Analysis and Compensation of the Reaction Lag of Evaluators in Continuous Emotional Annotations

Soroosh Mariooryad and Carlos Busso
Multimodal Signal Processing (MSP) Laboratory, Department of Electrical Engineering
The University of Texas at Dallas, Richardson TX 75080, USA
Email: soroosh.ooryad@utdallas.edu, busso@utdallas.edu

***Abstract*—Defining useful emotional descriptors to characterize expressive behaviors is an important research area in affective computing. Recent studies have shown the benefits of using continuous emotional evaluations to annotate spontaneous corpora. Instead of assigning global labels per segments, this approach captures the temporal dynamic evolution of the emotions. A challenge of continuous assessments is the inherent reaction lag of the evaluators. During the annotation process, an observer needs to sense the stimulus, perceive the emotional message, and define his/her judgment, all this in real time. As a result, we expect a reaction lag between the annotation and the underlying emotional content. This paper uses mutual information to quantify and compensate for this reaction lag. Classification experiments on the SEMAINE database demonstrate that the performance of emotion recognition systems improve when the evaluator reaction lag is considered. We explore annotator-dependent and annotator-independent compensation schemes.**

***Keywords*—Emotion Recognition, Continuous Emotion Annotation**

I. Introduction

Defining reliable emotional labels is a critical aspect in the area of affective computing [1], [2], [3], [4]. The labels are used as ground truth to train and evaluate emotion recognition systems. Therefore, the success of these systems is tied to the quality of the emotional labels. The common approach in the field of affective computing is the use of discrete labels such happiness and anger, or continuous emotional attributes such as activation (calm versus active) and valence (negative versus positive). These labels are usually assigned to specific segments (e.g., words, chunks or sentences) by multiple external observers who annotate the perceived emotion of the stimulus.

One limitation of assigning a label per segment is that the emotional descriptors cannot capture emotional variations within the segment (e.g., sharp changes, hot spots). While decreasing the length of the segments to smaller units such as words can solve part of the problem [5], it is not clear whether evaluators can perceive the emotional content from short segments. An appealing, alternative approach is to track the perceived emotional content of the stimuli continuously over time (e.g., many labels per second) [3], [4]. This approach is possible with *graphical user interfaces* (GUIs) that record the position of a cursor controlled by the user over a coordinate system with emotional dimensions as axes. Examples of these GUIs are FEELTRACE [6] and *Gtrace* [4] for evaluating audiovisual recordings and *MoodSwings* [7] and *EmuJoy* [8] for evaluating emotional music.

Scherer proposed the use of an adapted version of the Brunswik's lens model to study vocal communication of emotions [9]. This model makes an explicit distinction between the encoding (speaker), the transmission, and the representation (listener) of paralinguistic messages. The speaker encodes his/her communicative goals (trait or state), resulting in the production of distal indicators. These distal cues are manifested through modulation over various communication channels (e.g., speech and gestures). These distal indicators are transmitted to the listener, who perceives the information, referred to as proximal percepts and makes inferences about their attributes. These proximal precepts are mapped into neural representation, resulting in perceptual judgments. In this context, the assumption that evaluators can provide instantaneous assessments that are tightly aligned with the emotional content of the stimulus is not realistic (see Fig. 1). The evaluators have to sense the stimuli, perceive the emotional message, and define their judgment, before moving the cursor to the selected location. This process introduces a delay between the emotional annotations and emotional content of the stimulus [3]. This paper studies this delay, which is referred to as the *evaluator's reaction lag*, using mutual information framework. We present experimental results on the SEMAINE database [10], demonstrating that an emotion recognition system can significantly improve its performance when this delay is considered.

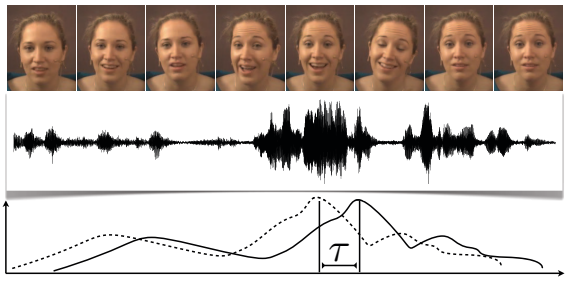


Fig. 1. Illustration of the evaluators' reaction lag between the annotations (solid line) and the underlying emotional content (dashed line).

II. Background

A. Related Work

Since global discrete affective labels associated to multimedia content cannot represent localized events, recent studies have considered continuous emotional annotations [4]. This approach facilitates the analysis of complete dialogs without the need of pre-segmenting the interactions into turns. It also offers the opportunity to study emotional modulation over different segments (e.g., sentence, phrase, chunk, word, syllable, phoneme). Because of these advantages, continuous emotional annotations have been used to evaluate the emotional content of TV programs [11] and movie clips [12]. However, this type of annotation introduces unique challenges [3].

978-0-7695-5048-0/13 $26.00 © 2013 IEEE
DOI 10.1109/ACII.2013.21

We believe that one of the main problems of continuous emotional annotations is the delay between the labels and the emotional content of the stimulus. This delay is caused by the reaction time that an evaluator takes from sensing the emotional cues to adjusting the evaluation scores. Nicolle et al. [13] identified this problem in their studies using the SEMAINE corpus. They assumed a linear relationship between the user's facial features and the annotated scores. They proposed a correlation-based measurement to find the delay, which was separately measured over four emotion primitive dimensions (activation, valence, expectation and power). They compared the correlation of facial features with the delayed emotional annotations across different delay intervals. They presented the probability distribution of the delay. The most probable delay was between three to six seconds for different primitive dimensions. It is interesting that the emotion recognition performance on the SEMAINE database are usually low [14], [15], [16]. While the corpus provides spontaneous recording with ambiguous emotions that are hard to predict, the use of labels that are not representative of the instantaneous expressive behaviors may cause lower performances.

Other related studies attempt to compensate for inter-evaluator delays. Nicolaou et al. [17] estimated constant shifts between the evaluations to minimize their mean square error. The authors later introduced the *dynamic probabilistic canonical correlation with time warping* (DPCTW) approach to compensate for localized shifts between annotations of different evaluators [18].

B. SEMAINE Database

We use the *sustained emotionally colored machine-human interaction using nonverbal expression* (SEMAINE) database [10] to analyze the *evaluator's reaction lag*. This database uses the *sensitive artificial learner* (SAL) technique [19] to elicit emotional reactions of users interacting with an operator. The operator has a given personality and his/her goal is to induce emotional reactions on the users. The SEMAINE database was recorded using different implementations of the SAL technique: solid SAL (i.e., human operator), semi-automated SAL (a virtual character operator controlled by humans), and automated-SAL (i.e., a virtual character operator controlled by a dialog management system). This study considers the solid SAL portion of the corpus. Using teleprompter screens the user and operator interact with each other while sitting in separate rooms. The speech and videos are simultaneously recorded.

The user's behaviors are continuously annotated over time in terms of activation (i.e., active versus passive), and valence (i.e., negative versus positive). The database also provides continuous annotations of other aspects including power, and expectation dimensions, and continuous labels describing the intensity of categorical attributes such as happiness, anger, interested, and thoughtfulness. We only consider activation and valence as the emotional descriptors, which are the most common dimensions used in related studies. The emotional evaluation is conducted with the FEELTRACE toolkit [6], which consists of a GUI displaying the activation/valence space. The annotator moves the cursor controlled by the computer's mouse over the GUI as he/she watches and evaluates the emotional content of the video. In contrast to segment level evaluations, these continuous annotations capture the temporal evolution of the perceived emotional content. Each conversation session is annotated by multiple evaluators (2 to 8). These annotations are mapped into the interval [-1, +1].

TABLE I. THE LIST OF *action units* (AUS) EXTRACTED BY CERT [20].

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

C. Facial and Acoustic Features

The user's videos include the profile and frontal views. We use the *computer expression recognition toolbox* (CERT) [20] to automatically extract facial features from the frontal view videos. The CERT toolkit makes use of a frame-by-frame processing method, yielding robust performance against different illumination settings. CERT estimates the values of *action units* (AUs), defined in the *facial action coding system* (FACS) [21]. The AUs describe movements of individual facial muscles or a group of facial muscles. Table I lists the 20 AUs used to represent facial expressions. In addition, we estimate three head rotation parameters using CERT (i.e., pitch, yaw, roll movements). Out of 94 currently released SEMAINE sessions, only 52 sessions are emotionally annotated. In eight of these sessions (sessions 82, 88-91, 95-97), CERT does not detect the user's face. Hence, only the remaining 44 sessions are used for the analysis and classification experiments. There are nine unique speakers in this portion of the corpus.

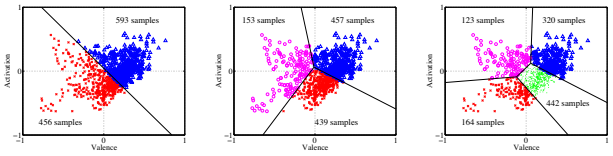
The emotion classification experiments use the exhaustive set of acoustic features introduced for the Interspeech 2011 speaker state challenge [22]. This set contains 4368 turn level features extracted from a set of frame-level features (i.e., statistics or functional extracted from a given speech segment). This set includes prosodic, spectral and voice quality features, and it is extracted with openSMILE [23]. Detailed description of this feature set can be found in Schuller et al. [22].

III. ANALYSIS OF THE EVALUATORS' REACTION LAG

As mentioned in Section II-B, the analysis uses the SEMAINE database, which was continuously evaluated with the FEELTRACE toolkit. We choose this corpus given that it is one of the largest naturalistic emotional corpora with continuous emotional annotations. This section describes the proposed approach to capture the reaction lag of the evaluators (Sec. III-A), and the results of the analysis which provide the optimum delay between the emotional annotation and the underlying emotional content (Sec. III-B).

A. Proposed Approach

The goal of this paper is to estimate the optimum delay between the continuous frame-by-frame emotional annotations and the underlying emotional content of the recordings. The proposed framework relies on mutual information, which is defined in Equation 1. The variables X and Y are characterized by their *probability mass functions* (PMFs) $p(x)$ and $p(y)$ and their joint PMF $p(x, y)$. Mutual information measures the dependency between two discrete random variables. Therefore, it provides an appealing approach to capture the dependency



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

Fig. 2. The clusters obtained with the K-means algorithm on the activation-valence space considering no delay on the annotations (τ=0).

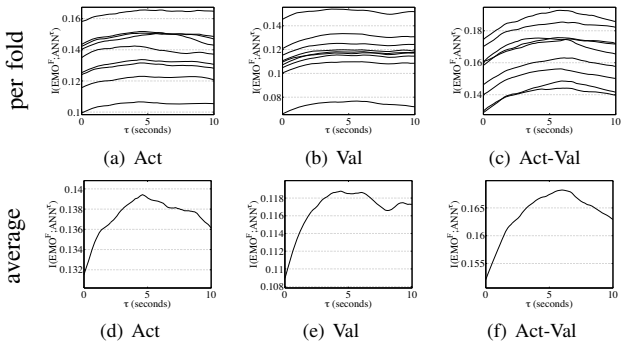
between the emotional content of the stimulus (EMO) and a τ-sec-shifted version of the emotional annotations (ANN^{τ}). The optimum reaction lag is given by the shift $\hat{\tau}$ that maximizes the dependency between these variables (Eq. 2).

$$I[X;Y] = \sum_{y\in Y}\sum_{x\in X} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \quad (1)$$

$$\hat{\tau} = \arg_{\tau}\max I[EMO;ANN^{\tau}] \quad (2)$$

While ANN^{τ} is derived by shifting the emotional annotations, the underlying emotional content EMO is not available. However, this metric can be approximated by estimating a continuous emotional profile describing the deviations in the features from normal behaviors. This study only uses facial features (EMO^{F}) for the following reasons. First, AUs measure deviations from normal facial poses. Therefore, they are adequate to represent the emotional profile. Second, it is possible to estimate facial features even when the subject is listening, while acoustic features are only available during speaking turns. Finally, conducting the analysis on facial expression allows us to approximate the underlying emotional content on a frame-by-frame basis. Prosodic features such as F0 contour and energy are suprasegmental features, in which the emotional information is conveyed by changes on the temporal patterns rather than on the actual values of the features. This is the reason that emotion classifiers from speech usually rely on statistics of the acoustic features estimated from longer segments (e.g., sentences, words, chunks). In contrast, AUs describe the actual facial expressions displayed on the given frames. Notice that emotion perception is also affected by speech properties, which is not be captured by our approach. Likewise, the estimation of the emotional profile depends on the robustness of CERT to provide reliable AUs information. In spite of these limitations, the experimental results validate the proposed approach, which increases the emotion recognition performance (Sec. IV).

The variables EMO^{F} and ANN^{τ} are discrete. Therefore, we use the K-means algorithm to define clusters for the continuous values of facial features and emotional annotations. We estimate the required PMFs from these nonuniform bins. For facial features, EMO^{F}, we use $K \in \{2,4,6,8,10,16,20\}$ over the joint feature space defined by the 23 facial features (Sec. II-C). We report the average results across these settings. For the τ-sec-shifted emotional annotations, ANN^{τ}, we follow a similar approach. For a given τ, we estimate the mean value of the emotional annotations across evaluators. We assume that the inter-evaluator differences in reaction lag is negligible. Notice that we relax this assumption in Section IV-C by considering different reaction's lags across evaluators. Then, we estimate the mean value for each sentence. We transform the continuous values of the activation/valence space into discrete



(a) Act (b) Val (c) Act-Val

(d) Act (e) Val (f) Act-Val

Fig. 3. Analysis of the *evaluator's reaction lag* in the emotional annotations of the SEMAINE database. The figures show the mutual information between facial features and τ-sec-delayed emotion annotations. The results are presented for each fold for (a) activation, (b) valence and (c) joint activation-valence space, and the average across all folds for (d) activation, (e) valence and (d) joint activation-valence space.

classes using the K-means algorithm with $k \in \{2,3,4\}$. This clustering is separately done for each dimension, and on the joint activation-valence space. For example, Figure 2 shows the clusters built on the activation-valence space when τ=0. The results of the analysis (Sec. III-B) and the classification experiments (Sec. IV) are reported for each dimension and for the joint activation-valence space. Notice that previous studies on emotion recognition have also used this approach to transform continuous emotional attributes into K discrete emotional classes [24], [25], [26].

B. Quantifying the Optimum Delay

We evaluate different delay values by changing τ from 0 to 10 sec in jumps of 40 ms. We assume that the reaction lag is a wise sense stationary random variable. Since the optimum evaluators' reaction lag will be used in the classification experiments (Sec. IV), we estimate $\hat{\tau}$ using a nine-fold speaker-independent, cross-validation scheme. In this approach, we only use the training set to estimate the delay (data from 8 speakers). Therefore, the classification results, which are evaluated on the testing set, are not biased by the selection of the delay. As a result, we estimate a different $\hat{\tau}$ for each of the nine folds.

Figures 3 (a-c) show examples of the average mutual information observed for each fold across different values of τ. The figures describe the activation, valence and joint activation-valence space with K = 3. These figures clearly illustrate the increase in mutual information between the emotional annotations and the emotional content on the facial features as τ increases from zero. We observe that the curves reach a plateau followed by a gradual decrease in the mutual information. This pattern is clearly seen in Figures 3 (d-f), which show the average of the individual curves across folds. This pattern is consistent with the probability distribution of the delay reported by Nicolle et al. [13].

We define the optimum evaluators' reaction lag as the minimum value of τ for which the derivative of $I[EMO^{F};ANN^{\tau}]$ is non-positive. While this approach does not always give the maximum value of the mutual information, it favors shorter delays by selecting the beginning of the plateau region. We estimate $\hat{\tau}$ for each fold generating nine different delays for

TABLE II. MEAN AND STANDARD DEVIATION OF THE ESTIMATED *evaluator's reaction lags* ACROSS THE NINE FOLDS FOR ACTIVATION, VALENCE AND JOINT ACTIVATION-VALENCE SPACE (SEE FIG. 3).

Attribute	K = 2		K = 3		K = 4	
	mean	std	mean	std	mean	std
Act	2.27	0.82	2.84	1.21	3.94	1.55
Val	3.48	0.66	3.68	0.86	3.37	0.79
Act-Val	3.61	0.52	4.98	0.84	4.43	1.36

activation, valence, and the activation-valence space. Table II reports the mean and standard deviation of the estimated delays for each of these dimensions, and for different values of emotional clusters (K). The table reports reaction lags in the annotations ranging from 2-5 secs. Given that the median duration of the sentences in this corpus is 2.76 sec, it is very important to consider this delay in emotion recognition experiments.

IV. EMOTION RECOGNITION RESULTS

This section presents emotion classification experiments that validate the need for compensating the reaction lag of the evaluators. We build classifiers that are trained and tested with and without shifting the emotional labels (nine-fold speaker independent cross-validation scheme). In addition to the optimum value of $\hat{\tau}$, we implemented fixed shifts of 1, 2 and 3 seconds. The experiments are conducted at the turn level with the segmentation provided in the SEMAINE corpus. We only consider the user's speaking turns with at least 300 ms duration, resulting in 1049 segments. Similar to the analysis section, the ground truth for the emotional labels are defined by averaging the annotation scores across the evaluators, and across the duration of the speaking turn. In this estimation, we consider the value of τ before estimating the average emotion evaluation scores. The values are then transformed into categorical classes using the clusters provided by the K-means algorithm for activation, valence and activation-valence space. Since the value of $\hat{\tau}$ varies across folds, the emotional labels used for the emotion recognition experiments may change for each cross validation.

All the classifiers are implemented with linear kernel *support vector machine* (SVM). We use the WEKA [27] implementation with *sequential minimal optimization* (SMO) training. The complexity parameter of SVM is set to c = 0.1 for all the classifiers. Since, the data is not uniformly distributed between the classes (see Fig. 2), the recognition results are reported in terms of *accuracy* (A), macro-average *precision* (P), macro-average *recall* (R) and macro-average *F-score* (F). Equation 3 gives the macro-average F-score. We use the large sample proportion hypothesis test to measure whether the differences in performance are statistically significant. We conduct separate experiments for emotion recognition from facial (Sec. IV-A) and acoustic (Sec. IV-B) features.

$$F = \frac{2PR}{P+R} \tag{3}$$

A. Emotion Recognition from Face

For each turn, a set of six statistics are extracted from each facial feature creating a 138D feature vector (i.e., [20 AUs + 3 head rotation] $\times$ 6 statistics). The statistics are the 1%, 25%, 75% and 99% quantiles, the mean and standard deviation. Table III reports the classification performances for different settings by considering different delays on the annotations (0, 1, 2, 3 and $\hat{\tau}$ seconds). The table shows improvement in performance in most of the experiments in which the emotional labels are shifted (33 out of 36 cases for accuracy; 32 out of 36 cases for F-score). On average across all the conditions, the optimal value of $\hat{\tau}$ gives the best performance in accuracy and F-score. In the activation-valence space, the optimal delay always gives statistically significant improvements in F-score ($p - value$ <0.0073) and accuracy ($p - value$ <0.0441) over the baseline settings without delay compensation.

B. Emotion Recognition from Speech

Given the size of the corpus and the high dimension of the acoustic feature vector, we use the *correlation feature selection* (CFS) method to reduce the number of acoustic features. We employ the WEKA's best first search algorithm implementation. This non-wrapper-based method selects the features with low correlation between themselves and high correlation with the target labels. The search method uses the correlation metrics to evaluate the effect of adding one feature at a time. This search is also enhanced by backtracking option. The feature selection is separately performed for each fold and for each delay using only the training set for $K \in \{2, 3, 4\}$. The average number of selected features for activation, valence and joint activation-valence spaces are 103 (σ = 15), 87 (σ = 17) and 106 (σ = 13), respectively. Table IV gives the speech emotion recognition experiments before and after compensating for the delays. The delayed annotations always improve the accuracy of the baseline (i.e., no delay), excepting the case for activation with K = 3. Even when we use fixed delays, we observe improvements in performances for most of the cases. In the activation-valence space, the increase in accuracy and F-score for the optimal delay is statistically significant when K = 2 and 4 ($p - value$ <0.0069). As an aside, it is interesting that we achieved better performance on valence than on activation, since acoustic features do not discriminate very well across valence dimension [28].

C. Pre-Aligning the Annotations of Evaluators

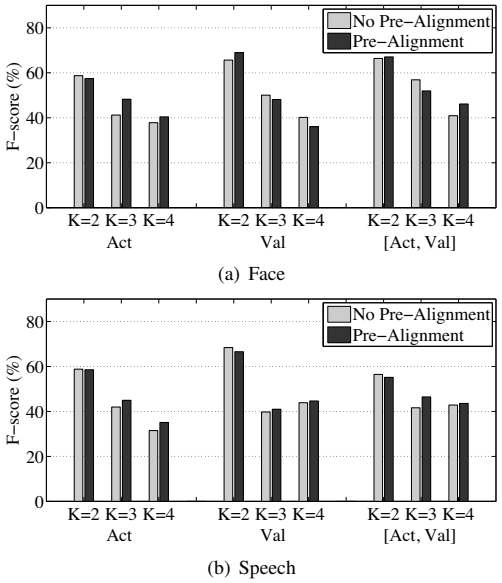
The analysis assumes that the *reaction lag* is consistent across evaluators. However, it is expected that the perception process will introduce delays that are evaluator-dependent [17], [18]. This section considers the variability in the reaction time across the evaluators. The approach consists in pre-aligning the evaluations before estimating the optimum value of τ. We consider two approximations: a) the phase between two different annotations is fixed across time (e.g., time-invariant), and b) this phase is less than one second. For each session, we select at random one of the annotations, which is set as the reference. Then, we estimate the best alignment by computing the correlation between this reference annotation and each of the other annotations, which are shifted from -1 sec to 1 sec. After the emotional annotations are aligned, we estimate the reaction lag between the average, aligned annotation scores and the facial features using the same approach described in Section III-A. The optimum values for τ are selected for each fold and used to replicate the recognition experiments presented in sections IV-A and IV-B.

TABLE III. FACIAL EMOTION RECOGNITION ON ACTIVATION, VALENCE AND ACTIVATION-VALENCE, CONSIDERING DIFFERENT NUMBER OF CLUSTERS (K). RESULTS ARE REPORTED FOR DELAYS OF 0, 1, 2, 3 AND $\hat{\tau}$ SEC, IN TERMS OF *Accuracy* (A), *average precision* (P), *average recall* (R) AND *F score* (F).

Attribute	lag	K = 2 [chances = 50%]				K = 3 [chances = 33.33%]				K = 4 [chances = 25%]			
		A	P	R	F	A	P	R	F	A	P	R	F
Activation	0	60.06	57.66	56.85	57.25	41.75	39.27	38.32	38.79	32.60	30.10	30.85	30.47
	1	60.27	58.80	58.97	58.88	46.35	48.17	46.72	47.43	34.17	32.62	32.21	32.41
	2	60.64	60.16	59.04	59.59	45.77	49.42	46.29	47.80	35.96	34.31	34.33	34.32
	3	61.06	54.12	54.33	54.22	45.50	46.13	45.36	45.74	30.92	34.10	32.18	33.11
	Optimal	58.24	58.71	58.72	58.71	47.70	40.82	41.61	41.21	35.95	36.08	39.65	37.78
Valence	0	67.49	69.19	67.82	68.50	53.00	46.96	50.65	48.74	34.13	28.90	28.09	28.49
	1	68.52	64.83	63.73	64.28	53.93	61.21	51.99	56.22	37.14	36.43	35.72	36.07
	2	72.40	74.13	70.38	72.21	54.62	57.29	47.20	51.76	37.71	26.48	27.18	26.83
	3	70.06	71.91	68.72	70.28	55.48	55.96	55.05	55.50	39.92	39.38	38.16	38.76
	Optimal	69.12	67.24	64.16	65.66	54.76	51.17	48.98	50.05	39.37	39.96	40.33	40.14
[Activation, Valence]	0	56.24	59.50	59.34	59.42	50.52	49.39	48.20	48.79	37.27	36.55	34.94	35.73
	1	59.50	65.61	65.94	65.77	44.05	55.72	44.82	49.68	39.35	38.58	41.54	40.01
	2	62.78	62.64	63.29	62.96	50.73	53.05	51.23	52.12	43.25	44.45	44.99	44.72
	3	63.11	64.09	63.30	63.69	54.99	52.50	51.15	51.82	40.22	36.02	36.71	36.36
	Optimal	64.12	65.74	67.03	66.38	59.17	57.32	56.44	56.88	40.90	40.34	41.50	40.91

TABLE IV. SPEECH EMOTION RECOGNITION ON ACTIVATION, VALENCE AND ACTIVATION-VALENCE, CONSIDERING DIFFERENT NUMBER OF CLUSTERS (K). RESULTS ARE REPORTED FOR DELAYS OF 0, 1, 2, 3 AND $\hat{\tau}$ SEC, IN TERMS OF *Accuracy* (A), *average precision* (P), *average recall* (R) AND *F score* (F).

Attribute	lag	K = 2 [chances = 50%]				K = 3 [chances = 33.33%]				K = 4 [chances = 25%]			
		A	P	R	F	A	P	R	F	A	P	R	F
Activation	0	55.3	58.92	58.62	58.77	42.62	42.35	44.4	43.35	29.73	30.14	30.33	30.23
	1	56.15	56.58	56.42	56.5	41.96	41.18	41.34	41.26	30.47	30.18	30.97	30.57
	2	57.35	57.31	57.32	57.31	42.45	44.83	46.65	45.72	30.62	28.85	30	29.41
	3	57.47	58.32	58.41	58.36	41.57	39.51	39.26	39.38	32.07	31.92	31.27	31.59
	Optimal	57.93	58.74	58.84	58.79	42.07	41.77	42.17	41.97	32.9	31.31	31.68	31.49
Valence	0	66.31	63.94	59.1	61.42	41.88	45	42.17	43.54	40.94	42.87	39.08	40.89
	1	67.09	66.74	66.95	66.84	42.36	42.45	45.07	43.72	41.42	38.61	38.75	38.68
	2	67.26	65.38	66.1	65.74	45.25	45.28	46.66	45.96	41.63	41.83	41.24	41.53
	3	68.13	64.06	62.4	63.22	43.29	40.33	40.59	40.46	43.63	41.13	40.17	40.64
	Optimal	67.47	67.85	68.95	68.4	43.1	39.77	39.76	39.76	42.77	44.23	43.53	43.88
[Activation, Valence]	0	47.99	49.46	49.48	49.47	42.42	43.5	43.65	43.57	37.38	36.98	38.22	37.59
	1	50.07	49.94	49.95	49.94	44.53	43.56	44.49	44.02	39.05	41.02	40.1	40.55
	2	52.02	54.96	55.16	55.06	45.52	47.69	45.3	46.46	41.49	39.48	38.04	38.75
	3	53.89	55.3	55.08	55.19	47.08	45.65	46.07	45.86	38.82	36.1	35.36	35.73
	Optimal	53.82	56.3	56.61	56.45	45.11	41.63	41.63	41.63	42.19	45.01	40.88	42.85



(a) Face

(b) Speech

Fig. 4. The F-score achieved with $\hat{\tau}$ for facial and speech emotion recognitions with and without pre-aligning the annotations. Pre-aligning the annotations improves slightly the performance.

The results after pre-aligning the emotional annotations are slightly better than the ones achieved without pre-alignment. Pairwise comparisons of the results on Tables III and IV with the corresponding ones obtained after the pre-alignment step (not reported in the paper) show 1.06% and 0.26% improvements, on average, in F-score for face and speech emotion recognition, respectively (across all conditions). Figure 4 shows the F-scores obtained for the optimal reaction lag $\hat{\tau}$ with and without the pre-alignment step. According to this figure, in most of the cases we observe an improvement in F-score. However, the difference in performance is small.

V. CONCLUSIONS AND FUTURE WORK

This paper studied the reaction lag of evaluators using continuous emotional annotations. The analysis is based on mutual information measurements, which clearly demonstrates the delay between the emotional annotations and the underlying emotional content. We conducted exhaustive emotion recognition experiments to validate the importance of considering the delay. The systems trained with shift-delayed emotional annotations achieved statistically significant improvements over baseline systems by compensating the evaluators' reaction lag.

Although the estimated delays are obtained by comparing the mutual information of annotation scores with only facial features, the improvements were observed in both facial and speech emotion recognition systems. Our future work will consider the estimation of emotional content using speech features. We will derive these metrics by contrasting emotional speech with the pattern observed in neutral speech using neutral models [29]. Likewise, the analysis shows evaluator-dependent patterns on the annotations. This variability should be considered to achieve more reliable emotional labels. Furthermore, the underlying assumption of this study is that the

delay is consistent across the evaluation sessions. However, the delay may be not only evaluator-dependent, but also time-variant. Therefore, we plan to explore approaches to estimate delays over localized segments. For example, we can couple the proposed approach with the DPCTW method presented by Nicolaou et al. [18]. These techniques can provide more robust labels to train emotion recognition systems.

ACKNOWLEDGMENT

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.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. Oxford University Press, New York, NY, USA, 2013.

[2] R. Cowie and R.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.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] 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.

[6] 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, September 2000, ISCA, pp. 19–24.

[7] Y.E. 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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)*, S. DMello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., vol. 6975/2011 of *Lecture Notes in Computer Science*, pp. 378–387. Springer Berlin / Heidelberg, Memphis, TN, USA, October 2011.

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

[17] M.A. 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, pp. 92–105, April-June 2011.

[18] M.A. 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.

[19] 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.

[20] M.S. Bartlett, G.C. Littlewort, M.G. Frank, C. Lainscsek, I. Fasel, and J.R. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, pp. 22–35, September 2006.

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

[22] 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.

[23] 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)*, Firenze, Italy, October 2010, pp. 1459–1462.

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

[25] 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.

[26] A. Metallinou, M. Wöllmer, A. Katsamanis, F. Eyben, B. Schuller, and S.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.

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

[28] C. Busso and T. Rahman, "Unveiling the acoustic properties that describe the valence dimension," in *Interspeech 2012*, Portland, OR, USA, September 2012, pp. 1179–1182.

[29] J.P. Arias, C. Busso, and N.B. Yoma, "Energy and F0 contour modeling with functional data analysis for emotional speech detection," in *Interspeech 2013*, Lyon, France, August 2013.