IEMOCAP: Interactive emotional dyadic motion capture database

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee and Shrikanth S. Narayanan
Speech Analysis and Interpretation Laboratory (SAIL)
University of Southern California, Los Angeles, CA 90089

October 15th, 2007

Abstract. Since emotions are expressed through a combination of verbal and non-verbal channels, a joint analysis of speech and gestures is required to understand expressive human communication. To facilitate such investigations, this paper describes a new corpus named the "interactive emotional dyadic motion capture database" (IEMOCAP), collected by the Speech Analysis and Interpretation Laboratory (SAIL) at the University of Southern California (USC). This database was recorded from ten actors in dyadic sessions with markers on the face, head, and hands, which provide detailed information about their facial expression and hand movements during scripted and spontaneous spoken communication scenarios. The actors performed selected emotional scripts and also improvised hypothetical scenarios designed to elicit specific types of emotions (happiness, anger, sadness, frustration and neutral state). The corpus contains approximately twelve hours of data. The detailed motion capture information, the interactive setting to elicit authentic emotions, and the size of the database make this corpus a valuable addition to the existing databases in the community for the study and modeling of multimodal and expressive human communication.

Keywords: Audio-visual database, dyadic interaction, emotion, emotional assessment, motion capture system

1. Introduction

One of the most interesting paralinguistic messages expressed during human interaction is the emotional state of the subjects, which is conveyed through both speech and gestures. The tone and energy of the speech, facial expressions, torso posture, head position, hand gestures, and gaze are all combined in a nontrivial manner, as they unfold during natural human communication. These communicative channels need to be jointly studied if robust emotional models are to be developed and implemented.

In this context, one of the major limitations in the study of emotion expression is the lack of databases with genuine interaction that comprise integrated information from most of these channels. Douglas-Cowie *et al.* analyzed some of the existing emotional databases [28], and concluded that in most of the corpora the subjects were asked to simulate ("act") specific emotions. While desirable from the viewpoint of providing controlled elicitation, these simplifications in data collection, however, discarded important information observed in real life scenarios [29]. As a result, the performance of emotion recognition significantly decreases when the automatic recognition models developed by such databases are used in real life applications [6], where a blend of emotions is observed [29, 27] (i.e., combinations of the "basic emotions" [31]). Another limitation of existing corpora is that the recorded materials often consist of isolated utterances or dialogs with few turns [28]. This setting neglects important effects of contextualization, which play a crucial role in how we perceive [19] and express emotions [29]. Likewise, most of the existing databases contain only the acoustic speech channel. Therefore, these corpora cannot be used to study the information that is conveyed through the other communication channels. Other limitations of current emotional databases are the limited number of subjects, and the small size of the databases [29]. Similar observations were also presented in the review presented by Ververidis and Kotropoulos [54].

Considering these limitations, a new audio-visual database was designed, which notably includes direct and detailed motion capture information that would facilitate access to detailed gesture information not afforded by the state of the art in video processing. In this database, which will be referred here on as the *interactive emotional dyadic motion capture database* (IEMOCAP), ten actors were recorded in dyadic sessions (5 sessions with 2 subjects each). They were asked to perform three selected scripts with clear emotional content. In addition to the scripts, the subjects were also asked to improvise dialogs in hypothetical scenarios, designed to elicit specific emotions (happiness, anger, sadness, frustration and neutral state). One participant of the pair was motion captured at a time during each interaction. Fifty-three facial markers were attached to the subject being motion captured, who also wore wristbands and a headband with markers to capture hand and head motion, respectively (see Figure 1). Using this setting, the emotions were elicited within a proper context, improving the authenticity of the captured emotional data. Furthermore, gathering data from ten different subjects increases the plausibility of effectively analyzing trends observed in this database on a more general level. In total, the database contains approximately twelve hours of data.

This corpus, which took approximately 20 months to collect (from the design to the post processing stages), is hoped to add to the resources that can help advance research to understand how to model expressive human communication. With this database, we hope to be able to expand and generalize our previous results about the relation-

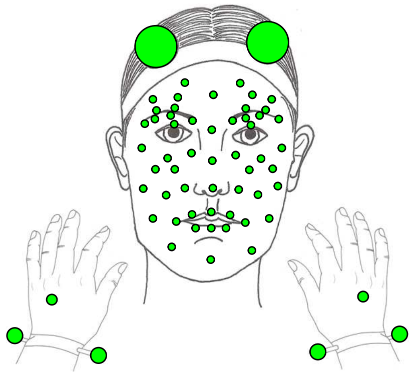


Figure 1. Marker layout. In the recording, fifty-three markers were attached to the face of the subjects. They also wore wristbands (two markers) and headband (two markers). An extra marker was also attached on each hand.

ship and interplay between speech, facial expressions, head motion and hand gestures during expressive speech, and conversational interactions [9, 11, 12, 13]. Likewise, we expect to model and synthesize different aspects of human behaviors, using an unified framework that properly take into consideration the underlying relationship between facial expressions and speech (e.g., head motion driven by speech prosody [7, 8]).

The rest of this paper is organized as follows. Section 2 presents a review of audio-visual databases that have been used to study emotions. Section 3 describes the design of the corpus presented in this paper. Section 4 explains the recording procedures of the database. Section 5 presents the various post processing steps such as reconstruction of the marker data, segmentation and emotional evaluation. Section 6 discusses how the IEMOCAP database overcomes some of the main limitations in the current state of the art emotional databases. It also comments on some of the research questions that can be studied using this data. Finally, Section 7 presents the conclusion and final remarks.

2. Brief review of audio-visual databases

One of the crucial improvements that is needed to achieve major progress in the study of emotion expression is the collection of new databases that overcome the limitations existing in current emotional corpora. Douglas-Cowie *et al.* discussed the state of the art emotional databases [28], focusing on four main areas: scope (number of speakers, emotional classes, language, etc), naturalness (acted versus spontaneous), context (in-isolation versus in-context) and descriptors (linguistic and emotional description). They highlight the importance of having suitable databases with natural emotions recorded during an interaction rather than monologues. Other requirements for a good database are multiple speakers, multimodal information capture and adequate descriptors of the emotion contained in the corpus.

Given the multiple variables considered in the study of emotions, it is expected that a collection of databases rather than a single corpus will be needed to address many of the open questions in this multidisciplinary area. Unfortunately, there are currently few emotional databases that satisfy these core requirements. Some of the most successful efforts to collect new emotional databases to date have been based on broadcasted television programs. Some of these examples are the Belfast natural database [28, 29], the VAM database [37, 38] and the EmoTV1 database [1]. Likewise, movie excerpts with expressive content have also been proposed for emotional corpora, especially for extreme emotions (e.g., SAFE corpus [20]). Nevertheless, one important limitation of these approaches is the copyright and privacy problems that prevent the wide distribution of the corpora [23, 28]. Also, the position of the microphones and cameras, the lexical and emotional content, and the visual and acoustic backgrounds cannot be controlled, which challenge the processing of the data [23]. Other attempts to collect natural databases were based on recordings in situ (Genova Airport Lost Luggage database [48]), recording spoken dialogs from real call center (the CEMO [56], and CCD [44] corpora), asking the subjects to recall emotional experiences [2], inducing emotion with a Wizard of Oz approach in problem-solving settings using a human-machine interface (e.g., SmartKom database [50]), using games specially designed to emotionally engage the users (e.g., the EmoTaboo corpus [57]), and inducing emotion through carefully designed human-machine interaction (i.e., SAL [17, 23]). In the Humaine project portal, further descriptions of some of the existing emotional databases are presented [41].

Recording professional actors under controlled conditions can overcome many of the limitations of the aforementioned recording techniques. We have claimed in our previous work that good quality acted databases can be recorded, when suitable acting methodologies are used to elicit emotional realizations from experienced actors, engaged in dialogs rather than monologues [15]. The Geneva Multimodal Emotion Portrayal (GEMEP) [5] is a good example. Enos and Hirschberg argued that emotion arises as a consequence of what it is expected and what it is finally achieved [32]. They suggested that acted databases could produce more realistic emotions if this goal-oriented approach is suitably incorporated in the recording.

In order to make a unified analysis of verbal and nonverbal behavior of the subjects possible, the database should include the visual channel capturing gestures and facial expression in conjunction with the aural channel. Although there are automatic platforms to track salience features in the face using images (e.g., [21]), the level of the detailed facial information provided by motion capture data is not presently achievable using the state of art in video processing. This is especially notable in the cheek area, in which there are no salience feature points. To the best of our knowledge, few motion capture databases exist for the study of emotional expression. Kapur *et al.* presented an emotional motion capture database, but they targeted only body postures (no facial expressions) [42]. The USC Facial Motion Capture Database (FMCD), our previous audio-visual database, is another example [9]. This database was recorded from a single actress with markers attached to her face, who was asked to read semantically-neutral sentences expressing specific emotions. The two main limitations of this corpus are that the emotions were elicited in isolated sentences, and that only one speaker was recorded. The IEMOCAP database described in this paper was designed to overcome some of these basic limitations. The requirements considered in the design of the IEMOCAP database are listed below.

- The database must contain genuine realizations of emotions.
- Instead of monologues and isolated sentences, the database should contain natural dialogues, in which the emotions are suitably and naturally elicited.
- Many experienced actors should be recorded.
- The recording of the database should be as controlled as possible in terms of emotional and linguistic content.
- In addition to the audio channel for capturing verbal behavior, the database should have detailed visual information to capture the nonverbal information, all in a synchronized manner.
- The emotional labels should be assigned based on human subjective evaluations.

Notice that there are inherent tradeoffs between some of these requirements (e.g., naturalness versus control of expression content). The next sections describe how these requirements were addressed in the IEMOCAP database collection.

3. The design of the database

The IEMOCAP database was designed toward expanding our research in expressive human communication. In our previous work, we have analyzed the relationship between gestures and speech [12], and the interplay between linguistic and affective goals in these communicative channels [10, 11]. Our results indicated that gestures and speech present high levels of correlation and coordination, and that the emotional modulation observed in the different communicative channels is not uniformly distributed. In fact, our results indicated that when one modality is constrained by speech articulation, other channels with more degrees of freedom are used to convey the emotions [13]. As a result of the analysis, we have presented applications in automatic machine recognition and synthesis of expressive human behavior. For example, we have modeled and synthesized aspects of human behavior in virtual characters. In particular, we proposed an HMM based framework to synthesize natural head motions driven by acoustic prosodic features [7, 8]. In all these studies, the FMCD database was used. Since this database was recorded from a single subject, this new corpus will allow us to validate and expand our research. Section 6.2 provides details on the new directions that we are planning to explore with this database.

In our previous work, we have predominantly focused on happiness, anger, sadness and the neutral state [9, 11, 12, 13]. These categories are among the most common emotional descriptors found in the literature [47]. For this database, we decided also to include frustration, since it is also an important emotion from an application point of view. Therefore, the content of the corpus was designed to cover those five emotions. As will be discussed in Section 5, during the emotional evaluation, the emotional categories were expanded to include disgust, fear, excitement and surprise. The purpose of doing so was to have a better description of the emotions found in the corpus, notably in the spontaneous/unscripted elicitation scenarios.

The most important consideration in the design of this database was to have a large emotional corpus with many subjects, who were able to express genuine emotions. To achieve these goals, the content of the corpus and the subjects were carefully selected.

3.1. Material selection

Instead of providing reading material, in which the emotions are not guaranteed to be genuinely expressed during the recording [28], two different approaches were selected: the use of plays (scripted sessions), and improvisation based hypothetical scenarios (spontaneous sessions).

The first approach is based on a set of scripts that the subjects were asked to memorize and rehearse. The use of plays provides a way of constraining the semantic and emotional content of the corpus. Three scripts were selected after reading more than one hundred 10-minute plays. A theater professional supervised the selection given the requirement that the plays should convey the target emotions (happiness, anger, sadness, frustration and neutral state). In addition, these plays were selected so that they each consisted of a female and a male role. This requirement was imposed to balance the data in terms of gender. Since these emotions are expressed within a suitable context, they are more likely to be conveyed in a genuine manner, in comparison to the recordings of simple isolated sentences.

In the second approach, the subjects were asked to improvise based on hypothetical scenarios that were designed to elicit specific emotions (see Table I). The topics for the spontaneous scenarios were selected following the guidelines provided by Scherer *et al.* [49]. As reported in their book, the authors polled individuals who were asked to remember situations in the past that elicited certain emotions in them. The hypothetical scenarios were based on some common situations (e.g., loss of a friend, separation). In this setting, the subjects were free to use their own words to express themselves. By granting the actors a considerable amount of liberty in the expression of their emotions, we expected that the results would provide genuine realization of emotions.

A comparison of the advantages and disadvantages of these two elicitation approaches is given in our previous work [16].

3.2. Actors selection

As suggested in [28], skilled actors engaged in their role during interpersonal drama may provide a more natural representation of the emotions. Therefore, this database relied on seven professional actors and three senior students from the Drama Department at the University of Southern California. Five female and five male actors were selected, after reviewing their audition sessions. They were asked to rehearse the scripts under the supervision of an experienced professional (functioning as a director) who made sure the scripts were memorized and the intended emotions were genuinely expressed, avoiding exaggeration or caricature of the emotions.

The subjects were recorded in five dyadic recording sessions, each of which lasted approximately six hours, including suitable rest periods.

Table I. Scenarios used for eliciting unscripted/unrehearsed interactions in the database collection. The target emotions for each subject are given in parenthesis (Fru = Frustation, Sad = Sadness, Hap = Happiness, Ang = Anger, Neu = Neutral).

	Subject 1 (with markers)	Subject 2 (without markers)
1	(Fru) The subject is at the *Department of Motor Vehicles* (DMV) and he/she is being sent back after standing in line for an hour for not having the right form of IDs.	(Ang) The subject works at DMV. He/she rejects the application.
2	(Sad) The subject, a new parent, was called to enroll the army in a foreign country. He/she has to separate from his/her spouse for more than 1 year.	(Sad) The subject is his/her spouse and is extremely sad for the separation.
3	(Hap) The subject is telling his/her friend that he/she is getting married.	(Hap) The subject is very happy and wants to know all the details of the proposal. He/she also wants to know the date of the wedding.
4	(Fru) The subject is unemployed and he/she has spent last 3 years looking for work in his/her area. He/she is losing hope.	(Neu) The subject is trying to encourage his/her friend.
5	(Ang) The subject is furious, because the airline lost his/her baggage and he/she will receive only $50 (for a new bag that cost over $150 and has lots of important things).	(Neu) The subject works for the airline. He/she tries to calm the customer.
6	(Sad) The subject is sad because a close friend died. He had cancer that was detected a year before his death.	(Neu) The subject is trying to support his friend in this difficult moment.
7	(Hap) The subject has been accepted at USC. He/she is telling this to his/her best friend.	(Hap) The subject is very happy and wants to know the details (major, scholarship). He/she is also happy because he/she will stay in LA so they will be together.
8	(Neu) He/She is trying to change the mood of the customer and solve the problem.	(Ang) After 30 minutes talking with a machine, he/she is transferred to an operator. He/she expresses his/her frustration, but, finally, he/she changes his/her attitude.

Since the selected scripts have a female and a male role, an actor and an actress were recorded in each of the five sessions (see Fig. 2).

4. Recording of the corpus

For each of the sessions, fifty-three markers (diameter $\approx$ 4mm) were attached to the face of one of the subjects in the dyad to capture detailed facial expression information, while keeping the markers far from each other to increase the accuracy in the trajectory reconstruction step.

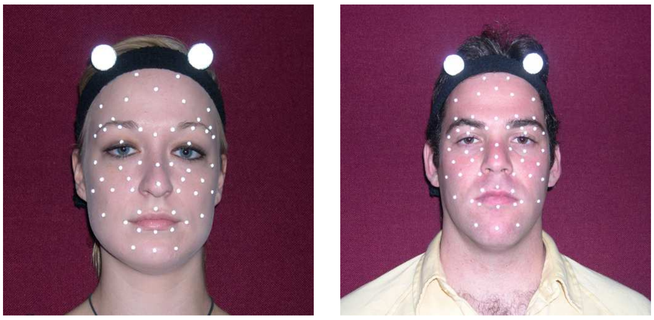


Figure 2. Two of the actors who participated in the recording, showing the markers on the face and headband.

Most of the facial markers were placed according to the feature points defined in the MPEG-4 standard [46, 52]. Figures 1 and 2 show the layout of the markers. The subject wore a headband with two markers on it (diameter $\approx$ 2.5cm). These markers, which are static with respect to the facial movements, are used to compensate for head rotation. In addition, the subject wore wristbands with two markers each (diameter $\approx$ 1cm). An extra marker in each hand was also added. Since only three markers are used in each hand, it is not possible to have detailed hand gestures (e.g., of fingers). Nevertheless, the information provided by these markers give a rough estimate of the hands' movements. In total, 61 markers were used in the recording (Fig. 1). Notice that the markers are very small and do not interfere with natural speech. In fact, the subjects reported that they felt comfortable wearing the markers, which did not prevent them from speaking naturally.

After the scripts and the spontaneous scenarios were recorded, the markers were attached to the other subject, and the sessions were recorded again after a suitable rest. Notice that the original idea was to have markers on both speakers at the same time. However, the current approach was preferred to avoid interference between two separate setups. The VICOM cameras are sensitive to any reflected material in their field of view. Therefore, it is technically difficult to locate the additional equipments in the room without affecting the motion capture recording (computer, microphones, cameras). Furthermore, with this setting, all the cameras were directed to one subject, increasing the resolution and quality of the recordings.

The database was recorded using the facilities of the John C. Hench Division of Animation & Digital Arts (Robert Zemeckis Center) at USC. The trajectories of the markers data were recorded using a VICON motion capture system with eight cameras that were placed approximately one meter from the subject with markers, as can be

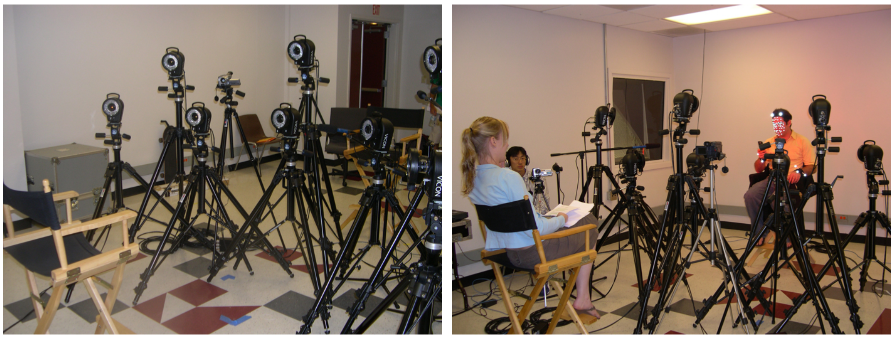


Figure 3. VICON motion capture system with 8 cameras. The subject with the markers sat in the middle of the room, with the cameras directed to him/her. The subject without the markers sat outside the field of view of the VICON cameras, facing the subject with markers.

seen in Figure 3. The sample rate of the motion capture system was 120 frames per second. To avoid having gestures outside the volume defined by the common field of view of the VICOM cameras, the subjects were asked to be seated during the recording. However, they were instructed to gesture as naturally as possible, while avoiding occluding their face with the hands. The subject without the markers was sitting out of the field of view of the VICON cameras to avoid possible interferences. As a result of this physical constraint, the actors were separated approximately three meters from each other. Since the participants were within the *social distance* as defined by Hall [39], we expect that the influence of proxemics did not affect their natural interaction. At the beginning of each recording session, the actors were asked to display a neutral pose of the face for approximately two seconds. This information can be used to define a neutral pose of the markers.

The audio was simultaneously recorded using two high quality shotgun microphones (Schoeps CMIT 5U) directed at each of the participants. The sample rate was set to 48KHz. In addition, two high-resolution digital cameras (Sony DCR-TRV340) were used to record a semi frontal view of the participants (see Fig. 5). These videos were used for emotion evaluation, as will be discussed in Section 5.

The recordings were synchronized by using a clapboard with reflective markers attached to its ends. Using the clapboard, the various modalities can be accurately synchronized with the sounds collected by the microphone, and the images recorded by the VICON and digital cameras. The cameras and microphones were placed in such a way that the actors could face each other, a necessary condition for natural interaction. Also, the faces were within the line of sight - not talking

Table II. Segmentation statistics of the IEMOCAP database speech. Comparative details for popular spontaneous spoken dialog corpora are also shown.

	IEMOCAP			Other spontaneous corpora	
	All turns	Scripted	Spontaneous	Switchboard-I	Fisher
Turn duration [sec]	4.5	4.6	4.3	4.5	3.3
Words per turn	11.4	11.4	11.3	12.3	9.9

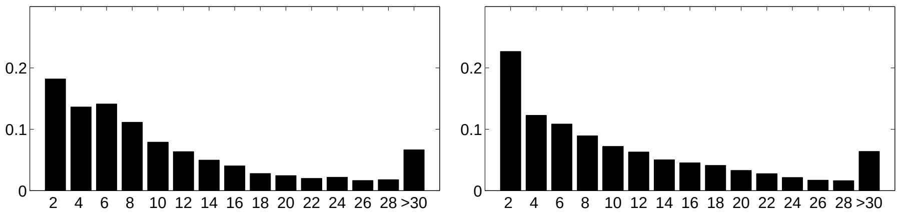
to the back of a camera. In fact, the actors reported that the side conditions of the recording did not affect their natural interaction.

5. Post processing

5.1. Segmentation and transcription of the data

After the sessions were recorded, the dialogs were manually segmented at the dialog turn level (speaker turn), defined as continuous segments in which one of the actors was actively speaking (see Figure 5 which shows two turns with emotional evaluation). Short turns showing active listening such as “mmhh” were not segmented. Multi-sentence utterances were split as single turns. For the scripted portion of the data (see Section 3.1), the texts were segmented into sentences in advance and used as reference to split the dialogs. This segmentation was used only as guidance, since we did not require having the same segmentation in the scripts across sessions. In total the corpus contained ten thousand and thirty nine turns (scripted session: 5255 turns; spontaneous sessions: 4784 turns) with an average duration of 4.5 seconds. The average value of words per turn was 11.4. The histograms of words per turn for the scripted and spontaneous sessions are given in Figure 4. These values are similar to the turn statistics observed in well-known spontaneous corpora such as Switchboard-1 Telephone Speech Corpus (Release 2) and Fisher English Training Speech Part 1 (see Table II).

The professional transcription of the audio dialogs (i.e., what the actors said) was obtained from Ubiqus [53] (see Table III for an example). Then, forced alignment was used to estimate the word and phoneme boundaries. Conventional acoustic speech models were trained with over 360 hours of neutral speech, using the Sphinx-III speech recognition system (version 3.0.6) [40]. Although we have not rigorously evaluated the alignment results, our preliminary screening suggests that the boundaries are accurate, especially in segments with no speech



(a) Scripted session (b) Spontaneous sessions

Figure 4. Histogram with the number of words per turns (in percentage) in the scripted and spontaneous sessions.

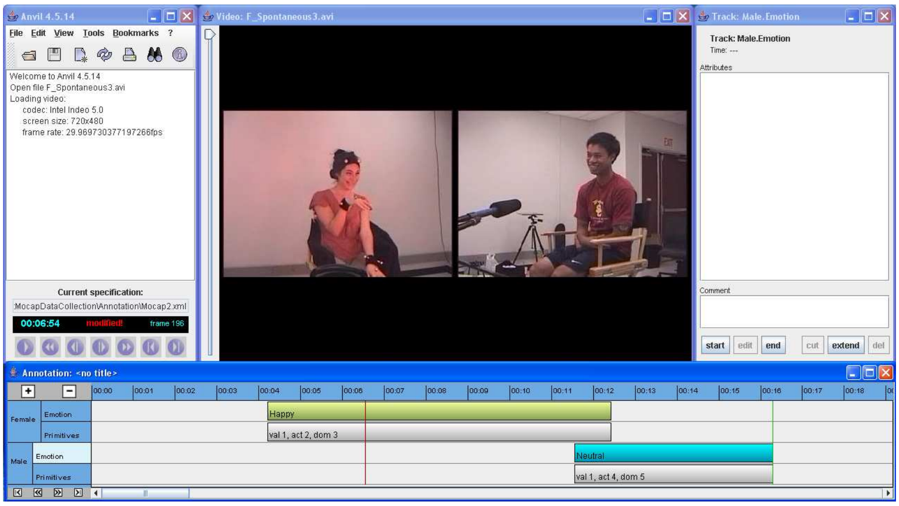
Table III. Example of the annotations for a portion of a spontaneous session (third scenario in Table I). The example includes the turn segmentation (in seconds), the transcription, the categorical emotional assessments (three subjects) and the attribute emotional assessment (valence, activation, dominance, two subjects).

Seg. [sec]	Turn	Transcription	Labels	[v,a,d]
[05.0 – 07.8]	F00:	Oh my God. Guess what, guess what, guess what, guess what, guess what, guess what?	[exc][exc][exc]	[5,5,4][5,5,4]
[07.8 – 08.7]	M00:	What?	[hap][sur][exc]	[4,4,3][4,2,1]
[08.9 – 10.8]	F01:	Well, guess. Guess, guess, guess, guess.	[exc][exc][exc]	[5,5,4][5,5,4]
[11.1 – 14.0]	M01:	Um, you–	[hap][neu][neu]	[3,3,2][3,3,3]
[14.2 – 16.0]	F02:	Don't look at my left hand.	[exc][hap][hap;exc]	[4,3,2][5,4,3]
[17.0 – 19.5]	M02:	No. Let me see.	[hap][sur][sur]	[4,4,4][4,4,4]
[20.7 – 22.9]	M03:	Oh, no way.	[hap][sur][exc]	[4,4,4][5,4,3]
[23.0 – 28.0]	F03:	He proposed. He proposed. Well, and I said yes, of course. [LAUGHTER]	[exc][hap][hap;exc]	[5,4,3][5,5,3]
[26.2 – 30.8]	M04:	That is great. You look radiant. I should've guess.	[hap][hap][exc]	[4,4,3][5,3,3]
[30.9 – 32.0]	F04:	I'm so excited.	[exc][exc][exc]	[5,4,3][5,5,3]
[32.0 – 34.5]	M05:	Well, Tell me about him. What happened	[hap][exc][exc]	[4,4,3][4,4,4]

overlaps. Knowing the lexical content of the utterances can facilitate further investigations into the interplay between gestures and speech in terms of linguistic units [11, 12, 13].

5.2. Emotional annotation of the data

In most of the previous emotional corpus collections, the subjects are asked to express a given emotion, which is later used as the emotional label. A drawback of this approach is that it is not guaranteed that the recorded utterances reflect the target emotions. Additionally, a given display can elicit different emotional percepts. To avoid these problems, the emotional labels in this corpus were assigned based on



```
Welcome to Anvil 4.5.14
Open file F_Spontaneous3.avi
Loading video:
    codec: Intel Indeo 5.0
    screen size: 720x480
    frame rate: 29.969730377197266fps
```

Current specification: :MocapDataCollection\Annotation\Mocap2.xml

Track: Male.Emotion
Time: ---

		Emotion	Primitives
Female		Happy	val 1, act 2, dom 3
Male		Neutral	val 1, act 4, dom 5

Figure 5. ANVIL annotation tool used for emotion evaluation. The elements were manually created for the turns. The emotional content of the turns can be evaluated based on categorical descriptors (e.g., happiness, sadness) or primitive attribute (e.g., activation, valence).

agreements derived from subjective emotional evaluations. For that purpose, human evaluators were used to assess the emotional content of the database. The evaluators were USC students who were fluent English speakers.

Different methodologies and annotation schemes have been proposed to capture the emotional content of databases (i.e., Feeltrace tool [24], *Context Annotation Scheme* (MECAS) [27]). For this database, two of the most popular assessment schemes were used: discrete categorical based annotations (i.e., labels such as happiness, anger, and sadness), and continuous attribute based annotations (i.e., activation, valence and dominance). These two approaches provide complementary information of the emotional manifestations observed in corpus.

The *"annotation of video and spoken language"* tool ANVIL [43] was used to facilitate the evaluation of the emotional content of the corpus (see Fig. 5). Notice that some emotions are better perceived from audio (e.g., sadness) while others from video (e.g., anger) [26]. Also, the context in which the utterance is expressed plays an important role in recognizing the emotions [19]. Therefore, the evaluators were asked to sequentially assess the turns, after watching the videos. Thus, the acoustic and visual channels, and the previous turns in the dialog were available for the emotional assessment.

One assumption made in this evaluation is that, within a turn, there is no transition in the emotional content (e.g., from *frustration* to *anger*). This simplification is reasonable, since the average duration of the turns is only 4.5 seconds. Therefore the emotional content is expected to be kept constant. Notice that the evaluators were allowed to tag more than one emotional category per turn, to account for mixtures of emotions (e.g., *frustration* and *anger*), which are commonly observed in human interaction [27].

Categorical emotional descriptors

Six human evaluators were asked to assess the emotional content of the database in terms of emotional categories. The evaluation sessions were organized so that three different evaluators assessed each utterance. The underlying reason was to minimize evaluation time for the preliminary analysis of the database. The evaluation was divided into approximately 45-minute sessions. The evaluators were instructed to have a suitable rest between sessions.

As mentioned in Section 3, the database was designed to target anger, sadness, happiness, frustration and neutral state. However, some of the sentences were not adequately described with only these emotion labels. Since the interactions were intended to be as natural as possible, the experimenters expected to observe utterances full of excitement, fear and other broad range of mixed emotions that are commonly seen during natural human interactions. As described by Devillers *et al.*, emotional manifestations not only depends on the context, but also on the person [27]. They also indicated that ambiguous emotions (non-basic emotions) are frequently observed in real-life scenarios. Therefore, describing emotion is an inherent complex problem. As a possible way to simplify the fundamental problem in emotion categorization, an expanded set of categories was used for emotion evaluation. On the one hand, if the number of emotion categories is too extensive, the agreement between evaluators will be low. On the other hand, if the list of emotions is limited, the emotional description of the utterances will be poor and likely less accurate. To balance the tradeoff, the final emotional categories selected for annotation were anger, sadness, happiness, disgust, fear and surprise (known as basic emotions [31]), plus frustration, excited and neutral states.

Figure 6 shows the Anvil emotion category menu used to label each turn. Although it was preferred that the evaluators chose only a single selection, they were allowed to select more than one emotional label to account for blended emotions [27]. If none of the available emotion categories were adequate, they were instructed to select *other* and write their own comments. For the sake of simplicity, majority voting was used for emotion class assignment, if the emotion category with the highest votes was unique (notice that the evaluators were allowed to tag more than one emotion category). Under this criterion, the evaluators

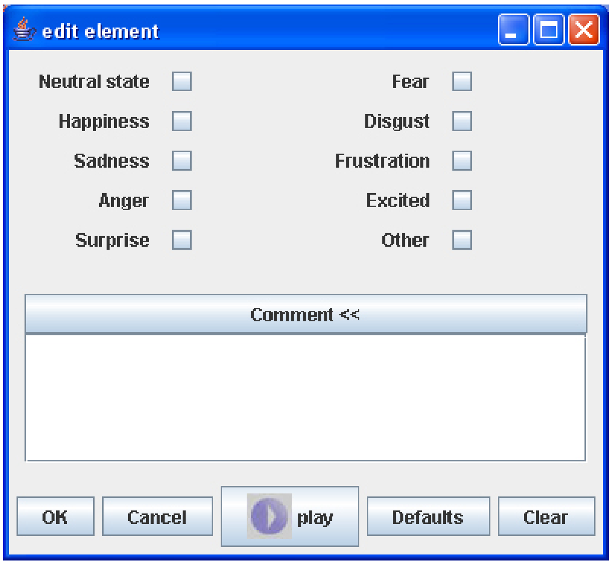


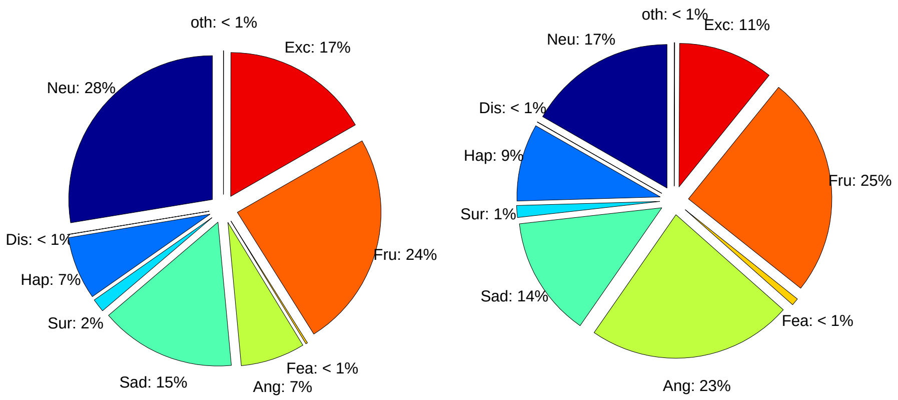
Figure 6. ANVIL emotion category menu presented to the evaluators to label each turn. The evaluators could select more than one emotions and add their own comments.

reached agreement in 74.6% of the turns (scripted session: 66.9%; spontaneous sessions: 83.1%). Notice that other approaches to reconciling the subjective assessment are possible (e.g., entropy based method [51], and multiple labels [27]).

Figure 7 shows the distribution of the emotional content in the data for the turns that reached agreement. This figure reveals that the IEMOCAP database exhibits a balanced distribution of the target emotions (happiness, anger, sadness, frustration and neutral state). As expected, the corpus contains few examples of other emotional categories such as fear and disgust.

Using the assigned emotional labels as ground truth, the confusion matrix between emotional categories in the human evaluation was estimated. The results are presented in Table IV. On average, the classification rate of the emotional categories was 72%. The table shows that some emotions such as neutral, anger and disgust are confused with frustation. Also, there is an overlap between happiness and excitement.

To analyze the inter-evaluator agreement, Fleiss' *Kappa* statistic was computed [34] (see Table V). The result for the entire database is $\kappa = 0.27$. The value of the Fleiss' *Kappa* statistic for the turns in which the evaluators reached agreements according to the criterion mentioned before is $\kappa = 0.40$. Since the emotional content of the database mainly span the target emotions (see Fig. 7), the *Kappa* statistic was recalculated after clustering the emotional categories as



(a) Scripted session (b) Spontaneous sessions

Figure 7. Distribution of the data for each emotional category. The figure only contains the sentences in which the category with the highest vote was unique (Neu = neutral state, Hap = happiness, Sad = sadness, Ang = anger, Sur = surprise, Fea = fear, Dis = disgust, Fru = Frustation, Exc = Excited and Oth = Other).

Table IV. Confusion matrix between emotion categories estimated from human evaluations (Neu = neutral state, Hap = happiness, Sad = sadness, Ang = anger, Sur = surprise, Fea = fear, Dis = disgust, Fru = frustation, Exc = excited and Oth = other).

Emotional labels	Neu	Hap	Sad	Ang	Sur	Fea	Dis	Fru	Exc	Oth
Neutral state	0.74	0.02	0.03	0.01	0.00	0.00	0.00	0.13	0.05	0.01
Happiness	0.09	0.70	0.01	0.00	0.00	0.00	0.00	0.01	0.18	0.01
Sadness	0.08	0.01	0.77	0.02	0.00	0.01	0.00	0.08	0.01	0.02
Anger	0.01	0.00	0.01	0.76	0.01	0.00	0.01	0.17	0.00	0.03
Surprise	0.01	0.04	0.01	0.03	0.65	0.03	0.01	0.12	0.09	0.01
Fear	0.03	0.00	0.05	0.02	0.02	0.67	0.02	0.05	0.15	0.00
Disgust	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.17	0.17	0.00
Frustation	0.07	0.00	0.04	0.11	0.01	0.01	0.01	0.74	0.01	0.02
Excited	0.04	0.16	0.00	0.00	0.02	0.00	0.00	0.02	0.75	0.00

follows. First, happiness and excited were merged since they are close in the activation and valence domain. Then, the emotional categories fear, disgust and surprise were relabeled as *other* (only for this evaluation). Finally, the labels of the remaining categories were not modified. With this new labeling, the Fleiss' *Kappa* statistic for the entire database and for the turns that reached agreement are $\kappa = 0.35$ and $\kappa = 0.48$, respectively. These levels of agreement, which are considered as fair/moderate agreement, are expected since people have different perception and

Table V. Fleiss' *Kappa* statistic to measure inter-evaluator agreement. The results are presented for all the turns and for the turns in which the evaluators reached agreement

	Original labels		Recalculated labels	
Session	All turns	Reached agreement	All turns	Reached agreement
Entire database	0.27	0.40	0.35	0.48
Scripted sessions	0.20	0.36	0.26	0.42
Spontaneous sessions	0.34	0.43	0.44	0.52

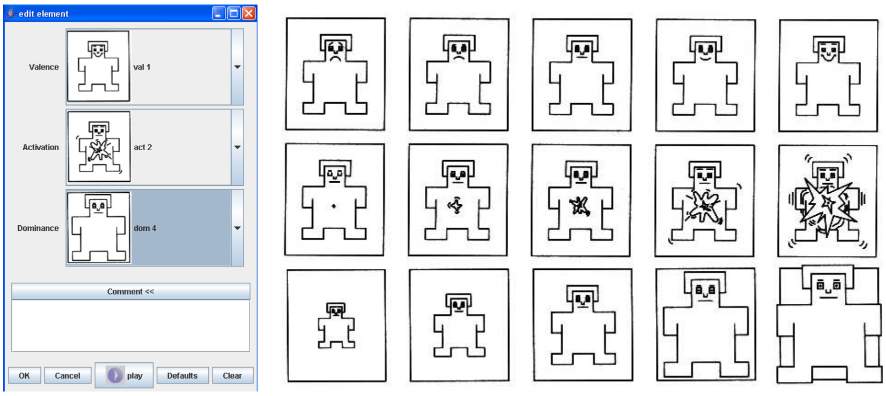
interpretation of the emotions. These values are consistent with the agreement levels reported in previous work for similar tasks [27, 37, 51]. Furthermore, everyday emotions are complex, which may cause poor inter-evaluator agreement [29].

Table V also provides the individual results of the Fleiss Kappa statistic for the scripted and spontaneous sessions. The results reveal that for the spontaneous sessions the levels of inter-evaluator agreement are higher than in the scripted sessions. While spontaneous sessions were designed to target five specific emotions (happiness, anger, sadness, frustration and neutral state), the scripted sessions include progressive changes from one emotional state to another, as dictated by the narrative content of the play. Within a session, the scripted dialog approach typically elicited a wider range of ambiguous emotion manifestations. As a result, the variability of the subjective evaluations increases, yielding to lower level of inter-evaluator agreement. Further analysis comparing scripted and spontaneous elicitation approaches are given in [16].

Continuous emotional descriptors

An alternative approach to describe the emotional content of an utterance is to use primitive attributes such as valence, activation (or arousal), and dominance. This approach, which has recently increased popularity in the research community, provides a more general description of the affective states of the subjects in a continuous space. This approach is also useful to analyze emotion expression variability. The readers are referred to [22], for example, for further details about how to describe emotional content of an utterance using such an approach.

The *self-assessment manikins* (SAMs) were used to evaluate the corpus in terms of the attributes *valence* [1-negative, 5-possitive], *activation* [1-calm, 5-excited], and *dominance* [1-weak, 5-strong] [33, 37] (Fig. 8). This scheme consists of 5 figures per dimension that describe progressive changes in the attribute axis. The evaluators are asked to



Valence val 1
Activation act 2
Dominance dom 4

(a) (b)

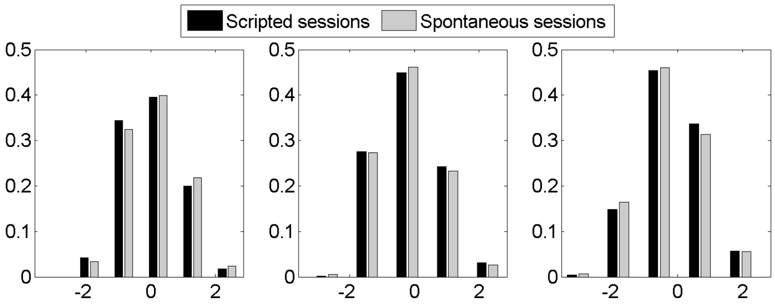
Figure 8. (a) ANVIL attribute-based menu presented to the evaluators to label each turn. (b) Self-assessment manikins. The rows illustrate *valence* (top), *activation* (middle), and *dominance* (bottom).

select the manikin that better describes the stimulus, which is mapped into an integer between 1 to 5 (from left to right). The SAMs system has been previously used to assess emotional speech, showing low standard deviation and high inter-evaluator agreement [36]. Also, using a text-free assessment method bypasses the difficulty that each evaluator has on his/her individual understanding of linguistic emotion labels. Furthermore, the evaluation is simple, fast, and intuitive.

Two different evaluators were asked to assess the emotional content of the corpus using the SAMs system. At this point, approximately 85.5% of the data have been evaluated. After the scores were assigned by the raters, speaker dependent z-normalization was used to compensate for inter-evaluator variation. Figure 9 shows the distribution of the emotional content of the IEMOCAP database in terms of valence, activation and dominance. The histograms are similar to the results observed in other spontaneous emotional corpus [37]).

The Cronbach alpha coefficients were computed to test the reliabilities of the evaluations between the two raters [25]. The results are presented in Table VI. This table shows that the agreement for valence was higher than for the other attributes.

Categorical levels do not provide information about the intensity level of the emotions. In fact, emotional displays that are labeled with the same emotional category can present patterns that are significantly different. Therefore, having both types of emotional descriptions provide complementary insights about how people display emotions and how these cues can be automatically recognized or synthesized for better human-machine interfaces.



(a) Valence (b) Activation (c) Dominance

Figure 9. Distribution of the emotional content of the corpus in terms of (a) valence, (b) activation, and (c) dominance. The results are separately displayed for scripted (black) and spontaneous (gray) sessions.

Table VI. Inter-evaluator agreement of the attribute based evaluation measured with the Cronbach alpha coefficient

		Cronbach's Alpha	
Session	Valence	Activation	Dominance
Entire database	0.809	0.607	0.608
Scripted sessions	0.783	0.602	0.663
Spontaneous sessions	0.820	0.612	0.526

5.3. Self- emotional evaluation of the corpus

In addition to the emotional assessments with naïve evaluators, we asked six of the actors who participated in the data collection to self-evaluate the emotional content of their sessions using categorical (i.e., sadness, happiness) and attribute (i.e., activation, valence) approaches. This self emotional evaluation was performed only in the spontaneous/unscripted scenarios (see Section 3.1). Table VII compares the self-evaluation (*"self"*) results with the assessment obtained from the rest of the evaluators (*"others"*). For this table, the emotional labels obtained from majority voting are assumed as ground truth. The turns in which the evaluators did not reach agreement were not considered. The results are presented in terms of classification percentage. Surprisingly, the results show significant differences between the emotional perceptions between naïve evaluators and the actors. Although the emotional labels were estimated only with the agreement between naïve evaluators -and therefore the recognition rates are expected to be

Table VII. Comparison of the recognition rate in percentage between the evaluation by *self* and *others* for the spontaneous/unscripted scenarios (categorical evaluation). The results are presented for six of the actors (e.g., $F03$ = female actress in session 3).

	F01	F02	F03	M01	M03	M05	Average
Self	0.79	0.58	0.44	0.74	0.57	0.54	0.60
Others	0.76	0.80	0.79	0.81	0.80	0.77	0.79

higher- this table suggests that there are significant differences between both assessments.

In our recent work, we studied in further detail the differences between the evaluations from the naïve raters and the self-assessments in terms of inter-evaluator agreement [14]. We analyzed cross evaluation results across the actors and the naïve evaluators by estimating the differences in reliability measures when each of the raters was excluded from the evaluation. The results also revealed a mismatch between the expression and perception of the emotions. For example, the actors were found to be more selective in assigning the emotional labels to their turns. In fact, the kappa value decreased when the self-evaluations were included in the estimation. Notice that the actors are familiar with how they commonly convey different emotions. Unlike the naïve evaluators, they were also aware of the underlying protocols to record the database. Further analysis from both *self* and *others* evaluations is needed to shed light into the underlying differences between how we express and perceive emotions.

5.4. Reconstruction of marker data

The trajectories of the markers were reconstructed using the VICON iQ 2.5 software [55]. The reconstruction process is semi-automatic, since a template with the markers' positions has to be manually assigned to the markers. Also, the reconstruction needs to be supervised to correct the data when the software is not able to track the markers. Cubic interpolation was used to fill gaps when the number of consecutive missing frames for each marker was less than 30 frames (0.25 second).

Unfortunately, some of the markers were lost during the recording, mainly because of sudden movements of the subjects, and the location of the cameras. Since natural interaction was encouraged, the recording was not stopped when the actors performed sudden movements. The cameras were located approximately one meter from the subject to successfully capture hand gestures in the recording. If only facial expressions were recorded, the cameras had to be placed close to the subjects' faces to increase the resolution. Figure 10 shows the markers, in which the percentage of missing frames was higher than 1% of the corpus. The markers with higher percentages are associated with the eyelids and the hands. The reason is that when the subjects had their eyes open, the eyelids' markers were sometime occluded. Since the purpose of these markers was to infer eye blinks, missing markers are also useful information to infer when the subjects' eyes blinked. The main problem of the hands' markers was the self-occlusion between hands.

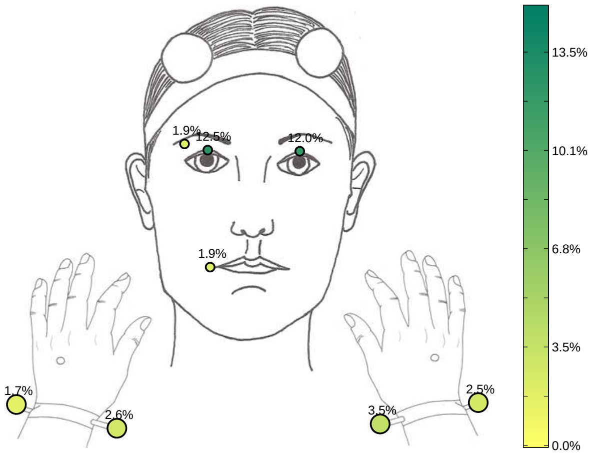


Figure 10. Percentage of the markers that were lost during the recording. The figure only shows the markers that have more than 1% of missing values. Dark colors indicate higher percentage.

After the motion data were captured, all the facial markers were translated to make a nose marker at the local coordinate center of each frame, removing any translation effect. After that, the frames were multiplied by a rotational matrix, which compensates for rotational effects. The technique is based on *Singular value Decomposition* (SVD) and was originally proposed by Arun *et al.* [3]. The main advantage of this approach is that the 3D geometry of every marker is used to estimate the best alignment between the frames and a reference frame. It is robust against markers' noise and its performance overcomes methods that use few "static" markers to compensate head motion.

In this technique, the rotational matrix was constructed for each frame as follows: A neutral facial pose for each subject was chosen as a reference frame, which was used to create a 53×3 matrix, M_{ref}, in which the row of M_{ref} has the 3D position of the markers. For the frame t, a similar matrix M_t was created by following the same marker order as the reference. After that, the SVD, UDV^T, of matrix $M_{ref}^T \cdot M_t$ was calculated. Finally, the product of VU^T gave the rotational matrix, R_t, for the frame t [3].

$$M_{ref}^T \cdot M_t = UDV^T \tag{1}$$

$$R_t = VU^T \tag{2}$$

The markers from the headband were used to ensure good accuracy in the head motion estimation. After compensating for the translation and rotation effects, the remaining motion between frames corresponds to local displacements of the markers, which largely define the subject's facial expressions.

6. Discussion

6.1. IEMOCAP database: advantages and limitations

As mentioned in Section 2, Douglas-Cowie *et al.* defined four main issues that need to be considered in the design of a database: scope, naturalness, context and descriptor [29]. In this section, the IEMOCAP database is discussed in terms of these issues.

Scope: Although the number of subjects suggested by Douglas-Cowie *et al.* is greater than ten [28], the number used in this database may be a sufficient initial step to draw useful conclusions about inter-personal differences. To have this kind of comprehensive data from ten speakers marks a small first, but hopefully an important, step in the study of expressive human communication (e.g., equipment, markers, number of modalities). The detailed motion capture information will be important to better understand the joint role in human communication of modalities such as facial expression, head motion and hand movements in conjunction with the verbal behavior. Also, twelve hours of multimodal data will provide a suitable starting point for training robust classifiers and emotional models.

Naturalness: As mentioned by Douglas-Cowie *et al.*, the price of naturalness is lack of control [28]. The use of the motion capture system imposed even greater constraints on the recording of natural human interaction. In this corpus, this tradeoff was attempted to be balanced by selecting appropriate material to elicit the emotions during dyadic interactions. On the one hand, the linguistic and emotional content was controlled with the use of scripts (for the plays). On the other hand, it is expected that with this social setting, genuine realizations of emotions that are not observed either in monologues or in read speech corpus can be observed. According to [28], this database would be labeled as semi-natural since actors were used for the recording, who may exaggerate the expression of the emotions. However, based on the setting used to elicit the emotions and the achieved results, we consider that the emotional quality of this database is closer to natural than those from prior elicitation settings. As suggested by Cowie *et al.*, we are planning to evaluate the naturalness of the corpus by conducting subjective assessments [23].

Context: One of the problems in many of the existing emotional databases is that they contain only isolated sentences or short dialogs [28]. These settings remove the discourse context, which is known to be an important component [19]. In this corpus, the average duration of the dialogues is approximately 5 minutes in order to contextualize the signs and flow of emotions. Since the material was suitably designed from a dialog perspective, the emotions were elicited with adequate context. The emotional evaluations were also performed after watching the sentences in context so that the evaluators could judge the emotional content based on the sequential development of the dialogs.

Descriptors: The emotional categories considered in the corpus provide a reasonable approximation of the emotions content observed in the database. These emotions are the most common categories found in previous databases. Also, adding the primitive based annotation (valence, activation, and dominance) improves the emotional description of the collected corpus by capturing supplementary aspects of the emotional manifestation (i.e., intensity and variability). Lastly, with the detailed linguistic transcriptions of the audio part of the database, the emotional content can be analyzed in terms of various linguistic levels, in conjunction with the nonverbal cues.

In sum, the IEMOCAP was carefully designed to satisfy the key requirements presented in Section 2. As a result, this database addresses some of the core limitations of the existing emotional databases.

6.2. Open questions suitable for inquiry using IEMOCAP database

The IEMOCAP corpus can play an important role in the study of expressive human communication. In this section, some of the open questions that could be addressed with this corpus are discussed.

Using the IEMOCAP database, gestures and speech from different subjects can be analyzed toward modeling personal styles. For example, by learning inter-personal similarities, speaker-independent emotion recognition systems can be designed (e.g., building models for the features that are emotionally salient across speakers [10]). By using models based on inter-personal differences, human-like facial animation with specific personality can be generated [4].

This corpus is suitable to study the dynamic progression of emotions (especially for the spontaneous scenarios). Since each sentence was emotionally evaluated, it will be interesting to study when the subjects move from one emotion to another and the nature of such audio-visual indicators. From an application point of view, this is an interesting problem, since detecting when a user is changing his/her affective state can be used to improve human-machine interfaces.

This corpus can enable studying the relation between high-level linguistic functions and gestures. For example, one could model gestures that are generated as discourse functions (e.g., head nod for "yes") to improve facial animation [18]. These discourse-based models can be combined with our natural head motion framework to synthesize head motion sequences that respond to the underlying semantic content of what is spoken [8, 7]. Since the data contains spontaneous dialogs and detailed marker information, this corpus is suitable to address these kinds of questions.

This database can also be used to study which areas in the face are used to modulate the affective state of the speakers in a dynamic fashion [9, 11]. Although facial information is obtained from a motion capture system, we hope that the results from the analysis of this data can guide the design of automatic multimodal emotion recognition systems.

The IEMOCAP database was designed for two-person dialogs. Therefore, it is suitable to extend the analysis for dyadic interaction. Active listeners respond with non-verbal gestures that form an important part of the interaction. These gestures appear in specific structures of the speaker's words [30]. This implies that the speech of the active speaker is linked to the listener's gestures, which can be exploited to improve human machine interfaces (e.g., *Virtual Rapport* [35]). We are also interested in analyzing the influence of the gestures of one subject on the behavior of the other subject. For example, this corpus is particularly useful to analyze multimodal cues observed during competitive and cooperative interruptions [45]. With the advances in human-machine interfaces, these studies will play an important role in dialog understanding and user modeling.

These are some of the questions that we plan to explore in our own future research with the IEMOCAP database as the cornerstone resource.

7. Conclusions

This paper presented the *interactive emotional dyadic motion capture database* (IEMOCAP) as a potential resource to expand research in the area of expressive human communication. This corpus provides detailed motion capture information for head, face, and to some extent, the hands in dyadic interactions. In total, ten actors recorded three selected scripts, and dialogs in fictitious scenarios designed to elicit specific emotions (happiness, sadness, anger and frustration). Since the emotions were elicited within the context of discourse, the database provides realizations of more natural expressive interactions, compared to previous elicitation techniques for acted corpora. This database can play an important role in understanding and modeling the relation between different communicative channels used during expressive human communication, and contribute to the development of better human-machine interfaces.

Acknowledgements

This work was supported in part by funds from the National Science Foundation (NSF) (through the Integrated Media Systems Center, an NSF Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award), the Department of the Army, and a MURI award from the Office of Naval Research. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies. The authors wish to thank Joshua Izumigawa, Gabriel Campa, Zhigang Deng, Eric Furie, Karen Liu, Oliver Mayer, and May-chen Kuo for their help and support.

References

1. Abrilian, S., L. Devillers, S. Buisine, and J.C.Martin: 2005, 'EmoTV1: Annotation of Real-life Emotions for the specification of Multimodal Affective Interfaces'. In: *11th International Conference on Human-Computer Interaction (HCI 2005).* Las Vegas, Nevada, USA, pp. 195–200.
2. Amir, N., S. Ron, and N. Laor: 2000, 'Analysis of an emotional speech corpus in Hebrew based on objective criteria'. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.* Newcastle, Northern Ireland, UK, pp. 29–33.
3. Arun, K., T. Huang, and S. Blostein: 1987, 'Least-squares fitting of two 3-D point sets'. *IEEE Trans. Pattern Anal. Mach. Intell.* **9**(5), 698–700.
4. Arya, A., L. Jefferies, J. Enns, and S. DiPaola: 2006, 'Facial actions as visual cues for personality'. *Computer Animation and Virtual Worlds* **17**(3-4), 371–382.

5. Bänziger, T. and K. Scherer: 2007, 'Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus'. In: A. Paiva, R. Prada, and R. Picard (eds.): *Affective Computing and Intelligent Interaction (ACII 2007), Lecture Notes in Artificial Intelligence 4738.* Berlin, Germany: Springer-Verlag Press, pp. 476–487.
6. Batliner, A., K. Fischer, R. Huber, J. Spilker, and E. Nöth: 2000, 'Desperately seeking emotions or: actors, wizards and human beings'. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.* Newcastle, Northern Ireland, UK, pp. 195–200.
7. Busso, C., Z. Deng, M. Grimm, U. Neumann, and S. Narayanan: 2007a, 'Rigid Head Motion in Expressive Speech Animation: Analysis and Synthesis'. *IEEE Transactions on Audio, Speech and Language Processing* **15**(3), 1075–1086.
8. Busso, C., Z. Deng, U. Neumann, and S. Narayanan: 2005, 'Natural Head Motion Synthesis Driven by Acoustic Prosodic Features'. *Computer Animation and Virtual Worlds* **16**(3-4), 283–290.
9. Busso, C., Z. Deng, S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan: 2004, 'Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information'. In: *Sixth International Conference on Multimodal Interfaces ICMI 2004.* State College, PA, pp. 205–211.
10. Busso, C., S. Lee, and S. Narayanan: 2007b, 'Using Neutral Speech Models for Emotional Speech Analysis'. In: *Interspeech 2007 - Eurospeech.* Antwerp, Belgium, pp. 2225–2228.
11. Busso, C. and S. Narayanan: 2006, 'Interplay between linguistic and affective goals in facial expression during emotional utterances'. In: *7th International Seminar on Speech Production (ISSP 2006).* Ubatuba-SP, Brazil, pp. 549–556.
12. Busso, C. and S. Narayanan: 2007a, 'Interrelation between speech and facial gestures in emotional utterances: a single subject study'. *IEEE Transactions on Audio, Speech and Language Processing* **15**(8), 2331–2347.
13. Busso, C. and S. Narayanan: 2007b, 'Joint Analysis of the Emotional Fingerprint in the Face and Speech: A single subject study'. In: *International Workshop on Multimedia Signal Processing (MMSP 2007).* Chania, Crete, Greece, pp. 43–47.
14. Busso, C. and S. Narayanan: 2008a, 'The expression and perception of emotions: Comparing Assessments of Self versus Others'. In: *Interspeech 2008 - Eurospeech.* Brisbane, Australia.
15. Busso, C. and S. Narayanan: 2008b, '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, pp. 17–22.
16. Busso, C. and S. Narayanan: 2008c, 'Scripted dialogs versus improvisation: Lessons learned about emotional elicitation techniques from the IEMOCAP database'. In: *Interspeech 2008 - Eurospeech.* Brisbane, Australia.
17. Caridakis, G., L. Malatesta, L. Kessous, N. Amir, A. Raouzaiou, and K. Karpouzis: 2006, 'Modeling naturalistic affective states via facial and vocal expressions recognition'. In: *Proceedings of the 8th international conference on Multimodal interfaces (ICMI 2006).* Banff, Alberta, Canada, pp. 146–154.
18. Cassell, J., T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjalmsson, and H. Yan: 1999, 'Embodiment in Conversational Interfaces: Rea'. In:

International Conference on Human Factors in Computing Systems (CHI-99). Pittsburgh, PA, USA, pp. 520–527.

19. Cauldwell, R.: 2000, 'Where did the anger go? The role of context in interpreting emotion in speech'. In: *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*. Newcastle, Northern Ireland, UK, pp. 127–131.
20. Clavel, C., I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette: 2006, 'The SAFE Corpus: illustrating extreme emotions in dynamic situations'. In: *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*. Genoa, Italy, pp. 76–79.
21. Cohn, J., L. Reed, Z. Ambadar, J. Xiao, and T. Moriyama: 2004, 'Automatic analysis and recognition of brow actions and head motion in spontaneous facial behavior'. In: *IEEE Conference on Systems, Man, and Cybernetic*, Vol. 1. The Hague, the Netherlands, pp. 610–616.
22. Cowie, R. and R. Cornelius: 2003, 'Describing the emotional states that are expressed in speech'. *Speech Communication* **40**(1-2), 5–32.
23. Cowie, R., E. Douglas-Cowie, and C. Cox: 2005, 'Beyond emotion archetypes: Databases for emotion modelling using neural networks'. *Neural Networks* **18**(4), 371–388.
24. Cowie, R., E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor: 2001, 'Emotion recognition in human-computer interaction'. *IEEE Signal Processing Magazine* **18**(1), 32–80.
25. Cronbach, L.: 1951, 'Coefficient alpha and the internal structure of tests'. *Psychometrika* **16**, 297–334.
26. De Silva, L., T. Miyasato, and R. Nakatsu: 1997, 'Facial emotion recognition using multi-modal information'. In: *International Conference on Information, Communications and Signal Processing (ICICS)*, Vol. I. Singapore, pp. 397–401.
27. Devillers, L., L. Vidrascu, and L. Lamel: 2005, 'Challenges in real-life emotion annotation and machine learning based detection'. *Neural Networks* **18**(4), 407–422.
28. Douglas-Cowie, E., N. Campbell, R. Cowie, and P. Roach: 2003, 'Emotional speech: Towards a new generation of databases'. *Speech Communication* **40**(1-2), 33–60.
29. Douglas-Cowie, E., L. Devillers, J. Martin, R. Cowie, S. Savvidou, S. Abrilian, and C. Cox: 2005, 'Multimodal Databases of Everyday Emotion: Facing up to Complexity'. In: *9th European Conference on Speech Communication and Technology (Interspeech'2005)*. Lisbon, Portugal, pp. 813–816.
30. Ekman, P.: 1979, 'About brows: emotional and conversational signals'. In: M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog (eds.): *Human ethology: claims and limits of a new discipline*. New York, NY, USA: Cambridge University Press, pp. 169–202.
31. Ekman, P. and W. Friesen: 1971, 'Constants across cultures in the face and emotion'. *Journal of Personality and Social Psychology* **17**(2), 124–129.
32. Enos, F. and J. Hirschberg: 2006, 'A Framework for Eliciting Emotional Speech: Capitalizing on the Actors Process'. In: *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*. Genoa,Italy, pp. 6–10.
33. Fischer, L., D. Brauns, and F. Belschak: 2002, *Zur Messung von Emotionen in der angewandten Forschung*. Pabst Science Publishers, Lengerich.

34. Fleiss, J.: 1981, *Statistical methods for rates and proportions.* New York, NY, USA: John Wiley & Sons.
35. Gratch, J., A. Okhmatovskaia, F. Lamothe, S. Marsella, M. Morales, R. van der Werf, and L. Morency: 2006, 'Virtual Rapport'. In: *6th International Conference on Intelligent Virtual Agents (IVA 2006).* Marina del Rey, CA, USA.
36. Grimm, M. and K. Kroschel: 2005, 'Evaluation of natural emotions using self assessment manikins'. In: *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2005).* San Juan, Puerto Rico, pp. 381–385.
37. Grimm, M., K. Kroschel, E. Mower, and S. Narayanan: 2007, 'Primitives-based evaluation and estimation of emotions in speech'. *Speech Communication* **49**(10-11), 787–800.
38. Grimm, M., K. Kroschel, and S. Narayanan: 2008, 'The Vera AM Mittag German audio-visual emotional speech database'. In: *IEEE International Conference on Multimedia and Expo (ICME 2008).* Hannover, Germany, pp. 865–868.
39. Hall, E.: 1966, *The hidden dimension.* New York, NY, USA: Doubleday & Company.
40. Huang, X., F. Alleva, H.-W. Hon, M.-Y. Hwang, K.-F. Lee, and R. Rosenfeld: 1993, 'The SPHINX-II speech recognition system: an overview'. *Computer Speech and Language* **7**(2), 137–148.
41. Humaine project portal: 2008. http://emotion-research.net/. Retrieved September 11th, 2008.
42. Kapur, A., A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. Driessen: 2005, 'Gesture-Based Affective Computing on Motion Capture Data'. In: *1st International Conference on Affective Computing and Intelligent Interaction (ACII 2005).* Beijing, China, pp. 1–8.
43. Kipp, M.: 2001, 'ANVIL - A Generic Annotation Tool for Multimodal Dialogue'. In: *European Conference on Speech Communication and Technology (Eurospeech).* Aalborg, Denmark, pp. 1367–1370.
44. Lee, C. and S. Narayanan: 2005, 'Toward detecting emotions in spoken dialogs'. *IEEE Transactions on Speech and Audio Processing* **13**(2), 293–303.
45. Lee, C.-C., S. Lee, and S. Narayanan: 2008, 'An Analysis of Multimodal Cues of Interruption in Dyadic Spoken Interactions'. In: *Interspeech 2008 - Eurospeech.* Brisbane, Australia.
46. Pandzic, I. and R. Forchheimer: 2002, *MPEG-4 Facial Animation - The standard, implementations and applications.* John Wiley & Sons.
47. Picard, R. W.: 1995, 'Affective Computing'. Technical Report 321, MIT Media Laboratory Perceptual Computing Section, Cambridge, MA,USA.
48. Scherer, K. and G. Ceschi: 1997, 'Lost Luggage: A Field Study of EmotionAntecedent Appraisal'. *Motivation and Emotion* **21**(3), 211–235.
49. Scherer, K., H. Wallbott, and A. Summerfield: 1986, *Experiencing emotion: A cross-cultural study.* Cambridge, U.K.: Cambridge University Press.
50. Schiel, F., S. Steininger, and U. Türk: 2002, 'The SmartKom Multimodal Corpus at BAS'. In: *Language Resources and Evaluation (LREC 2002).* Las Palmas, Spain.
51. Steidl, S., M. Levit, A. Batliner, E. Nöth, and H. Niemann: 2005, '"Of all things the measure is man" automatic classification of emotions and interlabeler consistency'. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, Vol. 1. Philadelphia, PA, USA, pp. 317–320.

52. Tekalp, A. and J. Ostermann: 2000, 'Face and 2-D Mesh animation in MPEG-4'. *Signal Processing: Image Communication* **15**(4), 387–421.

53. Ubiqus: 2008. http://www.ubiqus.com/. Retrieved September 11th, 2008.

54. Ververidis, D. and C. Kotropoulos: 2003, 'A State of the Art Review on Emotional Speech Databases'. In: *First International Workshop on Interactive Rich Media Content Production (RichMedia-2003).* Lausanne, Switzerland, pp. 109–119.

55. Vicon Motion Systems Inc: 2008, 'VICON iQ 2.5'. http://www.vicon.com/. Retrieved September 11th, 2008.

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

57. Zara, A., V. Maffiolo, J. Martin, and L. Devillers: 2007, 'Collection and Annotation of a Corpus of Human-Human Multimodal Interactions: Emotion and Others Anthropomorphic Characteristics'. In: A. Paiva, R. Prada, and R. Picard (eds.): *Affective Computing and Intelligent Interaction (ACII 2007), Lecture Notes in Artificial Intelligence 4738.* Berlin, Germany: Springer-Verlag Press, pp. 464–475.