Exploring Cross-Modality Affective Reactions for Audiovisual Emotion Recognition

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

Abstract—Psycholinguistic studies on human communication have shown that during human interaction individuals tend to adapt their behaviors mimicking the spoken style, gestures and expressions of their conversational partners. This synchronization pattern is referred to as entrainment. This study investigates the presence of entrainment at the emotion level in cross-modality settings and its implications on multimodal emotion recognition systems. The analysis explores the relationship between acoustic features of the speaker and facial expressions of the interlocutor during dyadic interactions. The analysis shows that 72% of the time the speakers displayed similar emotions, indicating strong mutual influence in their expressive behaviors. We also investigate the cross-modality, cross-speaker dependency, using mutual information framework. The study reveals a strong relation between facial and acoustic features of one subject with the emotional state of the other subject. It also shows strong dependency between heterogeneous modalities across conversational partners. These findings suggest that the expressive behaviors from one dialog partner provide complementary information to recognize the emotional state of the other dialog partner. The analysis motivates classification experiments exploiting cross-modality, cross-speaker information. The study presents emotion recognition experiments using the IEMOCAP and SEMAINE databases. The results demonstrate the benefit of exploiting this emotional entrainment effect, showing statistically significant improvements.

Index Terms—Entrainment, multimodal interaction, cross-subject multimodal emotion recognition, facial expressions, emotionally expressive speech.

1 Introduction

During human interaction, individuals tend to adapt their verbal and non-verbal behaviors, synchronizing their spoken style, gestures and expressions with the ones of their conversational partners. This phenomenon occurs in various aspects of the conversation, including choice of words [1], pronunciation [2], speaking rate [3], tone [4], [5], head motion [6], body gestures [7] and postures [8]. This effect is known as entrainment, alignment, adaptation or accommodation [9]. The study of entrainment provides opportunities to enhance human-machine interaction systems. For instance, a virtual agent was used to affect the speaking rate of a user when his/her speech was too fast or too slow [10]. By adapting the users' speaking rate, the performance of an *automatic speech recognition* (ASR) system can improve, given the decrease in mismatches between training and testing conditions. Likewise, studies have shown improvement in efficiency and user satisfaction when a spoken dialog system is entrained to the users' behaviors [11]. This study explores emotional entrainment effects in dyadic spontaneous interactions, and their implications on multimodal emotion recognition systems.

• *This study was funded by Samsung Telecommunications America and* National Science Foundation *(NSF) grant IIS 1217183.*
The authors are with the Multimodal Signal Processing (MSP) laboratory, The University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: soroosh.ooryad@utdallas.edu, busso@utdallas.edu).

The first goal of this study is to understand the emotional entrainment effect during spontaneous interactions. We present a thorough analysis using the *interactive emotional dyadic motion capture* (IEMOCAP) database. First, we study the co-occurrence of the emotional states of the speakers and listeners. The result shows that in 72% of the conversation turns the two subjects presented similar emotions. Given that the dialog partners' emotions are synchronized most of the time (i.e., mirroring behavior), we hypothesize that they display behaviors that are characteristic of the given joint emotional state. As a result, the expressive behaviors from one subject should be correlated with the behaviors of his/her conversation partner. To address this hypothesis, this study analyzes cross-subject emotional entrainment using mutual information. The analysis shows that the cues from one subject (i.e., acoustic or facial features) provide additional information about the emotional state of the dialog partner. Furthermore, we observe that the mutual information between the behaviors from conversation partners (i.e., paired condition) are significantly higher than the mutual information between the behaviors from subjects engaged in separate interactions (i.e., unpaired condition). The analysis also reveals that the information provided by modalities from one subject are complementary to the behaviors displayed by the other subject.

Motivated by the entrainment analysis, this study proposes to exploit cross-modality, cross-speaker information to improve the performance of an emo-

tion recognition system. The existence of this cross-modality entrainment suggests that the cues from one subjects can be used to obtain robust predictions of the emotional state of the other dialog partner. In this context, we are particularly interested in recognizing the emotional reactions of listeners. In these cases, only their facial expressions are available (assuming that a camera captures their faces without occlusion and with good illumination conditions). Therefore, having another complementary source of information can be very valuable (e.g., acoustic features from speakers). Notice that this task is related to the problem of monitoring the emotional reaction of users consuming multimedia content. Given the accelerated growth of social media and ubiquitous mobile devices, this problem is important.

To assess the benefit of utilizing the discussed mutual influence, several emotion classification experiments are conducted to recognize the expressive reactions from speakers and listeners. The first set of experiments consider the IEMOCAP corpus. We demonstrate that the emotion recognition accuracy of one subject improves when the emotion of the other dialog partner is known. Then, we implement cross-subject, cross-modality classification experiments, in which we recognize the emotion of the listener or speaker using features from both dialog partners. According to the large sample test of hypothesis about a population proportion, these classifiers achieve statistically significant improvements in performance over a classifier trained with only features estimated from the target subject. The accuracy and F-score in recognizing the listener's emotion increase by 5% (8.1% relative) and 7.5% (14.4% relative), respectively. Similarly, the accuracy and F-score of recognizing the speaker's emotion increase by 8.3% (15.4% relative) and 8.6% (16.6% relative), respectively.

The second set of experiments validates the proposed approach using the SEMAINE corpus, which comprises non-acted recordings using video cameras and microphones. These experiments consider classification tasks after clustering the activation-valence space from the original primitive-based evaluations. For the listener's emotion recognition problem, incorporating the dialog partner's facial and vocal cues improves the accuracy and F-score up to 8.1% (14.7% relative) and 8.1% (14.6% relative), respectively. Similarly, the facial expressions of the dialog partner enhances the accuracy and F-score of the predicted speaker's emotion up to 14.8% (29.4% relative) and 16.2% (33.3% relative), respectively. These results represent statistically significant improvement in performance over the systems trained with only features from the target subject. The evaluations support the advantage of exploiting cross-subject, cross-modality emotional entrainment in recognizing emotions.

The paper is organized as follows. Section 2 discusses related work on entrainment, especially in the context of emotions. Section 3 introduces the database, features and preprocessing steps. Section 4 describes the analysis on cross-modality emotional entrainment. The findings of the analysis motivate the classification experiments to recognize emotional states of both listeners and speakers, which are presented in Section 5. Section 5 also validates the benefits of using cross-subject, cross-modality features in dyadic recordings on the SEMAINE corpus. Section 6 discusses the findings and future directions of this study.

2 Related Works

2.1 Entrainment in Human-Machine Interaction

During spontaneous conversation, individuals tend to externalize similar verbal and nonverbal behaviors to promote effective communication (i.e., synchronization). In communication sciences, this effect is referred to as the reciprocity pattern [12]. Giles et al. [13] reported that during dyadic interactions the participants usually express similar nonverbal behaviors. Hinde [14] describes this phenomenon as analogous behaviors, which are expressed simultaneously or alternately during an interaction. This synchronization/regulation effect is known as entrainment. It is defined as becoming more similar to the dialog partner during the course of the interaction [15]. For instance, individuals tend to use similar terms, which suggests the existence of entrainment in their lexical choices [1].

The entrainment effect has been reported in different acoustic and prosodic features, including intensity [4], [9], fundamental frequency (F0) [5], [9], voice quality [9], duration and response latency [5], and speaking rate [3]. It is also observed in gestural behaviors including head motion [6], body gestures [7] and postures [8]. For example, Levitan et al. [9] proposed to compare the similarity in behaviors displayed by subjects engaged in a conversation (paired condition), with the similarity in behaviors displayed by subjects participating in separate conversations (unpaired condition). They study the entrainment effect in the speech preceding backchannels. They considered the number of common cues as a measure of similarity. These cues are defined in terms of intonation, intensity, fundamental frequency, duration and voice quality. They reported significantly higher number of common cues between the subjects interacting with each other.

The entrainment effect is observed not only during human-human interaction, but also in human machine/robot interaction [6], [10], [16], [17]. Breazeal [6] reported mutual regulation and entrainment in human robot interaction. They noticed adaptation effects in the body posture, head tilt and facial expression during the interactions. Bell et al. [10] reported that the speech rate of the users can be adapted using virtual characters, which was useful for ASR systems.

Ido et al. [16] studied lexical entrainment effect during human-robot interaction. They designed a robot to identify objects signaled by users through speech and gestures. They showed that the robot's confirmation statements can bias the users to employ easily recognizable terms, improving the speech recognition accuracy. Kanda et al. [17] developed a humanoid robot. They observed that users interacting with the robot made eye contact and imitated its gestures. They use these results to demonstrate the communication capabilities of the robot. These studies suggest that understanding the entrainment effect is important to improve the performance and efficiency of human machine interfaces.

2.2 Entrainment in Emotional Behaviors

Since acoustic and facial expressions are communicative channels to signal the emotions of the speakers [18], [19], we expect to observe the same adaptation pattern in emotions. Lee et al. [20] proposed the square of the correlation coefficient, mutual information and mean coherence as three measures to quantify entrainment in the context of emotional behaviors. They showed that the proposed measures provide discriminative information to classify between negative/positive conversations, due to the intrinsic higher level of entrainment during positive interactions.

The communication accommodation theory describes two types of adaptation behaviors [13]. The first type is convergence, which is defined as becoming similar to the conversational partner, in terms of communicative behaviors. The mimicking or mirroring behaviors observed during entrainment falls into this category [15]. The second pattern is divergence, which is defined as accentuating the differences in communicative behaviors.

The presence of entrainment effect in expressive behaviors displayed by conversational partners suggests that their mutual influences can be utilized to obtain more reliable assessments of their emotional states. We explored these ideas in our previous work [21]. We proposed a *dynamic Bayesian network* (DBN) to capture the mutual influence of the emotions between individuals during dyadic interactions (the emotions of one subject was conditioned on the predicted emotions of the other subject). Using only acoustic features, the study demonstrated the benefits of explicitly modeling the mutual emotional influence between speakers. Metallinou et al. [22] showed that the estimated emotions of the dialog interlocutor can improve the speaker emotion recognition, only when the interlocutor's vocal and facial cues are both available.

This study analyzes cross-modality entrainment (e.g., facial expression and acoustic features) using mutual information framework. Motivated by the findings, we propose novel cross-modality, cross-speaker emotion recognition experiments that improve the performance over baseline systems. To our knowledge, these directions have not been explored by other groups and represent important advancements in the area of multiparty emotion recognition.

Fig. 1. IEMOCAP data collection setting to capture spontaneous face-to-face interactions.

3 Corpus, Features and Preprocessing Steps

The entrainment analysis relies on the *interactive emotional dyadic motion capture* (IEMOCAP) database [23]. This section describes the corpus (Sec. 3.1), the emotional annotation (Sec. 3.2) and the facial and acoustic features (Sec. 3.3).

3.1 IEMOCAP database

The IEMOCAP corpus is an audiovisual database designed to study expressive human interactions [23]. It comprises five sessions of spontaneous conversation between professional actors (10 participants). In each session, an actor and an actress were asked to play three scripts and improvise eight hypothetical scenarios (e.g., getting married). The scripts and improvisation scenarios are carefully selected to elicit spontaneous emotional reactions [24]. These acting techniques are rooted in their theatrical training, producing realistic emotions evoked as a results of the interactions.

A VICON motion capture system is used to track markers attached to the face (53 markers), head (2 markers), and hands (6 markers) of the actors (Fig. 1). The placement of the facial markers followed the position of *feature points* (FPs) defined in the MPEG-4 standard, in most of the cases (Fig. 2(a)). The motion capture system provides detailed facial information at 120 frames per second. In each session, only one actor had markers to avoid interference between two separate setups (see Fig. 1). After collecting the script and improvisation recordings for one actor, the markers were placed on the other actor and the sessions were repeated. The audio is captured with two directional

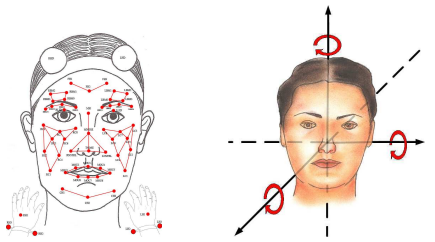


Fig. 2. (a) The IEMOCAP markers layout (53 facial markers). (b) 3D head rotation (pitch, roll and yaw).

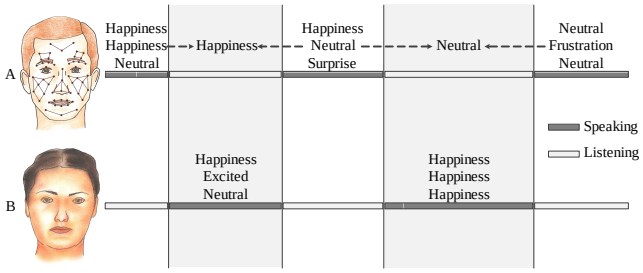


Fig. 3. Turn segmentation of the dialogs in the IEMOCAP corpus. For the turns in which the actor with markers (i.e., subject A) is not speaking (highlighted in gray), the emotion is interpolated using the emotional evaluations of adjacent segments.

shotgun microphones placed in the direction of the actors. The corpus comprises 12 hours of data.

3.2 Segmentation and Emotional Annotation

The data is transcribed and manually segmented into dialog turns. Six annotators were asked to assess the emotional contents of the actors during their speaking turns. The selected labels include happiness, anger, sadness, neutral, frustration, surprise, excited, fear and other. The subjective evaluation was conducted such that each turn was separately annotated by three evaluators. Notice that the emotions were elicited as dictated by the dialog, resulting in realistic, natural behaviors full of ambiguous, mixed emotions [25]. To be consistent with other studies using this corpus [26], [27], we consider only the most frequent emotional classes: happiness, anger, sadness, and neutral. Furthermore, happiness and excited are merged into a single class.

For a given turn, we are interested in studying the emotional states of both the speaker and the listener. A challenge associated with this goal is that the emotional evaluations were given to the turns in which the actors were speaking – the labels for the listeners' emotions are not directly available. This problem is described in Figure 3, which depicts a conversation sequence between subject A (with markers) and subject B (without markers). The dashed blocks in both tracks represent the segments in which the actors spoke (see legend in Fig. 3). These are the turns that were emotionally annotated. Even though the affective state of a subject can change in a short period of time, we approximate the emotions of subject A when he/she is listening with the emotional evaluations derived from his/her previous and following speaking turns (i.e., emotion interpolation). We consider all the emotional labels assigned by the evaluators to the surrounding turns – not just the consensus labels associated to these turns. Then, we assign the majority vote among these two sets as the emotional state of the listening segments. Consider the first listening turn of subject A in Figure 3. His/her previous turn received the labels *happiness* (2) and *neutral*, and his/her following turn received the labels *happiness*, *neutral* and *surprise*. We collect all these labels from the adjacent turns – i.e. *happiness* (3), *neutral* (2) and *surprise* (1) – and we assign the majority vote label which in this case is *happiness*. Similarly, the label for the second listening turn for subject A is *neutral* (3 out of 6).

We validate the emotion interpolation approach by comparing its agreement with the labels assigned by human evaluators. We asked three subjects to annotate the emotional content of 139 listeners' turns extracted from six dialogs across the actors (three improvisation, three scripts). We follow the same setting used during the original annotation of the corpus (e.g., 10 emotional classes, use of Anvil, sequential annotation of the dialogs – see [23]). The emotional labels are assigned to the turns according to the majority vote rule (in case of ties we use soft assignments). Then, we compared the agreement between the labels assigned by the interpolation and perceptual evaluators. We estimated the Cohen's kappa coefficient, achieving κ= 0.36 for 10 emotional categories. When we reduced the classes by merging happiness and excited, and relabeling the emotional classes with few samples as *other* (i.e., fear, disgust and surprise), the Cohen's kappa coefficient was κ= 0.44. For comparison, we estimated the Cohen's kappa coefficient between evaluators. The average results are κ=0.33 for 10 classes, and κ=0.38 for the reduced emotional set. This experiment reveals that the labels assigned by the interpolation approach are as consistent as the ones assigned by perceptual evaluations.

Given the recording setting of the corpus, we only consider the segments when the subject with markers (i.e., subject A) is listening – highlighted in gray in Figure 3. For these turns, facial marker information is available for subject A, and speech is available for subject B. For the other turns, we have facial information and speech for subject A, but no information for subject B (i.e., no speech, no facial markers). Hence, the segments when subject B is listening are not considered in the experiments, and their emotions are not interpolated.

Our previous studies showed high influence of spoken message on the variabilities of facial features

TABLE 1
Distribution of the emotional labels assigned to the actors' listening and speaking turns (portion of the IEMOCAP corpus).

emotion	happiness	anger	sadness	neutral	all
Listener	577	201	274	200	1252
Speaker	541	159	222	330	1252

in the mouth and jaw areas [26], [28]–[30]. To avoid capturing the anticipatory effect of articulation, the initial and ending 300 milliseconds of the listener's facial expressions are discarded. Also, the experiments do not consider the segments shorter than 500 milliseconds. These constraints limit the number of turns considered in this study (1252 turns). Table 1 shows the number of samples in each emotional class for both speakers and listeners using the aforementioned portion of the database.

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 [33].

Spectral LLDs
RASTA-style filtered auditory spectrum bands 1-26 (0-8kHz)
MFCCs 1-12
Spectral energy 25-650Hz, 1k-4kHz
Spectral roll-off point 0.25, 0.50, 0.75, 0.90
Spectral flux, entropy, variance, skewness, kurtosis, slope
Energy related LLDs
Sum of auditory spectrum (loudness)
Sum of RASTA-style filtered auditory spectrum
RMS Energy
Zero-crossing rate
Voice LLDs
F0
Probability of voicing
Jitter (local, delta)
Shimmer (local)

3.3 Facial and Acoustic Features

Facial features are extracted from the markers' information. First, the markers are translated and rotated using an approach based on *singular value decomposition* (SVD) described by Busso et al. [23]. After compensating for rotation and translation, the remaining movements of the facial markers correspond to facial expressions. The study uses as features the three dimensional location of the 53 facial markers and the head rotation parameters (i.e., pitch, roll and yaw). Figure 2(a) depicts the markers layout used to collect the motion capture data. Figure 2(b) shows the head rotation angles.

Given the differences in facial structure across actors and the variability in the actual placement of the reflective markers, it is crucial to normalize the facial features. For this purpose, we followed the facial normalization scheme proposed in our previous work [26]. The proposed approach adjusts the mean and standard deviation of the markers of each actor to match the ones of a reference actor. The female speaker in the first session is selected as the reference speaker. For each subject s, the marker m in direction $d \in \{X, Y, Z\}$ is mapped into the marker space of the reference subject (ref). Equation 1 gives the transformation, where μ and σ are the mean and standard deviation of the markers, respectively.

$$\acute{m_d^s} = (m_d^s - \mu_d^s) \times \frac{\sigma_d^{ref}}{\sigma_d^s} + \mu_d^{ref} \quad (1)$$

For each turn, seven high level statistics are extracted from the facial features: minimum, maximum, standard deviation, mean, median, lower quartile and upper quartile. Altogether, we create a 1,134 dimension feature vector for each turn ([53 markers $\times$ 3 dimensions + 3 head Euler angles] $\times$ 7 statistics). Due to the high dimension of this feature set, we used *correlation feature selection* (CFS) [31] criterion to reduce its dimension for analysis section (Sec. 4). This technique extracts a set of features having high correlation with the emotional labels, but low correlation between themselves. We used WEKA's best first search implementation to perform the selection [32]. This forward feature selection method is based on greedy hill-climbing approach, equipped with backtracking capability. The method sequentially expands the feature subset by adding a single feature. The subset is evaluated using the correlation criterion. If the path being explored does not improve in five consecutive steps, the previous subsets are considered for a different expansion path. Notice that this greedy feature selection approach is not a wrapper-based method depending on a particular classifier. The final feature set has 125 facial features.

This study uses the exhaustive acoustic feature set proposed for the Interspeech 2011 speaker state challenge [33]. This feature set comprises of sentence level functionals extracted from a set of frame-level features. Table 2 summarizes the frame-level features, referred to as *low level descriptors* (LLDs). The table presents spectral LLDs, energy related LLDs, and voice LLDs. The spectral feature comprises of RASTA-style filtered auditory spectrum, *Mel frequency cepstral coefficients* (MFCCs), and a set of statistics extracted at frame level, across spectral components. The statistics include energy in low band (25-650Hz) and high band (1k-4kHz), multiple roll-off points, flux, entropy, variance, skewness, kurtosis and slope. Spectral components are estimated with the short time *discrete Fourier transform* (DFT) amplitudes. The RASTA-style filtered auditory spectrum are estimated using the following steps: first, the *Mel filter bank* (MFB) are applied to the spectral components; then, the outputs are temporally filtered to remove non-speech components

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

(i.e., RASTA filtering); finally, equal loudness curve and loudness compression are applied to simulate the human auditory perception [34]. The extraction of the auditory spectrum includes all these steps, except the temporal filtering of the MFB coefficients. The X roll-off point is the frequency below which the signal energy drops the $X * 100\%$ of total signal energy. The energy related features include sum of auditory components before and after RASTA filters, *root mean square* (RMS) and zero-crossing rate. The voice LLDs include the fundamental frequency (F0), probability of voicing, jitter and shimmer.

Table 3 gives the set of sentence-level functionals, including 33 base functionals and 6 F0 functionals. Altogether, we estimate a 4368 dimensional feature vector from speech, which are extracted with the openSMILE toolkit [35]. A detailed description of the features can be found in Schuller et al. [33]. Similar to facial features, we implement CFS on the acoustic feature set, using the entire corpus, reducing its dimension to 210. We use this feature set for Section 4. The acoustic features are also normalized across the ten speakers in the database using Equation 1.

4 Cross-Modality Emotional Entrainment

This section studies cross-modality emotional entrainment and its effects on the acoustic and facial cues displayed by subjects during dyadic conversations. Previous studies on entrainment have proposed different metrics to study entrainment such as the number of common cues [9], absolute distance [3], correlation and *mutual information* (MI) [20]. Following the work of Lee et al. [20], this study uses mutual information. The proposed approach analyzes the mutual informations between behaviors observed across modalities and across subjects (e.g., facial expression of the listener versus the acoustic features of the speaker). We are also interested in analyzing the relation between modalities of one subject and the emotions of the other (e.g., acoustic features of the speaker versus the emotional state of the listener). Since we are studying the relation between heterogeneous modalities, we can not directly compare the similarities with metrics such as distance or correlation. Instead, we use mutual information to quantify the dependencies rather than similarities between modalities, which is a major difference between this study and previous works.

Equation 2 gives the mutual information for discrete variables X and Y, given their marginal and joint *probability mass functions* (PMFs). Facial and acoustic features provide continuous values. Therefore, we discretize the features using the K-means algorithm. Given the differences in the range across features, we apply z-normalization before estimating the clusters. The PMFs are estimated from the data.

$$I[X;Y] = \sum_{x\in X, y\in Y} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} \quad (2)$$

$$I[X;Y] = H[X] - H[X|Y] \quad (3)$$

$$H[X] = -\sum_{x\in X} p(x)\log p(x) \quad (4)$$

Levitan and Hirschberg [3] proposed to compare the similarity in behaviors between individuals during their interactions (paired condition), with the similarity in behaviors between individuals engaged in different conversations (unpaired condition). The proposed analysis follows a similar approach by comparing the mutual information in paired and unpaired conditions. Note that the unpaired conditions corresponds to randomly permuting emotional labels or acoustic/facial features from different turns, depending on the analysis (e.g., randomly pairing speaker's emotion with listener's emotion from different turns – see Fig. 4(a)). Figure 4 summarizes the four parts of the analysis, which will be described next. The nodes EMO_L and EMO_S are the emotional states of the listener and speaker, respectively. The node F_L describes the facial features of the listener. The node V_S represents the features from the speaker's voice.

4.1 Emotion Entrainment – Fig. 4(a)

According to the interpersonal adaptation theory, conversational partners tend to converge in the behaviors showing reciprocal and mirroring patterns. The exception occurs when the subjects decide to diverge in their behaviors to cope with a given situation [15]. Therefore, it is expected to observe similar emotional behaviors across dialog partners during spontaneous

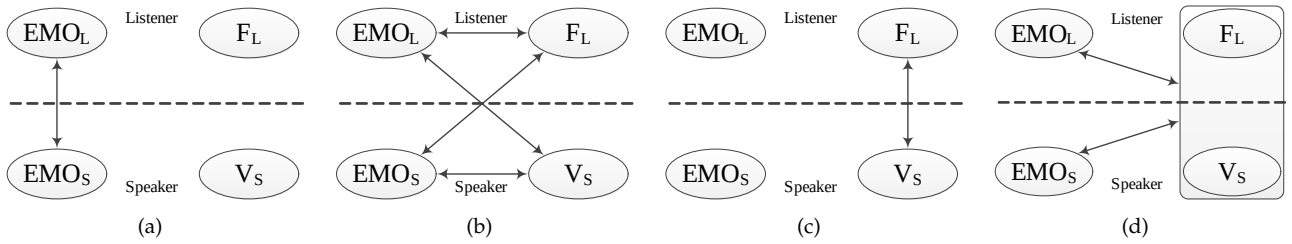


Fig. 4. The four aspects in dyadic interactions considered in the analysis. (a) dependency between the emotional states of the dialog partners (Sec. 4.1), (b) dependency between the emotion of one subject and the expressive behaviors of the other (Sec. 4.2), (c) dependency between heterogeneous behaviors from the dialog partners (Sec. 4.3), (d) effect of cross-subject multimodal information for emotion discrimination (Sec. 4.4). EMO_L: listener's emotion, EMO_S: speaker's emotion, F_L: listener's facial features, V_S: speaker's voice.

TABLE 4
Co-occurrence between the emotions displayed by speakers and listeners in the turns during spontaneous dialogs (IEMOCAP corpus) (Ang: anger, Hap: happiness, Sad: sadness and Neu: neutral)

		Listener			
		Ang	Hap	Sad	Neu
Speaker	Ang	**488**	3	11	39
	Hap	4	**113**	13	29
	Sad	4	2	**192**	24
	Neu	81	83	58	**108**

interactions. To investigate this hypothesis, we estimate the co-occurrence in the emotional labels assigned to listeners and speakers for each turn (see Fig. 4(a)). Table 4 shows that 72% of the time both conversation partners share the same emotion. This result supports the emotional adaptation hypothesis. If we account for the marginal distribution of the speaker's and listener's emotions (Table 1), and assuming their independence, the expected ratio of observing similar emotions by chance is 30%. Also, if we randomly pair the emotional labels of the subject s100 times, (i.e., sampling individual distributions), we observe 30.1% matching, on average. The large sample hypothesis test about a population proportion shows statistically significant differences ($p - values << 1e - 20$).

Table 4 shows that the co-occurrence of emotions between dialog partners decreases when one of them is in neutral state. Notice that neutral state is not always well defined and it is often confused with other emotions [36]. The table also shows that cases in which one subject displayed a non-neutral emotion (e.g., anger), and the other displayed a different non-neutral emotion (e.g., sadness) are uncommon 3%).

We estimate the entropy of the speaker's and listener's emotions, given the distributions provided in Table 1 (see Eq. 4). Their entropies are $H[EMO_S] = 1.85$ bits and $H[EMO_L] = 1.84$ bits, respectively. The mutual information between these two variables is 0.8 bits. Hence, the knowledge of the emotion from one subject provides important information about the

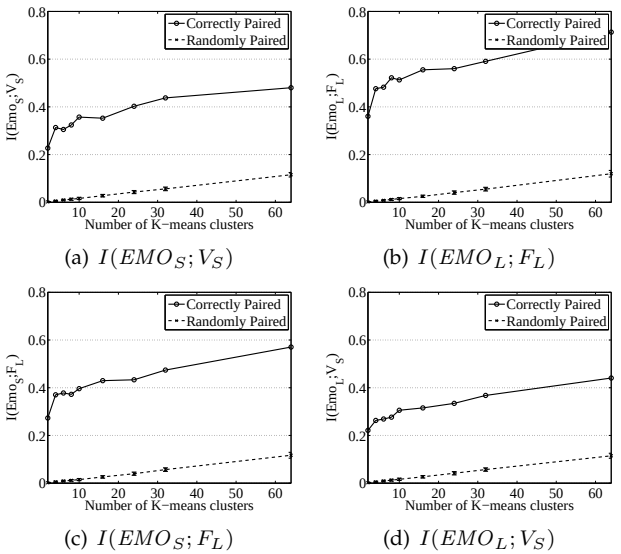


Fig. 5. Mutual information analysis of speaker's and listener's emotions with the speaker's voice and listener's facial expressions (IEMOCAP). (a) speaker's emotion and speaker's voice, (b) listener's emotion and listener's face, (c) speaker's emotion and listener's face, (d) listener's emotion and speaker's voice. Lines show correctly (—) and randomly (- - -) paired turns.

emotion of the other subject (see also the classification results in Sec. 5). For comparison, we estimate the mutual information between the emotions of speakers and listeners from different turns (emotional labels of the 1252 turns are randomly paired 100 times). The average mutual information for the unpaired condition is 0.005, which is significantly lower than the mutual information for the paired case ($p - values < 1e - 20$). These findings clearly confirm the effect of entrainment at the emotion level.

4.2 Cross-Subject Relation of Emotion and Modalities–Fig.4(b)

Given the aforementioned emotional synchronization patterns, we hypothesize that facial gestures of the listeners provide complementary information about

the speakers' emotions, and that the acoustic features of the speakers provide information about the emotion of the listeners (i.e., diagonal arrows in Fig. 4(b)). This cross-subject emotional entrainment is studied with mutual information (i.e., $I(EMO_L; V_S)$, and $I(EMO_S; F_L)$). As a reference, we also report the mutual information of the subjects' emotions and their corresponding acoustic/facial features (i.e., $I(EMO_L; F_L)$, and $I(EMO_S; V_S)$ (i.e., horizontal arrows in Fig. 4(b)). The unpaired conditions in the reference experiments correspond to randomly pairing emotions of each subject with his/her expressive cues. The PMFs are estimated using different number of bins during the K-means algorithm. Notice that as we increase the number of clusters (K), fewer samples are assigned to the clusters. In the extreme, the distribution of the samples tends to the uniform distribution, artificially maximizing the entropy (Eq. 4). This case yields a one-to-one mapping between samples of the two variables, which reduces the conditional entropy, $H[X|Y]$, to zero (see Eq. 3). Therefore, increasing the number of clusters intrinsically increases the mutual information. With 1252 turns, the maximum number of clusters was set to $K_{\max}$ = 64.

Figure 5(a) shows the mutual information between the speaker's voice and speaker's emotion. Figure 5(b) shows the corresponding values between the listener's face and listener's emotions. These values (solid lines) are compared against the mutual informations between emotions and acoustic/facial features from randomly paired turns (unpaired condition - dashed line). These values are the average results over randomly pairing the 1252 turns, 100 times. These figures show strong connection between the features from a subject and his/her emotions, which validates various studies showing the value of using acoustic and facial features for recognizing emotions [18], [27].

Figures 5(c) and 5(d) show the mutual information in cross-subject settings. Notice that in these two figures the dashed line gives the mutual information in unpaired conditions following the aforementioned approach (i.e., emotion from one subject and acoustic/facial features from the other subject in the randomly paired turns). Figure 5(c) suggests that the facial expression of the listener provides valuable cues to describe the speaker's emotions. Likewise, Figure 5(d) suggests that the speaker's voice provides discriminative information to distinguish the listener's emotions. These results are significantly higher than the corresponding values for unpaired conditions. Section 4.4 demonstrates that the cross-subject information is complementary to the subject's own cues.

4.3 Cross-Modality, Cross-Subject Entrainment – Fig. 4(c)

This section directly studies the mutual information between the speaker's voice and the listener's facial expressions (i.e., $I(F_L; V_S)$ – Fig. 4(c)). The analysis measures the mutual information from cross-modality features extracted from both paired turns, $I^{paired}(F_L; V_S)$, and the average achieved by randomly pairing all the samples 100 times, $I^{unpaired}(F_L; V_S)$. We estimate the difference between these values, as described in Equation 5.

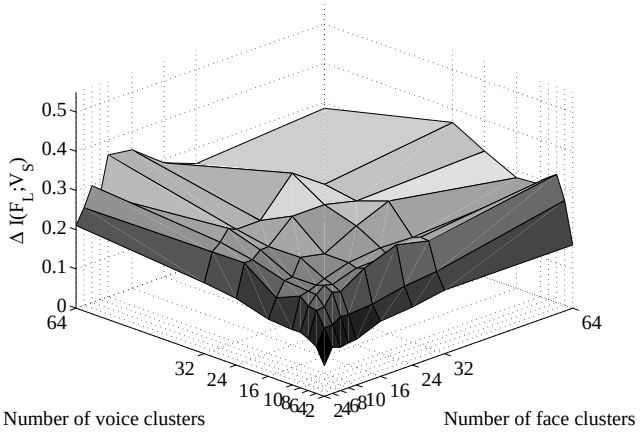


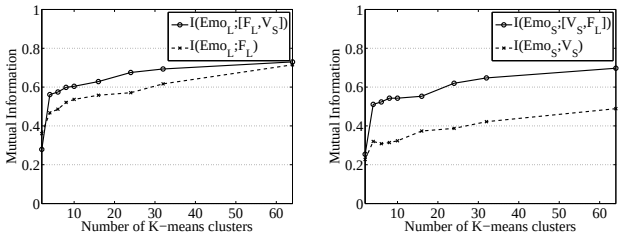
Fig. 6. Difference between mutual information of speaker's voice and listener's face in correctly and randomly paired turns – IEMOCAP corpus ($\Delta I(F_L; V_S)$).

$$\Delta I(F_L; V_S) = I^{paired}(F_L; V_S) - I^{unpaired}(F_L; V_S) \quad (5)$$

Different number of bins are used for facial and acoustic features. Figure 6 shows the results. The difference in mutual information between these two conditions is consistently positive across different number of bins. Therefore, the mutual information in the paired condition is always greater than the average values for the unpaired conditions. This analysis highlights the important coupling between cross-modality features extracted from different subjects. We believe that this finding is attributed to emotional entrainment. Conversational partners tend to display similar emotions (Sec. 4.1), producing expressive cues that are characteristic of the given emotional state. The emotions are manifested in both subjects across their modalities including facial expressions [19] and acoustic features [18], producing coupled behaviors across subjects.

4.4 Complementariness of Cross-Subject Behaviors – Fig.4(d)

The previous results highlight the connection between nonverbal behaviors of one subject and the emotions displayed by the other subject. An important question is to determine whether the cross-subject behaviors are complementary to or redundant with the own behaviors displayed by the subject. To address this question, we compare the mutual information in single modality setting with the mutual information in cross-subject multi-modality setting.



(a) Listener (b) Speaker

Fig. 7. Complementary nature of cross-subject behaviors in the IEMOCAP database. The figure compares the mutual information between a subject's emotion and his/her behaviors (dashed line), with the mutual information between a subject's emotion and a feature set that include cross-subject behaviors (solid line).

Figure 7(a) compares the mutual information between the listener's emotion and the listener's face, $I(EMO_L; F_L)$, with the mutual information between the listener's emotion and the multimodal information provided by the listener's face and speaker's voice, $I(EMO_L; [F_L, V_S])$. To build the distribution of this multimodal data, the listener's facial features and speaker's acoustic features are concatenated into a single vector before performing the K-means algorithm. Likewise, Figure 7(b) compares the mutual information between the speaker's emotion and speaker's voice, $I(EMO_S; V_S)$, with the mutual information between speaker's emotion and the cross-subject, cross-modality features, $I(EMO_S; [V_S, F_L])$. Both figures show an increase in mutual information in the cross-subject multimodal settings (solid lines). These results indicate that cross subject behaviors provide complementary information about the displayed emotion during dyadic interactions. Section 5 validates these results in emotion recognition experiments.

In summary, the results show that the emotion of one subject is related with the behaviors of the other subject. Furthermore, the cross-subject cues provide complementary information. These results have direct impact on multimodal emotion recognition problem, which is demonstrated in Section 5.

5 Cross-Subject Multimodal Emotion Recognition

This section explores the insights from the analysis in multimodal emotion recognition evaluations. We conduct the experiments on the IEMOCAP corpus (Sec. 3.1). The use of motion capture markers to represent facial cues is not practical in many emotion recognition applications. Therefore, we also consider the SEMAINE database – a non-acted multimodal emotional corpus (Sec. 5.2). For this database, the facial features are directly estimated from video recordings.

5.1 Results on the IEMOCAP Database

The evaluation assesses the improvement in emotion recognition performance when we consider cross-subject multimodal information. We separately consider both speaker's emotions and listener's emotions recognition tasks. The experiments are conducted using leave-one-speaker-out cross-validation (speaker independent training/testing partitions). For each of the 10 folds, CFS is used to select facial and acoustic features for the classification problems using only the training set. Therefore, we have 10 feature sets, with an average of 115 facial ($\sigma = 9$), and 197 acoustic ($\sigma = 11$) features. The evaluation uses linear kernel *support vector machine* (SVM) with *sequential minimal optimization* (SMO). The soft margin parameter c is selected by optimizing the baseline classifiers: SVM$_L$ (F_L) that recognizes the listener's emotions using his/her facial features; and, SVM$_S$ (V_S) that recognizes the speaker's emotions using his/her acoustic features. For each of the 10 folds, this parameter optimization is conducted exclusively on the training set (9 subjects). We evaluate different values of c (i.e., 0.001, 0.01, 0.1 and 1) by splitting the training set to build the classifiers (8 subjects) and to validate the results (1 subject). We implement all possible permutations of subjects across folds. In 85% of the cases, the best performance was obtained with c=0.1. For consistency, this value is used for the entire evaluation experiments. By not using the testing set for feature selection and parameter optimization, the reported results are accurate and unbiased.

Given that the data is not emotionally balanced (see Table 1), we estimate the precision rate for each emotional class (i.e., fraction of retrieved samples for one emotional class that are relevant). Then, we estimate and report the *average precision* (P) across classes. Likewise, we estimate the recall rate for each emotional class (i.e., fraction of relevant samples that are correctly classified). We report the *average recall* (R) across classes. With these values, we calculate the *F-score* (F) using Equation 6. In addition, we report the *accuracy* (A) of the classifiers.

$$F = \frac{2PR}{P+R} \tag{6}$$

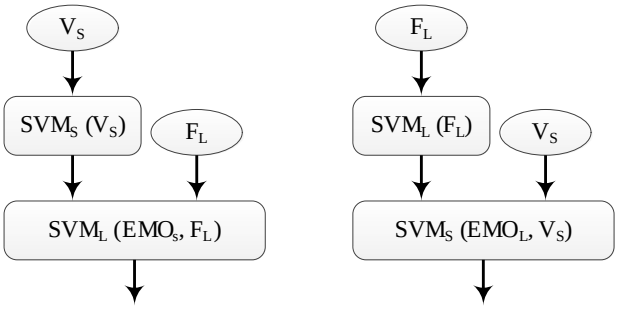
5.1.1 Recognition of the Listener's Emotion

Table 5 reports the results of the listener's emotion classification task under different conditions. The first row shows the baseline classifier, which is trained with only the facial features extracted from the listeners – SVM$_L$ (F_L). The average recall is 52.1%, which is slightly lower than the average recall reported in a previous study on facial expression that used a larger portion of this corpus [36] (see explanation in Sec. 3.2 on the reduced number of turns considered here).

We illustrate the emotional adaptation effect by recognizing the listeners' emotions using only the

TABLE 5
Results of emotion recognition of the listeners' turns for different settings (IEMOCAP corpus). The results are given in terms of Accuracy (A), Precision (P), Recall (R), and F-Score (F) (F_L: listener's face, EMO_S: speaker's emotion, V_S: speaker's voice).

Method	A	P	R	F
SVM$_L$ (F_L) [baseline]	62.30	52.01	52.10	52.05
SVM$_L$ (EMO_S)	70.21	67.16	62.81	64.91
SVM$_L$ (F_L, EMO_S)	72.28	65.43	64.34	64.88
Cascade SVM$_L$ (F_L, V_S)	66.21	57.16	57.47	57.31
SVM$_L$ (V_S)	55.03	45.93	45.16	45.54
SVM$_L$ (F_L, V_S)	**67.33**	**59.28**	**59.79**	**59.53**



(a) Listener's emotion (b) Speaker's emotion

Fig. 8. Cross-subject emotion recognition with cascade SVMs. The dialog partner's emotion is used as feature to recognize the target subject's emotion.

speakers' emotions – SVM$_L$ (EMO_S). Table 5 shows that this classifier achieves an accuracy of 70.2%. Notice that this condition outperforms the baseline without using any feature describing the listeners' behaviors. When a classifier is trained with the speakers' emotion and the facial features from the listeners, SVM$_L$ (F_L, EMO_S), the classifier achieves the best accuracy (72.3%) and F-score (64.9%) rates. These results highlight the importance of considering the emotional state of the dialog partners, as discussed in Section 4.1.

In many real applications, the speakers' emotion is not available and needs to be estimated. Following this direction, we consider both explicit and implicit modeling of the speaker's emotions to recognize the listener's emotions. We propose a cascade SVM in which we explicitly estimate the speaker's emotion using his/her acoustic features (see Fig. 8(a)). The output of this classifier and the facial features from the listeners are used as input to recognize the listener's emotion – Cascade SVM$_L$ (F_L, V_S). Table 5 shows that all the performance metrics for this configuration are higher than the ones achieved by the baseline classifier by at least 3.9% (absolute).

We also explore the case in which the speaker's emotion is implicitly incorporated in the classifiers by directly using the speaker's behaviors. First, we evaluate the performance of the classifier when we consider only features extracted from the speaker's voice – SVM$_L$ (V_S). This classifier achieves an accuracy of 55%, which is lower than the baseline classifier. However, the performance is significantly higher than chances (25%). This result demonstrates the discriminative power of the speaker's voice to distinguish the listener's emotion. It also supports the analysis presented in Section 4.2 (see Fig. 5(d)). Then, we train a classifier with heterogeneous features describing the speaker's voice and listener's faces – SVM$_L$ (F_L, V_S). This cross-modality, cross-subject classifier improves the baseline accuracy and F-score to 67.3% and 59.5%, respectively. A large sample hypothesis test about a population proportion indicates that the gain in F-score is statistically significant ($p - value < 0.0001$). The improvement in accuracy is also statistically significant ($p - value < 0.0042$). Although the speaker's emotion is unknown, the speaker's voice provides complementary information to recognize listener's emotion, which is consistent with the findings in the analysis section (see Fig. 7(a)).

TABLE 6
Average precision and recall of the classifiers for each emotional class (IEMOCAP corpus).

	Listener's emotion							
	precision (%)				recall (%)			
	Hap	Ang	Sad	Neu	Hap	Ang	Sad	Neu
SVM$_L$ (F_L) [baseline]	77.0	43.5	56.2	31.4	82.8	38.3	68.3	19.0
SVM$_L$ (F_L, V_S)	77.8	57.4	68.3	33.6	81.5	57.7	77.0	23.0

	Speaker's emotion							
	precision (%)				recall (%)			
	Hap	Ang	Sad	Neu	Hap	Ang	Sad	Neu
SVM$_S$ (V_S) [baseline]	57.2	61.4	54.1	39.1	71.5	44.0	63.1	24.0
SVM$_S$ (V_S, F_L)	72.8	62.5	61.1	44.9	73.2	56.6	69.4	42.4

Table 6 provides the precision and recall rates per emotion for the baseline, SVM$_L$ (F_L), and the best cross-speaker, cross-modality scheme, SVM$_L$ (F_L, V_S). Our previous study on facial emotion recognition showed that anger is often confused with sadness and happiness is often confused with neutral class [37]. The confusion between these pairs is reduced in the acoustic domain. Even though the speech is coming from the dialog partner, Table 6 indicates that the precision and recall rates improve when the speaker's voice is incorporated in the system. Therefore, the discrimination power of acoustic signal of the conversation partner reduces the confusion between these emotions. This result validates the complementary nature of cross-subject cues (Sec. 4.4).

5.1.2 Recognition of the Speaker's Emotion

We follow a similar approach to recognize the speaker's emotion. Table 7 reports the results. The baseline classifier is trained with features describing the speaker's voice – SVM$_S$ (V_S). Although the portion of the corpus used in the evaluation is different,

TABLE 7
Results of emotion recognition of the speakers' turns, for different settings (IEMOCAP corpus). The results are given in terms of Accuracy (A), Precision (P), Recall (R), and F-Score (F) (V_S: speaker's voice, EMO_L: listener's emotion, F_L: listener's face).

Method	A	P	R	F
SVM$_S$ (V_S) [baseline]	53.99	52.93	50.64	51.76
SVM$_S$ (EMO_L)	71.96	66.22	70.12	68.11
SVM$_S$ (V_S, EMO_L)	74.04	69.74	71.42	70.57
Cascade SVM$_S$ (V_S, F_L)	62.46	60.29	59.07	59.67
SVM$_S$ (F_L)	54.55	46.15	45.95	46.05
SVM$_S$ (V_S, F_L)	**62.30**	**60.32**	**60.40**	**60.36**

the average recall of our baseline (50.6%) is similar to the one reported in a previous study using only acoustic features (50.7%) [36].

When the listener's emotion is known, the speaker's emotion can be recognized with 72% accuracy – SVM$_S$ (EMO_L). When the speaker's voice and the listener' emotions are used, the classification accuracy improves to 74% – SVM$_S$ (V_S, EMO_L). When the listener's emotion is explicitly estimated using a cascade SVM$_S$ (V_S, F_L) (see Fig. 8(b)), we achieve a 62.5% accuracy. The improvement over the baseline for both metrics is over 7.9% (absolute), which is statistically significant, according to the proportion hypothesis test ($p - value < 1e - 10$). When we implicitly incorporate the listeners' emotion by adding features describing their facial expression, the classifier achieves 62.3% accuracy – SVM$_S$ (V_S, F_L). These results represent improvements over 8.3% (absolute), which are statistically significant ($p-value < 1e-10$). These results validate the relationship observed in the analysis between the speaker's emotion and listener's facial expressions (see Fig. 5(c)).

Similarly, Table 6 provides the precision and recall rates per emotion for SVM$_S$ (V_S) and SVM$_S$ (V_S, F_L). Our previous study showed high confusion in the acoustic domain between anger and happiness, and between sadness and neutral state [37], [38]. Table 6 shows that adding features describing the listener's facial expression increases the precision and recall rates of neutral state, happiness and sadness. The recall rate for anger is also increased. The complementary information of conversation partner in cross-modality settings can compensate for the intrinsic limitations observed in single modalities to discriminate between specific emotions (e.g., limitations of acoustic features to describe valence dimension [39]).

5.2 Results on the SEMAINE Database

This section validates the analysis on cross-subject, cross-modality affective entrainment in more natural recordings (i.e., non-acted corpus recorded with video cameras). For this purpose, we present emotion classification experiments using the *sustained emotionally colored machine-human interaction using nonverbal expression* (SEMAINE) database [40]. This multimodal corpus was collected using the *sensitive artificial learner* (SAL) technique [41] to engage users in emotional conversations with an operator. The operator can be a virtual character (i.e., semi-automatic SAL and automated SAL) or another human (i.e., solid SAL). This study only uses the solid SAL portion of the corpus, which provides spontaneous dialogs between two individuals. While the operator portrays a character with a specified mood, the users' reactions are purely non-acted. The user and operator sit in separate rooms, interacting through teleprompter screens. Their facial expressions and speech are simultaneously recorded. Therefore, unlike the IEMOCAP corpus (see Fig. 1), the SEMAINE corpus provides simultaneous recordings displaying the facial expressions from both subjects (we do not have the constraints described in Sec. 3.2).

This corpus provides frontal videos of the individuals' faces. The study relies on the *computer expression recognition toolbox* (CERT) [42] to extract facial features. CERT automatically extracts *action units* (AUs), defined in the *facial action coding system* (FACS) [43]. AUs describe the facial movements of individual muscles or groups of muscles. The toolkit processes the video frame-by-frame, providing high accuracy and robustness against different illumination conditions. Notice that the facial markers' layout in the IEMOCAP approximately follows the positions of the *feature points* (FPs) defined in the MPEG-4 standard for facial animation [44]. This standard also defines a set of *facial animation parameters* (FAPs) to modulate the facial appearance by moving the FPs. These FAPs are derived from the definition of the AUs. Therefore, there is a close relationship between the markers' trajectory – features on the IEMOCAP corpus – and the AUs – features on the SEMAINE corpus.

The classification experiments consider 20 AUs and 3 head rotation parameters provided by CERT (see Table 8). Similar to the approach used with the facial markers, we estimate seven statistics from these features at turn level (minimum, maximum, standard deviation, mean, median, lower quartile and upper quartile). Altogether, a turn is represented with a 161 dimensional facial feature vector (i.e., [20 AUs + 3 head rotation] $\times$ 7 statistics). Notice that these facial features are extracted from both the user's and operator's recordings. For acoustic features, we extract the same set described in Section 3.3 (Tables 2 and 3).

The user's emotional reactions are annotated in terms of activation (i.e., active versus passive) and valence (i.e., positive versus negative) dimensions, using the FEELTRACE toolkit [45] (other emotional attributes are also available). Instead of turn level assessments, this annotation scheme continuously captures the perceived emotional primitives values, as the annotators move the mouse cursor over a *graphical*

TABLE 8
The list of *action units* (AUs) extracted by CERT [42].

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

user interface (GUI) displaying the activation/valence space. For each dimension, the scores are mapped into the range [-1, +1]. The emotion annotations are performed by multiple labelers (2 to 8) over the entire sessions. Given that the focus of this corpus is on the user's reactions, there are few sessions in which the operator's videos are emotionally evaluated. Therefore, the classification experiments in this study consider only the user's emotions. The emotional labels include turns when the user is both speaking and listening. Therefore, this corpus is suitable for the proposed cross-subject, cross-modality evaluation.

Only 52 out of 94 currently released sessions have emotional labels. During eight of these sessions, the CERT toolkit did not correctly detect the user's face (sessions 82, 88, 89, 90, 91, 95, 96 and 97). Hence, this study considers interactions from 44 sessions. These sessions are split into turns using the provided segmentation. The dialog turns are manually segmented. We consider only turns which are at least 300 ms. For the turns when the user is listening, the initial and ending 100 milliseconds of the segments are discarded to avoid capturing articulation (provided that the remaining segment is at least 300 ms). Altogether, we consider 1884 turns, in which the user is listening in 835 segments and speaking in the remaining 1049 segments. The emotional ground truth for each of these turns is calculated by averaging the scores across evaluators and across frames (see plot in Fig. 10).

One drawback of using a continuous frame-by-frame evaluation toolkit such as FEELTRACE is the delay between the stimulus and the annotated labels. The delay is caused by the intrinsic reaction time between the perception of the expressive behaviors and the annotation of the stimuli (i.e., moving the cursor). Nicolle et al. [46] studied this delay on four emotion attributes (activation, valence, expectation and power) in the SEMAINE database using correlation analysis. They reported average delays between three to six seconds. Following a similar approach, we propose to estimate the optimal delay with the mutual information between the frame-level facial features (F) and the τ-sec-shifted emotional annotations (E_τ), $I(F;E_\tau)$ (Eq. 2). We rely only on facial features since the acoustic features are not always available during the course of

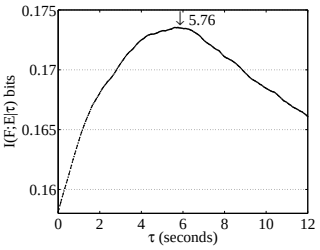
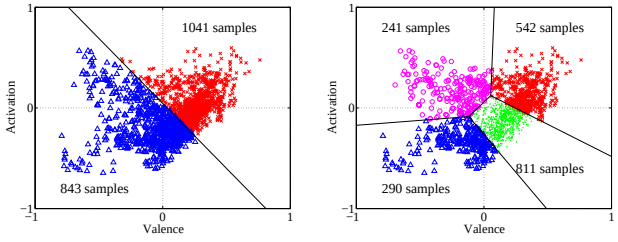


Fig. 9. Analysis of the delay between the emotion annotations and facial features for the SEMAINE database. The optimum delay is 5.76 seconds.



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

Fig. 10. Clusters obtained by the K-means algorithm in the valence-activation space (SEMAINE). The figure gives the number of turns assigned to each class.

an interaction. The PMFs for the emotion (activation-valence space) and facial features are estimated from the data using nonuniform bins created with the K-means algorithm. Figure 9 depicts the average mutual information for different delays achieved across different number of clusters. Maximizing this mutual information with respect to τ yields the optimal delay, which in our case is 5.76 seconds. The shape of the curve in Figure 9 and the optimal delay are consistent with the findings reported by Nicolle et al. [46]. Accordingly, the emotional evaluations are shifted in 5.76 seconds for the classification experiments.

Instead of dealing with continuous emotional attributes, we created K emotional clusters in the activation-valence space by using the K-means algorithm. Previous studies on affective computing have used this approach to transform a regression problem into a K-class recognition problem [21], [47]. We reported the classification results for K = 2 and K = 4. Figures 10(a) and 10(b) show the corresponding classes obtained on the entire database.

The classification experiments follow the settings described in Section 3.1 (i.e., SVM-SMO with c = 0.1). The selected portion of the database contains nine users. We train and test the classifiers using a leave-one-speaker-out cross-validation. The feature sets are reduced using CFS, using the training set of each fold. For K=2, CFS selects an average of 29 facial (σ=6) and 94 acoustic (σ=10) features. For K=4, CFS selects an average of 39 facial (σ=5) and 98 acoustic (σ=11) features. The feature selection is performed with au-

TABLE 9
Emotion classifications results on the SEMAINE corpus, for segments when the user is speaking and listening. Results are reported for K = 2 and K = 4 in terms of Accuracy (A), Precision (P), Recall (R), and F-Score (F).

User's state	Features		K = 2 [chance level = 50%]				K = 4 [chance level = 25%]			
	User	Operator	A	P	R	F	A	P	R	F
Listening	Face	-	55.36	55.51	55.53	55.52	**47.35**	30.06	30.30	30.18
	Face	Face	59.19	29.29	59.33	59.31	41.55	28.49	28.28	28.38
	Face	Voice	61.16	61.63	61.56	61.59	46.86	**33.56**	**30.34**	**31.87**
	Face	Face,Voice	**63.50**	**63.58**	**63.64**	**63.61**	43.28	30.23	29.77	30.00
Speaking	Face	-	53.62	52.54	52.50	52.52	47.17	35.13	34.38	34.75
	Voice	-	50.35	48.46	48.55	48.50	39.52	26.04	26.16	26.10
	Face, Voice	-	53.92	53.36	53.39	53.37	49.16	39.28	36.67	37.93
	Face	Face	64.95	64.41	64.49	64.45	51.44	42.58	38.32	40.34
	Voice	Face	**65.14**	**64.62**	**64.71**	**64.66**	**52.73**	**40.22**	**38.60**	**39.39**
	Face, Voice	Face	62.26	61.59	61.56	61.57	50.84	41.70	38.26	39.91

diovisual features from the users. However, the same feature set is estimated from the operator's behaviors. Table 9 reports the user's emotion classification experiments during the listening and speaking segments, considering the two emotional space clusterings.

5.2.1 Recognition of User's Emotion while Listening

For $K = 2$ (50% chance level), a classifier trained with only the user's facial expressions achieved an accuracy of 55.4% (turns when the user was listening). Incorporating features describing the operator's face, voice or both significantly improves the accuracy by at least 3.8%. The best performance is obtained when features describing the operator's voice and facial expression are added to the user's facial features ($A = 63.5\%$). This result represents statistically significance improvement over the performance of the classifier trained with only the facial features of the user ($p-value < 0.0001$ – population proportion test).

5.2.2 Recognition of User's Emotion while Speaking

During the segments when the user is speaking, the face is the only modality available for the operator. From the user, we extract his/her facial and acoustic features. Table 9 provides the performance for different combinations. There are three baseline classifiers for which only features from the user are used (i.e., face, voice or both modalities). The baseline classifiers trained with features describing the user's face achieve 53.6% and 47.2% accuracy rates for $K = 2$ and $K = 4$, respectively. The user's acoustic features do not provide significant discriminant information to recognize his/her emotion.

Table 9 indicates that adding cross-speaker information (i.e., the operator's face) improves the accuracy and F-score rates in all the settings, both for $K = 2$ and $K = 4$. When $K = 2$, the addition of features describing the operator's face yields statistically significant improvements for accuracy and F-score ($p-value < 0.0001$), across classifiers. The best performance is achieved when only acoustic cues of the user and facial expressions of the operator are employed ($A = 65.1\%$). For $K = 4$, the accuracy improves in the three cross-subject, multimodal settings. The best performance is achieved by incorporating user's voice and operator's facial expressions ($A = 52.7\%$). These results validate the benefits of using cross-subject features for multiparty emotion recognition.

6 Conclusions and Discussion

This paper analyzed cross-subject multimodal emotional entrainment and the implications on the design of emotion recognition systems for dyadic human interaction. We presented a thorough analysis to unveil the relation between the emotional states of dialog partners, and the relation between their expressive behaviors. The analysis reveals that most of the time the conversational partners present similar emotions (i.e., converging behaviors). Using mutual information as a metric of dependency, the study shows that the behaviors from one subject provide complementary information about the emotional state of the other subject. Motivated by these findings, we presented cross-subject multimodal emotion recognition experiments. We reported result on the IEMOCAP and SEMAINE databases. In both corpora, we consistently observed statistically significant improvements in the classifiers, when the feature set included features describing cross-subject behaviors.

We observe that the improvement in performance in recognizing the speaker's emotions is consistently higher than the one in recognizing the listener's emotions. This pattern is also observed in Figure 7, which shows that the listener's facial expression provides more complementary information about the speaker's emotion, than the speaker's voice provides about the listener's emotion. From a human communication perspective, this pattern should be studied further.

As mentioned, the findings of this work are relevant to the problem of monitoring the emotional reactions of users interacting with a device such as smart TV, cellphones, tablets, or computers. When the user consumes multimedia content, his/her facial expression is the only modality that is available to recognize his/her emotions. Studies have shown

traits of entrainment between human and machine interfaces (robots and avatars) [6], [10], [16], [17]. Therefore, we expect to observe similar cross-subject multimodal affective entrainment between the behaviors conveyed in multimodal content (i.e., movies and video-blog) and the user emotional reactions. We are studying the benefits of using the emotions conveyed in the multimodal content to recognize the emotional reactions of the user. For these practical applications, the speech and video streams should be automatically segmented and processed. We are investigating the use of fixed windows, which will minimize the need of pre-segmenting the data. This approach has been successfully used in speech emotion recognition [48].

Studies have shown that the adaptation patterns are context-dependent, since the dialog partners can display converging or diverging behaviors as dictated by the interaction [13]. We mostly observe converging behaviors in the IEMOCAP corpus. However, we expect to see diverging patterns in other scenarios. For instance, a representative working in a customer center should display behaviors that reduce the frustration or anger of the customers. While the emotions of the dialog partners may not be same, their behaviors are still related, which can provide complementary information even with diverging behaviors. We will explore these scenarios in our future work.

ACKNOWLEDGMENTS

The authors thank the Machine Perception Lab (MPLab) at UCSD for providing the CERT package.

REFERENCES

[1] S. Brennan, "Lexical entrainment in spontaneous dialog," in *International Symposium on Spoken Dialogue (ISSD-96)*, October 1996, pp. 41–44.

[2] M. E. Babel, "Phonetic and social selectivity in speech accommodation," Ph.D. dissertation, University of California Berkeley, Department of Linguistics, Berkeley, CA, USA, Spring 2009.

[3] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *12th Annual Conference of the International Speech Communication Association (Interspeech'2011)*, Florence, Italy, August 2011, pp. 3081–3084.

[4] M. Natale, "Convergence of mean vocal intensity in dyadic communication as a function of social desirability," *Journal of Personality and Social Psycholoxy*, vol. 32, no. 5, pp. 790–804, November 1975.

[5] R. Coulston, S. Oviatt, and C. Darves, "Amplitude convergence in children's conversational speech with animated personas," in *International Conference on Spoken Language Processing (ICSLP 2002)*, vol. 4, Denver, CO, USA, September 2002, pp. 2689–2692.

[6] C. Breazeal, "Regulation and entrainment in human-robot interaction," *International Journal of Robotics Research*, vol. 21, no. 10-11, pp. 883–902, October-November 2002.

[7] L. Mol, E. Krahmer, and M. Swerts, "Alignment in iconic gestures: Does it make sense?" in *International Conference on Auditory-Visual Speech Processing (AVSP 2009)*, Norwich, United Kingdom, September 2009, pp. 3–8.

[8] T. L. Chartrand and J. A. Bargh, "The chameleon effect: The perception-behavior link and social interaction," *Journal of Personality and Social Psychology*, vol. 76, no. 6, pp. 893–910, June 1999.

[9] R. Levitan, A. Gravano, and J. Hirschberg, "Entrainment in speech preceding backchannels," in *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2009)*, vol. 2, Portland, OR, USA, June 2009, pp. 113–117.

[10] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *15th International Congress of Phonetic Sciences (ICPhS 03)*, Barcelona, Spain, August 2003, pp. 2453–2456.

[11] R. Porzel, A. Scheffler, and R. Malaka, "How entrainment increases dialogical effectiveness," in *Proceedings of the International Conference on Intelligent User Interfaces (IUI 2006), Workshop on Effective Multimodal Dialogue Interfaces*, Sydney, Australia, January 2006.

[12] P. A. Andersen and L. K. Guerrero, *Handbook of Communication and Emotion: Research, Theory, Applications, and Contexts*. New York, NY, USA: Academic Press, October 1997.

[13] H. Giles, J. Coupland, and N. Coupland, *Contexts of Accommodation: Developments in Applied Sociolinguistics*. New York, NY, USA: Cambridge University Press, September 1991.

[14] R. A. Hinde, *Towards understanding relationships*. New York, NY, USA: Academic Press Inc, December 1979.

[15] J. Burgoon, L. Stern, and L. Dillman, *Interpersonal Adaptation: Dyadic Interaction Patterns*. New York, NY, USA: Cambridge University Press, October 1995.

[16] T. Iio, M. Shiomi, K. Shinozawa, T. Miyashita, T. Akimoto, and N. Hagita, "Lexical entrainment in human-robot interaction: can robots entrain human vocabulary?" in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009)*, St. Louis, MO, USA, October 2009, pp. 3727–3734.

[17] T. Kanda, H. Ishiguro, M. Imai, and T. Ono, "Development and evaluation of interactive humanoid robots," *Proceedings of the IEEE*, vol. 92, no. 11, pp. 1839– 1850, November 2004.

[18] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in *8th International Conference on Spoken Language Processing (ICSLP 04)*, Jeju Island, Korea, October 2004, pp. 2193–2196.

[19] P. Ekman and E. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS)*. New York, NY, USA: Oxford University Press, 1997.

[20] C.-C. Lee, M. Black, A. Katsamanis, A. Lammert, B. Baucom, A. Christensen, P. Georgiou, and S. Narayanan, "Quantification of prosodic entrainment in affective spontaneous spoken interactions of married couples," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 793–796.

[21] 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.

[22] A. Metallinou, A. Katsamanis, and S. Narayanan, "A hierarchical framework for modeling multimodality and emotional evolution in affective dialogs," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 2401–2404.

[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] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," 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. 17–22.

[25] 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.

[26] S. Mariooryad and C. Busso, "Factorizing speaker, lexical and emotional variabilities observed in facial expressions," in *IEEE International Conference on Image Processing (ICIP 2012)*, Orlando, FL, USA, September-October 2012, pp. 2605–2608.

[27] A. Metallinou, C. Busso, S. Lee, and S. Narayanan, "Visual emotion recognition using compact facial representations and viseme information," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 2474–2477.

[28] C. Busso and S. Narayanan, "Interplay between linguistic and affective goals in facial expression during emotional utterances," in *7th International Seminar on Speech Production (ISSP 2006)*, Ubatuba-SP, Brazil, December 2006, pp. 549–556.

[29] ——, "Interrelation between speech and facial gestures in emotional utterances: a single subject study," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2331–2347, November 2007.

[30] S. Mariooryad and C. Busso, "Feature and model level compensation of lexical content for facial emotion recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, April 2013.

[31] M. A. Hall, "Correlation based feature-selection for machine learning," Ph.D. dissertation, The University of Waikato, Hamilton, New Zealand, April 1999.

[32] 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.

[33] 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.

[34] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, April 1990.

[35] 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.

[36] A. Metallinou, S. Lee, and S. Narayanan, "Decision level combination of multiple modalities for recognition and analysis of emotional expression," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, Dallas, TX, USA, March 2010, pp. 2462–2465.

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

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

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

[40] 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.

[41] 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.

[42] 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, pp. 22–35, September 2006.

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

[44] I. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation - The standard, implementations and applications*. John Wiley & Sons, November 2002.

[45] 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.

[46] 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.

[47] 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.

[48] J. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011)*, Prague, Czech Republic, May 2011, pp. 4940–4943.

Soroosh Mariooryad (S'2012) 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 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) lab at UTD. 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 worked on statistical speech enhancement and fingerprint recognition.

Carlos Busso (S'02-M'09) 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 Chile in 2003. 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 co-author 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, and machine learning methods for multimodal processing.