolle et al. [28] and our previous work based on non-parametric distributions [13]. The results indicate that the reaction lag for expectation is the shortest one among the emotional dimensions ($\sim$1 sec). The longest delay is observed for power ($\sim$10 sec). However, the mutual information values in these two dimensions are significantly lower than the ones for activation and valence. The results suggest that the relation between facial features and these annotations is weak. In fact, studies using facial features have reported lower classification results in the SEMAINE database for these dimensions [45]. Given the lower reliability of expectation and power, this study only focuses on activation and valence, which are the most common dimensions used to describe emotions (e.g., the Circumflex of Affect [46]). Figure 4 shows the mutual information profiles when we jointly consider the activation-valence space (e.g., $EMO = [Act, Val]$). The figure shows similar patterns.

TABLE 5
Mean and standard deviation of the estimated evaluators' reaction lags across folds (see Fig. 3), and the estimated evaluators' reaction lags for the entire data.

Dimension	Across Folds		All Data
	mean	std	
Activation	4.57	0.81	5.44
Valence	4.25	0.82	4.08
[Activation, Valence]	5.70	1.63	5.60

We select $\hat{\tau}$ as the time shift that maximizes the mutual information profiles (Eq. 2). To favor shorter delays, we set $\hat{\tau}$ as the smallest local optimum (e.g., it may not be the global optimum). With this definition, we aim to identify the start of the plateaus in the mutual information profiles. Notice that $\hat{\tau}$, which is assumed constant, is set per emotional dimension, and per cross-validation fold.

Table 5 shows the mean and standard deviation of the extracted delays across folds, and also the extracted delays for the entire data (combining training and testing partitions), which are similar to the ones manually identified by our empirical analysis (Sec.



Act-Val: Per fold

Act-Val: Average

Fig. 4. Analysis of the evaluators' reaction lag when we jointly consider the activation-valence space.

TABLE 6
Estimated evaluators' reaction lags for each of the eight evaluators (R1-R8) in the SEMAINE corpus.

Dimension	Evaluator ID							
	R1	R2	R3	R4	R5	R6	R7	R8
Act	2.00	2.24	2.64	7.52	5.52	6.08	3.04	3.92
Val	2.08	3.84	3.12	10.88	3.44	6.00	1.92	1.44
Act-Val	2.72	3.52	3.28	11.04	5.52	5.92	1.92	1.76
Mean	2.27	3.20	3.01	9.81	4.83	6.00	2.29	2.37

4.1). Since the median duration of the sentences in the SEMAINE database is 2.76 secs, compensating for this delay is important to achieve reliable labels. Section 5 reports the classification results using this compensation scheme (denoted "Optimal ($\hat{\tau}$)").

4.3.2 Constant Evaluator-Dependent Reaction Lag across all Sessions

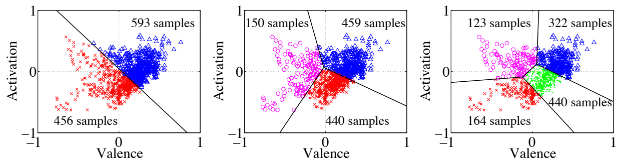
An alternative implementation of the proposed approach consists in estimating the reaction lag per evaluator, and per emotional dimension. Assuming that the reaction lag is evaluator-dependent is motivated by the inter-evaluator differences observed in the reaction lag presented in Section 4.1, and the results reported in previous studies [36], [37]. In this setting, each individual annotation is separately corrected by a constant evaluator-dependent $\hat{\tau}$ before aggregating the annotations.

Table 6 reports the estimated reaction lags for the eight evaluators in the SEMAINE corpus, using the proposed delay estimation technique. Section 5 reports classification results by correcting the annotations with this evaluator-dependent delay estimation approach (denoted "EvalDep"). Since each evaluator annotated sessions from both testing and training partitions, the annotations from the testing set are probably used in the estimation of $\hat{\tau}$. The estimated reaction lags across evaluators presents larger dynamic range (2-11 secs) than the ones with the evaluator-independent assumption (4-6 secs).

The findings observed with the data-driven approach are consistent with our empirical analysis in Section 4.1. Notice that the empirical analysis only includes events annotated by the first six evaluators (i.e., R1-R6). In both experiments, the evaluators R1, R2 and R3 have lower reaction time compared to R4, R5 and R6. Also, both analyses show that R4 has the longest reaction time among the evaluators. The proposed maximum mutual information criterion unveils the reaction lag for each evaluator. Also it captures the annotators' reliability through the maximum values obtained. This is a powerful approach, since it does not require the ground truth of the annotations to measure the annotators' reliability.

TABLE 7
Mean and standard deviation of the estimated evaluators' reaction lags across the nine folds for activation, valence and joint activation-valence after pre-aligning the annotations.

Statistics	Activation	Valence	Activation-Valence
Mean	3.90	4.68	5.60
Std	0.75	0.89	1.64



(a) $K = 2$ (b) $K = 3$ (c) $K = 4$

Fig. 5. The clusters obtained with the K-means algorithm on the activation-valence space without introducing any delay on the annotations ($\tau = 0$).

4.3.3 Constant Evaluator-Dependent Reaction Lag per Session (Pre-Aligning)

Instead of estimating the delay per evaluator (Sec. 4.3.2), we can pre-align the annotations of multiple evaluators in each session before estimating the common reaction lag. This approach compensates for both the inter-evaluator phase, and the evaluators' reaction lag. This approach assumes that the phase between annotations for each pair of evaluators is fixed and less than one second. For each session, one evaluator is randomly chosen as a reference and the other annotations are aligned with his/her annotations by maximizing the cross-correlation between them in the interval of [-1, 1] seconds. After aggregating the pre-aligned annotations, we estimate the optimal reaction lag using the approach described in Section 4.3.1. Table 7 gives the mean and standard deviation of the estimated delays after the pre-alignment step. Section 5 reports the classification results obtained using this compensation technique (denoted "Pre-Align").

5 Validation with Classification Experiments

This section presents emotion classification experiments with the SEMAINE database to study the implications of correcting the time-continuous annotations

TABLE 8
Facial and speech emotion recognition results on activation, valence and joint activation-valence space. K defines the number of emotional classes. The table reports results for: $\tau = 0, 1, 2, 3$ secs; the optimal delay $\hat{\tau}$ (“Optimal”, Sec. 4.3.1); the evaluator-dependent delay (“EvalDep”, Sec. 4.3.2); and the delay estimation with pre-alignment (“Pre-Alig”, Sec. 4.3.3). The results are presented in terms of *Accuracy* (A) and *F-score* (F).

Dimension	lag	Face						Speech					
		K = 2		K = 3		K = 4		K = 2		K = 3		K = 4	
		A	F	A	F	A	F	A	F	A	F	A	F
Activation	0 (baseline)	60.44	60.78	41.66	39.25	32.41	32.43	55.37	53.90	43.29	43.81	29.53	29.10
	1	60.27	60.74	46.45	44.08	32.53	32.90	56.62	54.75	42.43	43.08	28.38	28.19
	2	59.86	60.53	45.77	43.07	35.47	35.95	57.35	55.92	43.06	43.75	29.67	29.23
	3	59.10	59.59	45.40	43.11	34.05	35.30	57.47	56.08	41.57	42.28	31.93	31.53
	Optimal ($\hat{\tau}$)	61.56	61.92	47.63	45.45	39.53	39.82	59.68	58.30	40.60	41.19	31.99	31.28
	EvalDep	59.34	60.10	47.21	44.37	38.32	39.10	58.13	56.87	43.36	44.18	31.84	31.44
	Pre-Align	60.31	60.45	47.54	44.99	39.78	40.55	59.64	57.93	39.46	40.11	34.00	33.48
	Nicolle et al. [28]	60.41	58.62	45.38	45.83	35.27	35.23	59.25	58.67	41.02	45.35	31.01	29.13
Valence	0 (baseline)	67.49	67.80	53.00	51.01	34.13	34.61	66.44	66.48	41.68	40.64	40.94	38.32
	1	68.52	68.76	52.78	50.53	37.72	37.68	66.76	66.81	41.82	40.99	40.95	37.94
	2	72.40	72.66	54.42	52.23	37.71	37.95	66.71	66.79	41.22	40.62	41.22	38.71
	3	70.16	70.76	55.38	53.29	39.24	39.91	68.13	68.18	41.36	40.39	45.22	43.06
	Optimal ($\hat{\tau}$)	70.78	70.98	54.59	51.95	39.19	39.73	66.27	66.36	42.96	42.56	43.93	41.04
	EvalDep	71.36	71.71	54.06	52.89	38.12	38.51	67.75	67.89	46.54	45.71	44.44	41.36
	Pre-Align	69.41	69.90	53.27	51.45	39.60	40.40	67.76	67.76	40.88	40.65	44.29	41.97
	Nicolle et al. [28]	70.45	70.77	54.40	51.48	39.04	36.84	67.65	68.09	43.77	44.65	43.57	43.16
[Act., Val.]	0 (baseline)	55.77	56.04	50.71	51.12	37.46	36.23	48.05	48.31	42.75	45.52	36.85	35.91
	1	59.50	60.14	53.26	53.56	41.17	39.46	50.14	50.39	45.14	47.96	40.47	38.92
	2	62.88	63.35	54.23	54.27	43.63	43.26	52.22	52.45	45.66	48.17	41.08	39.33
	3	63.11	63.76	54.99	55.23	42.27	41.36	54.09	54.43	46.80	49.52	41.09	39.17
	Optimal ($\hat{\tau}$)	62.99	63.18	53.43	54.19	41.39	39.01	51.89	52.32	43.35	45.53	42.93	41.24
	EvalDep	61.00	61.63	55.52	56.10	42.52	40.88	52.57	52.91	46.75	49.26	42.62	40.70
	Pre-Align	62.65	62.91	53.78	54.45	42.03	39.94	52.91	53.26	45.83	48.13	43.66	41.67

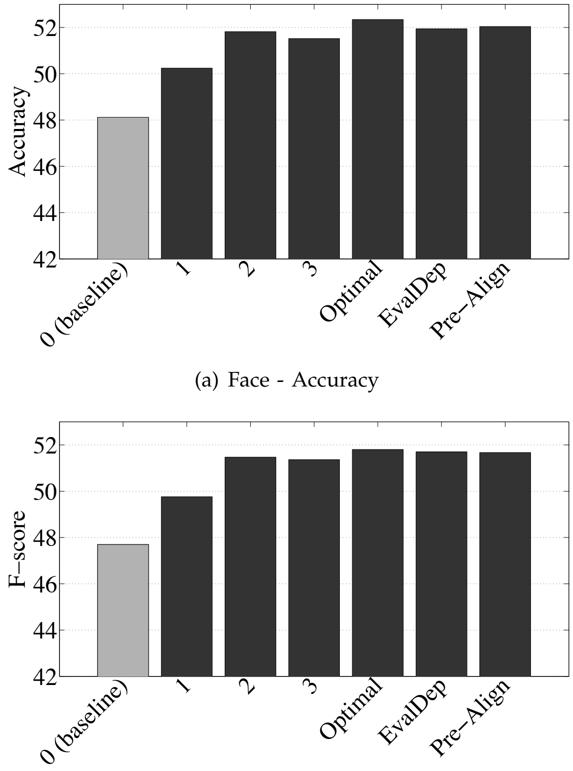
with the proposed methods. The evaluation considers the 44 sessions recorded from nine unique users. We consider the users' speaking segments that are at least 300 ms long (1049 segments). We use a data-driven approach to define the discrete emotional labels from the time-continuous annotations. This is a common approach to transform the problem into a multi-class classification task [47], [48], [49]. First, we estimate the average across the annotations over the duration of each segment (i.e., one value per segment). Notice that shifting the annotations directly affects the emotional value assigned to each segment. Then, we employ K-means algorithm over the average emotional values to define the desired discrete classes. Finally, the codebooks represent the discrete categories for the classification experiments. We separately implement this approach for activation, valence, and joint activation-valence space, using K = 2, 3 and 4. Figure 5 illustrates the clusters for the activation-valence space for different values of K, when no correction is made on the labels (i.e., the baseline setting).

The goal of this evaluation is to validate the benefits of compensating for the evaluators' reaction lag using the three implementations described in Section 4.3. The “Optimal ($\hat{\tau}$)” implementation (Sec. 4.3.1) only models the evaluators' reaction lag, which is assumed to be constant across evaluators and sessions. The second (“EvalDep”, Sec. 4.3.2) and third (“Pre-Align”, Sec. 4.3.3) implementations model the dependency of the annotators on the reaction lag using evaluator-dependent delays, and pre-alignment of the time-continuous labels, respectively. The classification results under these conditions are compared with the ones achieved with the baseline setting that does not compensate for the reaction lag (i.e., τ = 0; the common approach). In addition, we implement classification experiments in which the annotations are shifted with fixed delays. We consider τ = 1, 2 and 3 secs. We also consider the optimum delay proposed in Nicolle et al. [28] using correlation-based framework. The approximate delay values are (3.08 sec for valence, and 3.95 for activation.

Since the average emotional value assigned to each segment varies when the annotations are shifted, the emotional classes defined by the K-means algorithm also change. Therefore, we reestimate the clusters for each condition. Notice that initializing the K-means algorithm with randomly chosen seeds may result in non-homogeneous cluster sets for different time shifts. Therefore, we use the centroids obtained for the baseline setting as the initial seed to run K-means algorithm in the remaining conditions.

The classification evaluations are built with the *sequential minimal optimization* (SMO) implementation of the linear kernel *support vector machine* (SVM) provided by the WEKA toolkit [50]. The complexity parameter of the SVMs is set to c = 0.1 for all the settings. We implement the classifiers with a nine folds cross-validation approach, generating speaker-independent partitions for training and testing sets (i.e., data from one subject is included in either the training or testing sets, but not in both). Given the unbalanced classes obtained by K-means algorithm (see Fig. 5), we evaluate the classifiers with *accuracy* (A) and macro-average *F-score* (F). Macro-average F-score is given in Equation 5, where P and R are the average precision and average recall across the classes, respectively (i.e., we separately estimate the precision and recall values for each class, and then we take the average of these values across classes). We use the large sample proportion hypothesis test [51] to assess whether the improvements achieved by correcting the annotations are statistically significant. The proportion test assesses whether the ratios of samples that are correctly classified by two classification methods are significantly different. The number of samples in this test for our experiment is n=1049 (one-tailed z-test). Even though the reaction lags are only estimated with facial features, we evaluate emotion classification using facial (Sec. 5.1) and acoustic (Sec. 5.2) features. Table 8 reports the results.

$$F = \frac{2PR}{P+R} \tag{5}$$



(a) Face - Accuracy

(b) Face - F-score

Fig. 6. Emotion recognition performance with facial features by compensating for the evaluators' reaction lag across all conditions (K = 2, 3 and 4, for activation, valence, and joint activation-valence space).

5.1 Emotion Recognition from Face

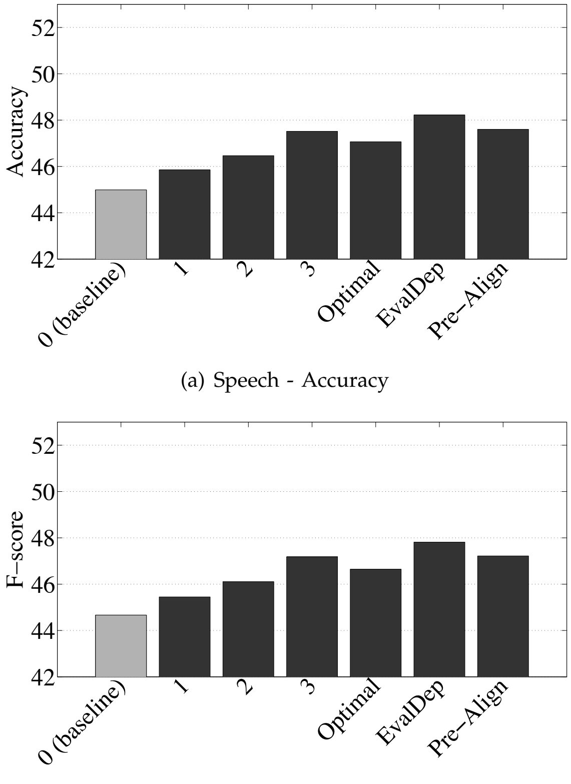
We build the classifiers with features derived from the AUs and head rotation at the sentence level. For each segment, we estimate a set of six statistics to obtain a 138D feature vector (i.e., [20 AUs + 3 head rotation] × 6 statistics). The selected statistics, which characterize the distribution of facial features, are the 1%, 25%, 75% and 99% quantiles, the mean and the standard deviation.

Table 8 gives the emotion classification results for activation, valence and the joint activation-valence space. We report the classification results when these emotional spaces are discretized into 2, 3 and 4 clusters (K). The rows "0 (baseline)" report the baseline results without compensating for the evaluators' reaction lag. The rows "1", "2", and "3" give the performance when fixed delays are included. The rows "Nicolle et al. [28]" report the performance when we use a fixed delay maximizing the correlation-based measure proposed by Nicolle et al. [28]. The rows "Optimal ($\hat{\tau}$)" report the results with the evaluator-independent optimal delay estimations. The rows "EvalDep" and "Pre-Align" give the results of the classifiers in which the time-continuous annotations are shifted with evaluator-dependent compensation schemes.

The table shows, in most of the cases, improvements in accuracy and F-score when a shift is applied to the time-continuous annotations (48 out of 54 cases for both accuracy and F-score). We observe that both metrics increase up to 7% (absolute) compared to the baseline performance. Even when fixed delays are used, the classifiers achieve better performance. These results highlight the importance of compensating for the evaluators' reaction lag. Figures 6(a) and 6(b) show the average accuracy and F-score values obtained across the three tasks (i.e., activation, valence and activation-valence) for different number of K-means clusters. On average, the optimal delay value $\hat{\tau}$ yields the best accuracy and F-score across all conditions. In some cases, however, Table 8 shows improvement by using the evaluator-dependent delay estimations. On average across all conditions, the three proposed annotation correction methods yield statistically significant improvements over the baseline setting for both accuracy (p-value<0.04) and F-score (p-value<0.035). The fixed delay of 3 secs also yields performances comparable to the proposed methods. Notice that the mutual information values for activation and valence for 3-sec are close to the optimal values (see Fig. 3). Therefore, it is not surprising that the classification accuracies are similar.

5.2 Emotion Recognition from Speech

The classification experiments with acoustic features follow a similar methodology as the one used for the classifiers trained with facial features. Since the set of



(a) Speech - Accuracy

(b) Speech - F-score

Fig. 7. Emotion recognition performance with acoustic features by compensating for the evaluators' reaction lag across all conditions (K = 2, 3 and 4, for activation, valence, and joint activation-valence space).

sentence level features for speech include more than 4000 features, we use the *correlation feature selection* (CFS) method to reduce the feature set. This approach is a greedy search method that seeks features that are correlated with the emotional labels, but are not correlated with each other. We use the WEKA's best first search implementation of this technique, which adds a feature at a time to the feature pool and evaluates it based on the correlation measure. This search method makes use of backtracking to avoid local optimum. Since the labels change with the compensation schemes, we independently run the feature selection algorithm for each cross-validation fold, each value of $K \in \{2, 3, 4\}$, each classification task (i.e., activation, valence and joint activation-valence), and each compensation scheme reported in Table 8. The average number of selected features across all conditions is 104 (σ = 14).

Table 8 shows the benefits of correcting the time-continuous labels with the evaluators' reaction lag. Similar to the results observed for facial features, a shift in the annotations improves the performance of the classifiers in most of the cases (44 out of 55 cases). We observe improvements in accuracy and F-score up to 6% (absolute) over the baseline setting. Figures 7(a) and 7(b) give the average performances in terms of accuracy and F-score across conditions, which clearly demonstrate the benefits of compensating for the evaluators' reaction lag. The best performances are achieved by evaluator-dependent lag estimations, which yield more than 7% relative improvement over the baseline. The fixed delay of 3 secs results in higher performance compared to our optimal estimation of delay ($\hat{\tau}$). Notice that the delay is only estimated based on facial expressions. We may have a better estimation of $\hat{\tau}$ if the acoustic features are effectively used to approximate this reaction lag, which is left as a future work.

6 Conclusions and Discussion

Assigning reliable labels to expressive behaviors is a crucial problem in affective computing, impacting the findings on analysis, recognition, synthesis and perception of emotions. This study analyzed and proposed compensation schemes for the evaluators' reaction lag in time-continuous annotations (i.e., reaction time between actual expressive behavior and annotators' responses). Using the maximum mutual information criterion, we quantified this reaction lag by finding the optimum delay that maximizes the relationship between the shifted time-continuous annotations and expressive behaviors represented with facial features. Our empirical and automatic analyses clearly demonstrated the effect of the reaction lag on the recordings of the SEMAINE database, showing delays longer than 3 secs.

The analysis was validated with classification experiments, which, on average, show over 7% relative improvement in accuracy when the annotations are shifted using the detected delays. The results in Table 8 show that the approach proposed by Nicolle et al. [28] also improves the baseline performance without delay compensation. The improvement in performance is even observed when a fixed shift of 3-sec is used. These results demonstrated that compensating for the evaluation reaction lag is important for training and testing emotion recognition systems with time-continuous traces. While, on average, the proposed optimal delay estimation yields slightly higher accuracies than other methods, the differences are not significant. Figure 3 shows that the mutual information values between 3-sec and the optimal delay are very similar. Therefore, it is expected to achieve similar classification results. The 3-sec delay approximately corresponds to the start of the plateau in the mutual information profiles for valence and activation. This suggests that detecting the start of the plateau can also be used to set the optimal delay.

Table 5 reports the estimated delays obtained by analyzing the entire database, under column "All Data". These values are reported as references for other researchers to compensate for the delay in this corpus.

6.1 Limitations and Future Research Directions

One of the limitations of the proposed method is neglecting the vocal cues in the estimation of the reaction lag. However, it is interesting that the classification results with acoustic features also improve when the time-continuous annotations are shifted with the reaction lag estimated with facial features. We believe that higher improvements can be achieved if $\hat{\tau}$ is also estimated with acoustic features. Vocal cues are important indicators perceived by the evaluators and need to be incorporated in the delay estimations. However, the dynamic nature of acoustic features due to the underlying linguistic content makes the delay estimation a challenging problem. Incorporating acoustic features in the estimation of the evaluators' reaction lag remains an open problem.

The proposed approach is based on mutual information which requires the probability density functions of the annotations and facial features. These statistics are estimated with a parametric model over several sessions, producing a constant time shift. However, the evaluators' reaction lag is probably time-variant, depending on many external factors (e.g., ambiguity of the expressive behavior and evaluators' fatigue). Following up with the idea of prealignment, an appealing approach is to replace our correlation-based approach (Sec. 4.3.3) with a nonlinear warping scheme that compensates for local delays (see for example the work of Nicolaou et al. [37], [38]). After this non-linear alignment, the proposed approach can still be used to estimate the global reaction lag.

Our analyses indicate that the evaluators' reaction lag depends on various factors including the annotation dimension (e.g., activation and valence), the intrinsic evaluator-dependent reaction time, and the nature of the stimuli (e.g., sudden changes versus long term shifts). We also hypothesize that the delay depends on the modality presented to the evaluators (e.g., face, speech, speech and face). This work only studied the dependency on the annotators, and expressive dimensional descriptors, which is a limitation of this work. The benefits of compensating for these factors were demonstrated with the empirical and automatic analyses, and with the automatic classification evaluations. Considering other sources of variability is part of our future directions to extract reliable annotations for emotion analysis.

Acknowledgments

This study was funded by Samsung Telecommunications America and NSF (IIS-1217104, IIS-1329659). We thank the MPLab at UCSD for providing the CERT package.

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.

[2] 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.

[3] 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.

[4] R. Cowie, G. McKeown, and E. Douglas-Cowie, "Tracing emotion: An overview," *International Journal of Synthetic Emotions*, vol. 3, no. 1, pp. 1–17, January-June 2012.

[5] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2011)*, Santa Barbara, CA, USA, March 2011, pp. 827–834.

[6] H. Schlosberg, "Three dimensions of emotion," *Psychological review*, vol. 61, no. 2, p. 81, March 1954.

[7] J. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of research in Personality*, vol. 11, no. 3, pp. 273–294, September 1977.

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

[9] D. Grandjean, D. Sander, and K. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and cognition*, vol. 17, no. 2, pp. 484–495, June 2008.

[10] A. Batliner, S. Steidl, and E. Nöth, "Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo emotion corpus," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Philadelphia, PA, USA, May 2008, pp. 28–31.

[11] 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.

[12] 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.

[13] 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.

[14] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[15] 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.

[16] Y. Kim, E. Schmidt, and L. Emelle, "MoodSwings: A collaborative game for music mood label collection," in *International Symposium on Music Information Retrieval (ISMIR 2008)*, Marrakech, Morocco, September 2008, pp. 28–31.

[17] F. Nagel, R. Kopiez, O. Grewe, and E. Altenmüller, "EMuJoy: Software for continuous measurement of perceived emotions in music," *Behavior Research Methods*, vol. 39, no. 2, pp. 283–290, May 2007.

[18] P. Ekman and W. Friesen, "Constants across cultures in the face and emotion," *Journal of Personality and Social Psychology*, vol. 17, no. 2, pp. 124–129, March 1971.

[19] P. Ekman, W. Friesen, M. Sullivan, A. Chan, A. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. LeCompte, T. Pitcairn, P. Ricci-Bitti, K. Scherer, and M. Tomita, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology*, vol. 53, pp. 712–717, October 1987.

[20] 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.

[21] L. Vidrascu and L. Devillers, "Real-life emotions in naturalistic data recorded in a medical call center," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, Genoa,Italy, May 2006, pp. 20–24.

[22] 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.

[23] 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.

[24] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[25] A. Metallinou, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database: A multimodal database of theatrical improvisation," in *Workshop on Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality (MMC 2010)*, Valletta, Malta, May 2010.

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

[27] N. Malandrakis, A. Potamianos, G. Evangelopoulos, and A. Zlatintsi, "A supervised approach to movie emotion tracking," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 2376–2379.

[28] 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.

[29] J. Arias, C. Busso, and N. Yoma, "Shape-based modeling of the fundamental frequency contour for emotion detection in speech," *Computer Speech and Language*, vol. 28, no. 1, pp. 278–294, January 2014.

[30] B. Schuller, M. Valstar, F. Eyben, G. McKeown, R. Cowie, and M. Pantic, "AVEC 2011- the first international audio/visual emotion challenge," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 415–424.

[31] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 183–196, April-June 2013.

[32] 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.

[33] J. C. Kim, H. Rao, and M. Clements, "Investigating the use of formant based features for detection of affective dimensions in speech," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 369–377.

[34] H. Meng and N. Bianchi-Berthouze, "Naturalistic affective expression classification by a multi-stage approach based on hidden Markov models," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 378–387.

[35] L. Cen, Z. L. Yu, and M. Dong, "Speech emotion recognition system based on L1 regularized linear regression and decision fusion," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 332–340.

[36] 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.

[37] M. Nicolaou, V. Pavlovic, and M. Pantic, "Dynamic probabilistic CCA for analysis of affective behaviour," in *European Conference on Computer Vision (ECCV 2012)*, Florence, Italy, October 2012, pp. 98–111.

[38] ——, "Dynamic probabilistic CCA for analysis of affective behaviour and fusion of continuous annotations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[39] E. Douglas-Cowie, R. Cowie, C. Cox, N. Amir, and D. Heylen, "The sensitive artificial listener: an induction technique for generating emotionally coloured conversation," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 1–8.

[40] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, pp. 22–35, September 2006.

[41] P. Ekman and W. Friesen, *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Palo Alto, CA, USA: Consulting Psychologists Press, 1978.

[42] B. Schuller, S. Steidl, A. Batliner, F. Schiel, and J. Krajewski, "The INTERSPEECH 2011 speaker state challenge," in *12th Annual Conference of the International Speech Communication Association (Interspeech 2011)*, Florence, Italy, August 2011, pp. 3201–3204.

[43] 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.

[44] J. Arias, C. Busso, and N. Yoma, "Energy and F0 contour modeling with functional data analysis for emotional speech detection," in *Interspeech 2013*, Lyon, France, August 2013, pp. 2871–2875.

[45] G. Ramirez, T. Baltrusaitis, and L. Morency, "Modeling latent discriminative dynamic of multi-dimensional affective signals," in *Affective Computing and Intelligent Interaction (ACII 2011)*, ser. Lecture Notes in Computer Science, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Memphis, TN, USA: Springer Berlin / Heidelberg, October 2011, vol. 6975/2011, pp. 396–406.

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

[47] C.-C. Lee, C. Busso, S. Lee, and S. Narayanan, "Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1983–1986.

[48] M. Wöllmer, F. Eyben, B. Schuller, E. Douglas-Cowie, and R. Cowie, "Data-driven clustering in emotional space for affect recognition using discriminatively trained LSTM networks," in *Interspeech 2009*, Brighton, UK, September 2009, pp. 1595–1598.

[49] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S. Narayanan, "Context-sensitive learning for enhanced audiovisual emotion classification," *IEEE Transactions on Affective Computing*, vol. 3, no. 2, pp. 184–198, April-June 2012.

[50] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, June 2009.

[51] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences*. Upper Saddle River, NJ, USA: Prentice-Hall, 2006.

Soroosh Mariooryad (S'12) received his B.S degree (2007) with high honors in computer engineering from Ferdowsi University of Mashhad, and his M.S degree (2010) in computer engineering (artificial intelligence) from Sharif University of Technology (SUT), Tehran, Iran. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of Texas at Dallas (UTD), Richardson, Texas, USA. From 2008 to 2010, he was a member of the Speech Processing Lab (SPL) at SUT. In 2010, he joined as a research assistant the Multimodal Signal Processing (MSP) laboratory at UTD. In summer 2013, he interned at Microsoft Research working on analyzing speaking style characterstics. His research interests are in speech and video signal processing, probabilistic graphical models and multimodal interfaces. His current research includes modeling and analyzing human non-verbal behaviors, with applications to speech-driven facial animations and emotion recognition. He has also worked on statistical speech enhancement and fingerprint recognition.

Carlos Busso (S'02-M'09-SM'13) is an Assistant Professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He received his B.S (2000) and M.S (2003) degrees with high honors in electrical engineering from University of Chile, Santiago, Chile, and his Ph.D (2008) in electrical engineering from University of Southern California (USC), Los Angeles, USA. 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 received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the coauthor of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests are in 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.