Lang Resources & Evaluation (2016) 50:497–521
DOI 10.1007/s10579-015-9300-0

ORIGINAL PAPER

The USC CreativeIT database of multimodal dyadic interactions: from speech and full body motion capture to continuous emotional annotations

Angeliki Metallinou[1] · Zhaojun Yang[2] · Chi-chun Lee[3] · Carlos Busso[4] · Sharon Carnicke[2] · Shrikanth Narayanan[2]

Published online: 17 April 2015

Abstract Improvised acting is a viable technique to study expressive human communication and to shed light into actors' creativity. The USC CreativeIT database provides a novel, freely-available multimodal resource for the study of theatrical improvisation and rich expressive human behavior (speech and body language) in dyadic interactions. The theoretical design of the database is based on the well-established improvisation technique of Active Analysis in order to provide naturally induced affective and expressive, goal-driven interactions. This database contains dyadic theatrical improvisations performed by 16 actors, providing detailed full body motion capture data and audio data of each participant in an interaction. The carefully engineered data collection, the improvisation design to elicit natural emotions and expressive speech and body language, as well as the well-developed annotation processes provide a gateway to study and model various aspects of theatrical performance, expressive behaviors and human communication and interaction.

Keywords Multimodal database · Theatrical improvisations · Motion capture system · Continuous emotion

✉ Zhaojun Yang
zhaojuny@usc.edu

Angeliki Metallinou
ametallinou@gmail.com

[1] Amazon Lab 126, Cupertino, CA 95014, USA

[2] University of Southern California, Los Angeles, CA 90089, USA

[3] National Tsing Hua University, Hsinchu, Taiwan

[4] University of Texas at Dallas, Richardson, TX 75080, USA

1 Introduction

Human interaction results from a complex interplay of communicative intent and goals, emotions and stance assumed, which are expressed and revealed through rich verbal and nonverbal behavior of body language, prosodic cues and spoken language. The study of human communication and expressive behaviors has attracted interest from multiple diverse domains including psychology, social and behavioral sciences, engineering, theater arts, and health care. This paper describes the design, collection and annotation process of a novel, multimodal database of dyadic interactions, carried out within a multidisciplinary setting. It is a result of the collaborative work between engineering, computer science and theater experts (at the University of Southern California, USC). This corpus consists of improvised dyadic interactions performed by pairs of actors. The multimodal database was collected using cameras, microphones and motion capture systems, thereby containing detailed audio-visual information of the body language and spoken language cues of both actors. The database which is transcribed and annotated with emotional information is being made freely available to the research community. This database was designed to serve two primary purposes. First, it aims to provide insights into the creative and cognitive processes of actors during theatrical improvisation. Second, it offers a carefully-designed and well-controlled opportunity to study and model expressive affective behaviors and natural human interaction.

The significance of studying improvisation in theater performance is that it is a form of real-time dynamic problem solving (Mendonca and Wallace 2007); it offers a window into the human creative processes in an interactive setting. Improvisation is a creative group performance where actors collaborate and coordinate in real-time to create a coherent viewing experience (Johnstone 1981). Improvisation may include diverse methodologies with variable levels of rules, constraints and prior knowledge, with respect to the script and the actor’s activities. Active Analysis, pioneered by Stanislavsky, introduces a goal-driven performance to elicit natural affective behaviors and interactions (Carnicke 2009). It is the primary acting technique utilized in the creation of the database described in this paper, providing a systematic way to investigate the creative processes that underlie improvisation in theater performance.

Improvised acting has been considered as a viable research methodology for studying human emotions and communication (Bänziger and Scherer 2007). Theater has been suggested as a model for believable agents; agents that may display emotions, intents and human behavioral qualities (Perlin and Goldberg 1996). Researchers have advocated the use of improvisation as a tool for eliciting naturalistic affective behavior for studying emotions and have also argued that improvised performances resemble real-life decision making (Busso and Narayanan 2008; Wallbott and Scherer 1986). The acting methodology can provide stable, well-controlled and repeatable emotion expressions which have been considered plausible and believable prototypes of a given emotion (Enos and Hirschberg 2006; Scherer et al. 2010). Furthermore, it has been suggested that experienced actors engaged in roles during dramatic interactions may provide a more natural representation of emotions, avoiding exaggeration or caricatures, compared to non-skilled actors or non-actors (Douglas-Cowie et al. 2003).

The USC CreativeIT database is a result of a close collaboration between theater experts, actors and engineers (Metallinou et al. 2010). Its theoretical design is based on the well established theatrical improvisation technique of Active Analysis. According to this technique, the interactions are goal-driven; the pair of actors in each interaction has a pair of predefined interaction goals, e.g., *to approach* versus *to avoid*, which they try to achieve through the appropriate use of body language, spoken language and vocal prosody. The predefined goals help to induce genuine realizations of emotions. In the collection of the database presented, we used Motion Capture technology to obtain detailed body language information of the actors, in addition to the use of microphones, video cameras and carefully designed post-performance interviews of the participants. Annotation of the data includes continuous emotional descriptors (activation, valence, dominance) as well as theatrical performance ratings. The database aims to facilitate the qualitative study of creative theatrical improvisation and to support advanced research on expressive human-human communicative interaction. It provides a valuable source to enhance the development of various applications such as multimodal emotion recognition, affective human gesture synthesis, and virtual human-machine interaction systems, e.g., Metallinou et al. (2013), Yang et al. (2014a, b, c).

The rest of this paper is organized as follows. Section 2 describes existing multimodal human interaction databases. Section 3 introduces the theatrical methodology and the design of the corpus presented in this paper. Section 4 describes the data collection process including multimodal data recording, post-processing and multimodal data stream synchronization. Section 5 describes the details of the data annotation process as well as annotation results, followed by database analysis in Sect. 6. We discuss current and possible future research directions with this database in Sect. 7. This paper ends with conclusions in Sect. 8.

2 Existing multimodal databases

In recent years, a number of multimodal emotion databases have been developed for supporting advanced research on affective analysis and computing. Douglas-Cowie et al. (2003) have comprehensively reviewed the development of emotional speech databases in Douglas-Cowie et al. (2003) from the perspective of scope, naturalness, context and descriptors. The VAM database presented in Grimm et al. (2008) containing the audio and video recordings of a TV talk show provides manual annotations for emotion attributes of activation, valence and dominance at the utterance level. It has been widely applied in the research field of automatic emotion recognition (Kanluan et al. 2008; Wu et al. 2011). The SEMAINE multimodal database consists of emotional conversations between human subjects and the computer conversational agent (proxy) SAL (McKeown et al. 2012). This database has also been incorporated in the HUMAINE corpus[1] which aims at providing diverse types of emotional data (naturalistic and induced) for research and development (Douglas-Cowie et al. 2007). The AFEW database described in Dhall et al. (2012) includes affective facial expressions extracted from movies. More recently, databases, such as DEAP (Koelstra et al. 2012), have been created with physiological signals of participants watching emotional videos. The MAHNOB-HCI database also includes other synchronized multimodal signals of eye gaze data, audio recordings as well as face videos (Soleymani et al. 2012).

As observed in Enos and Hirschberg (2006), valuable databases can be recorded by eliciting emotions from professional actors using goal-based and context-based theatrical techniques. A variety of acted databases have been collected in controlled settings for the study of emotion expressions. The Geneva Multimodal Emotion Portrayal (GEMEP) database was created using actor portrayal study and contains more than 7000 video-audio emotional sentences (Bänziger and Scherer 2007). To induce natural emotions, contextualized acting was applied in Anolli et al. (2005) for the collection of an audio–video database. Sneddon et al. (2012) collected induced natural emotion data of mild to moderate intensity levels from people with diverse cultures.

Since humans generally use non-verbal behavior to convey emotions, it is important to collect databases capturing gesture information, such as facial expressions, body gestures and postural cues, associated with audio data. Motion capture (MoCap) techniques, which provide much more detailed and accurate 3D gesture description than the state of art in video processing, have been applied to create multimodal corpora. Full body gestural data collected during the expression of four basic categorical emotions using Vicon systems (Kapur et al. 2005) underscored the early promise of the methodology. McKeown et al. (2013) devised laughter induction techniques to capture body movements related to laughter. To explore the affective content in body movement, Crane and Gross used an autobiographical memories paradigm to elicit emotions from participants while recording whole body motion data (Crane and Gross 2007). However, emotions are not only manifested by the multimodal communicative channels of an individual but also influenced by the social context, such as emotions, attitudes or communication goals and stance of one's interlocutor in a social interaction. To this end, Busso et al. (2008) had designed the interactive emotional dyadic motion capture database (IEMOCAP), which contains improvised and scripted dyadic interactions in the form of audio-visual data as well as Motion Capture data for facial expressions. A corpus of student-computer interactions has also been created in Grafsgaard et al. (2012), including Kinect videos and depth images of posture and hand-to-face gestures.

However, the above-mentioned databases are restricted to either only audio-video recordings, or part-body movements, such as facial expressions or hand gestures, or simple single-subject scenarios. The CreativeIT database is a novel, multimodal corpus that is distinct from, and complements, most of the existing corpora. The theoretical design of the database is based on the well-established improvisation technique of Active Analysis in order to provide naturally induced affective and

[1] HUMAINE is freely available at http://humaine-db.sspnet.eu/.

expressive, goal-driven interactions. This database provides audio-video recordings as well as detailed full body motion capture data of both participants in the dyadic interaction, as well as rich emotion annotations (continuous and discrete). The main characteristics of the CreativeIT database are summarized as follows:

- It uses formal acting techniques systematically for eliciting realistic emotional displays and expressive human behavior (speech and body language).
- It provides goal-driven improvised dyadic interactions with synchronized audio, video and detailed Motion Capture data on expressive full body movements.
- It contains rich annotation data on continuous emotional attributes and theatrical performance.
- The CreativeIT database is being made available freely to the research community.

3 Database design

3.1 Active analysis

The formal design of the CreativeIT database is based on the theatrical improvisation technique of Active Analysis pioneered by Stanislavsky (Carnicke 2009). In Active Analysis, the actors play conflicting forces that jointly interact. The balance of the forces determines the direction of the play. The scripts used in the case play the role of guiding the events (skeleton). The course of the play can be close or different from the script. This degree of freedom provides flexibility to work at different levels in the improvisation spectrum. A key element in Active Analysis is that actors are asked to keep a verb in their mind, while they are acting, which drives their actions. Therefore, the interaction and behavior of the actors may be more expressive and closer to being natural, which is crucial in the context of developing robust automatic emotion recognition systems. For instance, if the play suggests a confrontation between two actors, one of them may choose the verb *inquire* while the other may choose *evade*. If the verbs are changed (e.g., *persuade* vs. *confront*) the play will have a different development. By changing the verbs, the intensity of the play can be modified as well (i.e., *ask* vs. *interrogate*). As a result, different manifestations of communication goals, emotions and attitudes, such as in speech content, speech prosody and body language, can be naturally elicited through the course of the interaction. This flexibility allows us to explore the improvisation spectrum at different levels and makes Active Analysis a suitable technique to elicit expressive emotional manifestations.

3.2 Design of data collection

The USC CreativeIT database utilizes two different theatrical approaches, the two-sentence exercise and the paraphrase, both of which originate from the Active Analysis methodology. A post-performance interview was also performed after the recording.

In the two-sentence exercise, each actor is restricted to saying a particular sentence with a given verb. For example, one actor may say *Marry Me* with verb *confront*, and another one may say *I'll think about it* with verb *deflect*. Given the lexical constraint, the expressiveness and the flow of the play are primarily based on the prosodic and nonverbal behaviors of the actors. This type of controlled interaction can bring insights into how human/actors use their expressive behaviors, such as body language and prosody, to realize a communication goal. In addition, this approach is suitable for studying emotion modulation at a semantic level, since the same sentences are repeated different times with different emotional connotations.

In the paraphrase setting, the actors are asked to perform without a given script but with a known skeleton of a theme by using their own words and interpretation. Example plays that are used for paraphrase performance are *The Proposal* by Chekhov or *Taming of the Shrew* by Shakespeare. In this set of recordings, actors are not lexically constrained, resulting in a performance that is characterized by a more natural and free-flowing spoken interaction between the actors. Hence, the paraphrase interaction bears more resemblance to real-life interaction scenarios, compared to the two-sentence exercise. Behavioral analysis and findings on such sessions could possibly be extrapolated to natural human interaction and communication. An example paraphrase performance of a scene of *The Proposal* is shown in Table 1. For original lines from *The Proposal*, we refer interested readers to http://www.one-act-plays.com/comedies/proposal.html.

Finally, a brief interview of the actors is performed right after each performance. Examples of the questions asked are '*What verbs did you and the other actor use?*', '*What was the goal of your character?*', '*How would you describe your and the other actor's emotion during the interaction?*'. These questions are designed to help understand the cognitive planning process of the actors as they improvise on the scenes.

3.3 Session protocol

An expert on Active Analysis (professor at the USC School of Dramatic Arts, co-author SC) directed the actors during the rehearsal and the recording of these sessions. Prior to the scheduled data collection date, the actors had to go through a rehearsal with the director to become familiar with active analysis and the scene. Just before the recording of the paraphrase, there was another 5-min session to refresh actors' memory and to give the director a chance to remind the actors about the essence of the script.

The data collection protocol consists of the following steps:

1. Two-sentence exercise (unknown verbs, i.e., the actors were not privy to each other's verbs)
2. Two-sentence exercise, using the same sentences as previously but different verbs (known verbs)
3. Paraphrase of script (known verbs)

Table 1 An example of part of paraphrase performance about a scene of *The Proposal* by Anton Chekhov

Time (s)	Turn	Transcription
[12.4–17.1]	F1:	I am so sorry. You're totally right. It was your field
[17.2–18.7]	M1:	I am just, I mean, I am sorry, I got to collect myself, it is, I mean it is my field You understand, I mean I have documentation all down the family line
[18.8–23.4]	F2:	[SIGH] Okay, okay. [SILENCE] Ab, absolutely
[23.4–23.8]	M2:	Okay, I am glad
[23.9–29.1]	F3:	Of course. Of course I understand. I'm so sorry. Have a seat. I'm so sorry
[29.2–29.7]	M3:	I am fine here, thank you
[29.7–32.7]	F4:	Oh, okay. Um, [SILENCE] let's just
[32.6–33.1]	M4:	I mean you understand, it is my field, it is you
[33.2–35.4]	F5:	No. Of course. I mean, let's change the subject. Let's talk about something else
[35.4–36.9]	M5:	Okay, that is just fine
[37.0–39.4]	F6:	All right, uh, are you going hunting soon?
[39.3–41.5]	M6:	I am. I am going hunting for woodchucks but would you now know
[41.7–42.4]	F7:	Oh!
[42.6–46.6]	M7:	The worst thing has happened. You are familiar with my dog, Diviner?
[46.6–48.0]	F8:	Uh, I remember him, yeah
[48.1–51.6]	M8:	Well, unfortunately he has gone lame
[51.4–53.9]	F9:	What a surprise! No, no, you don't say
[54.0–62.0]	M9:	Yeah, it is actually, it is terrible. He was a wonderful dog [LAUGHTER] I got a steal on him. 125 rubles, do you believe that?
[62.1–62.8]	F10:	Excuse me?
[62.9–64.8]	M10:	It was 125 rubles, the best bargain I have ever had in my life
[65.0–69.0]	F11:	I'm sorry, I'm sorry. One more time, 125 rubles?
[69.1–69.8]	M11:	Yes

We also display sigh, laughter and silence at proper location in this transcription example, while their exact time stamps are annotated separately

4. Paraphrase of script, using the same script as previously but different verbs (known verbs)
5. Two-sentence exercise (unknown verbs)
6. Two-sentence exercise, using the same sentences as previously but different verbs (known verbs)

Verbs are chosen by the director prior to each performance. The verb pairs for interactions recorded in the database are summarized in Table 2, which introduce a large variety of communication goals. Unknown verbs indicate that actors are not aware of each other's verb prior to the performance. This setting provides a variety in the interaction dynamics of the two-sentence exercise. During the paraphrases, the verbs of actors are always known to each other in advance.

Table 2 Pairs of verbs chosen for each performance

to deflect–to pry
to seduce–to play the victim
to peal–to share excitement
to make her understand–to hold on to her
to unload on–to pry
to excite–to tease
to seek comfort–to push away
to put her down–to tease
to fix–to shut out
to instigate–to parry
to coddle–to stand up to him
to escape–to take care of him
to console–to annoy
to demand permission–to seduce
to bully–to observe
to make peace–to comfort
to deflect–to reprimand
to embarrass her–to hold out
to dig–to shut him out
to brace–to break down the wall
to shut her off–to seduce
to control–to patronize
to mock–to befriend
to seduce–to tease
to excite–to emasculate
to put her down–to force to surrender
to make her feel sorry–to seduce
to deflect–to force the issue
to confront–to reject
to put her off–to charm
to comfort–to get him to leave
to annoy–to provoke
to avoid–to provoke
to prove–to seduce
to seduce–to think it over
to reject–to convince
to persuade–to convince
to trap–to circle
to fight back–to accuse
to appease–to seduce
to seduce–to annoy
to seduce–to dodge
to trick–to belittle
to make her understand–to play
to shut her down–to bully
to try to understand–to piss her off

4 Data collection

In each session, each actor wore a special body suit and 45 markers were placed across his/her body, as illustrated in Fig. 1. A Vicon motion capture system consisting of 12 cameras retrieved the (x, y, z) positions of these markers at 60 frames per second. The performance of actors was also recorded by two HD Sony video cameras. These videos are used for annotation of emotion descriptors and performance ratings, which is described in Sect. 5. The Vicon MoCap cameras were located on a specially mounted frame close to the ceiling of the recording room, and the two HD cameras were placed at each corner of the room. The actors were instructed to gesture as naturally as possible and to stay within the field view of the Vicon and HD cameras. Since these professional actors are trained to work under different conditions as well as to be minimally affected by the particularities of the performance environment, wearing these markers did not appear to interfere with their natural performance. The audio data were simultaneously recorded through close-up microphones at 48 kHZ with 24 bits. Professional transcriptions of the audio data were also obtained from Ubiqus (2014, http://www.ubiqus.com). An example transcription is shown in Table 1. For each audio recording, we manually annotated four types of vocal behavior (verbal and nonverbal): speech, laughter,

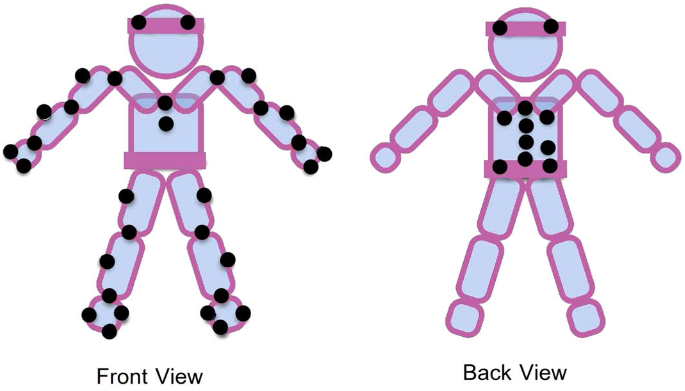


Fig. 1 Positions of motion capture (mocap) *markers* placed on the actors

sigh and silence. The annotations of speech and silence can help to segment the interactions into utterances, and the labels of laughter and sigh can bring insights about the role that nonverbal vocalization plays during human interaction. The four types of vocal behavior labels were not annotated directly in the transcriptions, but in separate files showing only the exact time stamps for each label. For better illustration, we also presented the annotations of sigh, silence and laughter at proper location in the example transcription in Table 1. The database consists of the recording of eight full sessions, each of which contains approximately one hour of audiovisual data. In total we have recorded 33 two-sentence exercises and 17 paraphrases performed by 16 actors (9 of whom are female).

After recording the sessions, we mapped each of the captured markers into the defined body model of a subject in a semi-automatic manner to reconstruct the trajectories of the markers. Since actors were asked to be expressive and natural with body language and gesture, occlusion of markers happens fairly often. Because of this, the computer software was unable to perform all the labeling automatically and accurately. For example, when two subjects are close to one another, one's hand marker may be labeled as the other person's shoulder if we rely solely on computer labeling. In order to obtain reliable and detailed marker information, the motion capture data were manually corrected frame by frame. The spline function was used to interpolate any missing markers. For each interaction performance, there are two subjects each with 45 markers, as well as around 5000–10,000 frames. Such post-processing of one actor in one performance may require approximately 1–2 h, which is a fairly time consuming task.

The multimodal data recordings, i.e., audio, video and MoCap data, were synchronized using a clapboard. Two markers were placed on the clapboard. A recording monitor clapped the clapboard to indicate the beginning of a performance. We marked the MoCap frame when the two markers on the clapboard 'coincide'. The clapping also resulted in a clap spike in the microphone audio data as well as a spike in the audio data extracted from the video file. In this way, the multimodal data recordings can be accurately synchronized based on the spike of the microphone audio data, the spike of the audio data from the video file and the marked MoCap frame in the MoCap data. At the beginning of the scene, one of the actors (typically the female) raised her hand to indicate 'speaker1' corresponding to the first 45 markers of the MoCap file. Accordingly, the other actor is 'speaker2' corresponding to the second 45 markers. This information is useful for post-processing the MoCap data files.

5 Data annotation

5.1 Continuous data annotation

The CreativeIT database contains a variety of multimodal expressions and interaction dynamics that continuously unfold during the improvisation. Therefore, it is difficult to define precise starting and ending times of expressions since those are produced multimodally, or to segment interactions into temporal units of homogenous emotional content. In unimodal databases, or databases that are spoken-dialog centric such as IEMOCAP (Busso et al. 2008) and VAM (Grimm et al. 2008), it seems natural to segment a conversation into utterances as basic units for examining emotional content. In contrast, the CreativeIT database contains many nonverbal emotional expressions that happen asynchronously to speech or when the participant is silent. Such observations motivate the use of continuous attributes as a natural way to describe the emotional flow of an interaction (Metallinou and Narayanan 2013).

The perceived emotional state for each participant was annotated in terms of the widely used dimensional attributes of activation, valence and dominance. This emotion representation is well-suited to describe the complex and ambiguous manifestations of the CreativeIT database, since these multimodal manifestations do not always have clear categorical descriptions. In addition, it has been pointed out that these dimensional representations can provide more robust agreement than other descriptors (Cowie et al. 2012). We used the Feeltrace software (Cowie et al. 2000) for annotations. Feeltrace incorporates both audio and visual inputs so that raters can comprehensively perceive actors' emotions. We collected annotations of perceived activation, valence and dominance for each actor at each frame in each performance. The annotations of the emotion attributes take continuous values in [−1, 1].

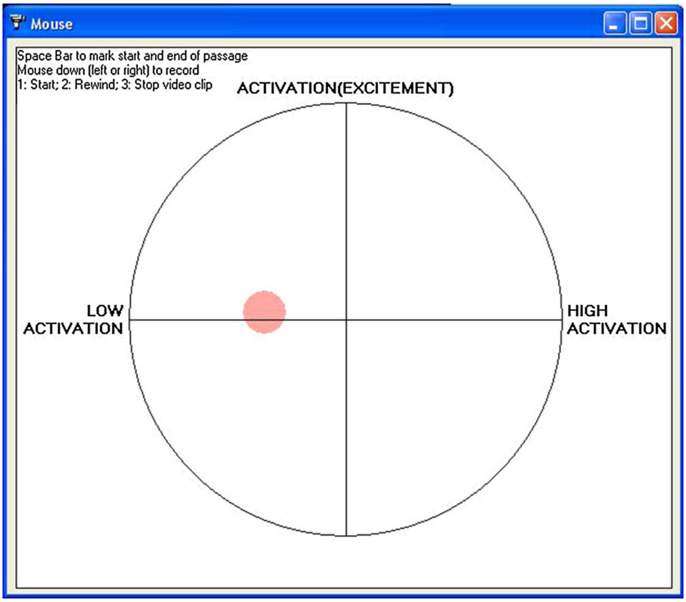
Annotation of emotional content is an inherently subjective task that depends, among others, on the individual's perception, experiences and cultural background. The use of continuous descriptors seems to increase the level of complexity of the emotional annotation task, as it requires a higher amount of attention and cognitive processing compared to non real-time, discrete annotation tasks. Apart from being a strenuous and time-consuming process, continuous annotation poses challenges in terms of obtaining inter-annotator agreement, as has been reported by several researchers. In Devillers et al. (2006) authors report that in 70 % of continuous valence annotations of TV clips, the inter-annotator correlations are above the 0.5 threshold, a percentage that reduces to 55 and 34 % for activation and dominance, respectively. In Malandrakis et al. (2011) authors report mean annotator correlations of 0.3 and 0.4 for valence and activation respectively for continuous self-annotations of perceived emotion while watching movies. Our pilot annotation of a subset of CreativeIT data resulted in median annotator correlations of around 0.5 for the three dimensional attributes (Metallinou et al. 2011).

For the annotation of the CreativeIT data, we recruited psychology students, most of whom have had previous experience in emotional annotation and were committed to weekly working requirements. Since the dimensional emotional attributes of interest are less intuitive for some annotators, compared to the categorical emotions, the definitions of activation, valence and dominance attributes were explained through examples. We clarified that ratings are subjective, however annotators should be able to rationally explain their decisions based on verbal or nonverbal characteristics of the interaction. Before performing annotation, annotators were trained on how to use Feeltrace. They performed their first annotations multiple times to familiarize themselves with the software, and were later encouraged to perform each annotation as many times as needed until they were satisfied with the result. In order to facilitate the annotation process, we wanted annotators to be familiar with the type and range of emotional manifestations that appear in the database. Therefore, as part of their training, they had to watch in advance about a fifth of the recorded performances, randomly selected across different sessions.

Since continuous annotations are performed in real-time, we expect subject-dependent delays due to perceptual processing, between the time when an event happens and when its emotional content is annotated. In order to reduce such delays, we modified the Feeltrace interface so that annotators can focus on one attribute each time, rather than two attributes that were in the original design of Feeltrace. The modified Feeltrace interface for activation annotation is presented in Fig. 2. The annotation is performed by moving the mouse, shown as a full red circle, along the horizontal line, while watching the performance video in a separate window. It is interesting to note that a one-dimensional version of the Feeltrace interface later became available (software Gtrace Cowie and Sawey 2011), indicating the general need for such a one-dimensional annotation tool. Finally, to further reduce delays due to perceptual processing, we also instructed annotators to watch each video multiple times so that they have a clear idea of the emotional content before starting the real-time annotation. It is noteworthy that though Fealtrace is a popular tracing tool, there still are some research questions need to be answered. For example, how the time delays between videos and the actual annotator responses depend on emotion dimensions? We believe our annotation data is suitable for these studies.

5.2 Discrete data annotation

We also collected discrete annotations of global emotional content of each performance. Emotional content was rated in terms of perceived activation, valence and dominance for each actor on a 9-point scale. Rating 1 denotes the lowest possible activation level, the most negative valence level, and the most submissive dominance level. Rating 9 indicates the highest possible activation level, most



Space Bar to mark start and end of passage
Mouse down (left or right) to record
1: Start; 2: Rewind; 3: Stop video clip

Fig. 2 Screenshot of the modified Feeltrace interface

positive valence level and most dominant level. Annotators were asked to give an overall rating that summarizes the particular attribute over the total recording. They were instructed to perform the overall rating right after completing the corresponding continuous annotation, such that they would have a recent impression of the annotated performance and their continuous assessment of its emotional content. The reason for collecting global annotations is two-fold: firstly we wanted to enrich our annotation with more standard discrete labels for potential future use; secondly, we want to study relations between global discrete and detailed continuous ratings provided by the same person, in order to shed light into the way humans summarize a perceived emotional experience. Understanding such perception mechanism is essential and relevant for behavioral science where human assessment is the core approach for various analyses (Humphrey 1924; Lindahl 2001).

5.3 Theatrical performance annotation

Apart from the annotations of activation, valence and dominance, the raters were also asked to evaluate how successful the actors were in conveying their target verbs. Specifically, the raters were told what the verb was for each actor in each recording, and rated a score ranging from one to nine for the actor's performance. This rating can be seen as a measure of the quality of the theatrical performance, which can facilitate future theatrical performance analysis.

5.4 Annotation results

The database contains 50 recordings, each rated for both actors in the dyad, therefore we have 100 actor-recordings. Seven annotators participated in total, rating overlapping portions of the database, so that each actor-recording could be rated by three or four annotators (88 out of the 100 actor-recordings were rated by three annotators). The resulting continuous annotations were pre-processed by low-pass filtering to remove high frequency noise artifacts. We used a equiripple low-pass filter with pass-band frequency at $\frac{1}{10}$ in normalized frequency, the stop-band frequency at $\frac{1}{5}$ in normalized frequency, the maximum magnitude in pass-band of 1 dB, and the attenuation in the stop-band of 60 dB. These parameters were chosen experimentally, after careful visual observation, since they aim to only remove minor noise in the data and perform slight smoothing. This section describes analysis of the annotation results that we obtained.

5.4.1 Interrater agreement

Evaluator agreement is a straightforward concept when dealing with discrete labels. For example we can say that two annotators agree if they choose the same label. For continuous annotations this concept becomes less straightforward. To choose an agreement metric, it is important to understand how raters behave when rating continuous attributes. Figure 3 shows an example of activation annotations by three annotators for the same actor-recording, and the average annotation over the three annotators. Although annotators agree on the trends of the activation curve (mean

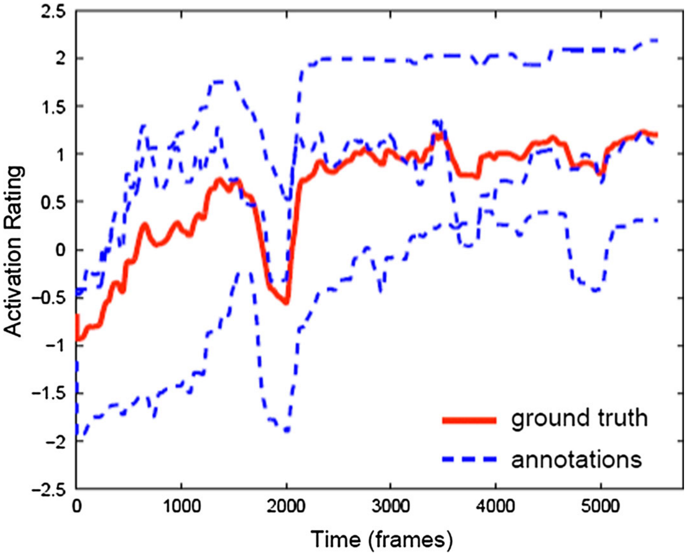


Fig. 3 Example of activation ratings by three annotators. Each *dashed curve* represents ratings from a different annotator. The *solid ground truth curve* is obtained from an annotator subset that yielded linear correlations greater than a specified threshold

Table 3 Measures of agreement of continuous ratings (Pearson's correlation) and discrete ratings (Cronbach's α) for activation, valence and dominance (without annotator selection)

Activation	Valence	Dominance
Continuous rating (mean Pearson's correlation)		
0.48	0.48	0.37
Discrete rating (Cronbach's α)		
0.72	0.78	0.67

correlation of 0.67), and recognize pronounced activation events, they do not agree on the actual activation values. Similar observations hold true for several of our obtained annotations. This suggests that people agree more when describing emotions in relative terms, e.g., whether there has been an increase or decrease, rather than in absolute terms. Rating perceived emotions in absolute terms seems more challenging because of individual-dependent internal scale when assessing an emotional experience. This motivates us to focus on the annotation trends, and to use correlation metrics, such as Pearson's correlation, to measure evaluator agreement. We can also observe in Fig. 3 that the interrater agreement on annotation trend differs over different time intervals. For example, the annotations agree well in the beginning of the interaction, but a higher interrater divergence appears after the valley at frame 2000 in the example shown. This observation indicates that our annotation data might be suitable for analyzing discernible patterns of agreement by defining distinct regions according to different agreement levels as proposed in Cowie et al. (2012).

To compute the emotional ground truth for each recording (especially for facilitating subsequent computational modeling), we need to aggregate the decisions of multiple annotators. However, the ratings of some annotators might appear inconsistent with those of the majority of annotators. This issue is common in emotional labeling with categorical labels, where the emotional ground truth is often computed based on majority voting and minority labels are ignored, e.g. Busso et al. (2008). Here, we extend this notion to continuous ratings, using correlations as a basis for agreement. Specifically, we set a cut-off threshold for defining acceptable interrater agreement.[2] For each actor-recording, we take the union of all annotator pairs with linear correlations greater than the threshold. The ground truth for the corresponding actor-recording is computed by averaging annotation across annotators in the selected subset. If no annotators are selected then we assume that there is no agreement for that recording. Our threshold is empirically set to 0.45, which is similar to the correlation threshold used in Devillers et al. (2006) for defining agreement. This process results in ground truth agreement in 80, 84 and 73 actor-recordings for the activation, valence and dominance class respectively, out of 100 in total. Interestingly, a comparable percentage of ground truth agreement (about 75 %) was reported for annotation of categorical labels using this majority voting

[2] An alternative viewpoint to majority voting schemes is to explicitly model the diversity in these inherently subjective ratings when the ground truth is hidden from direct observation such as that proposed in Audhkhasi and Narayanan (2013).

scheme, for the IEMOCAP database (Busso et al. 2008), an emotional database of improvised acting. To get an impression of annotator agreement over the entire database, we first compute the mean of the correlations between every pair of annotators per actor-recording (no annotator selection is performed here), and then average over all actor-recordings. The agreement measures are presented in Table 3. We can observe a positive correlation (around 0.4) for all the emotional attributes. Given the challenging nature of the task, these correlations still indicate an acceptable level of annotator consistency (Cowie et al. 2012).

The consistency of the discrete global annotations was also examined by computing the Cronbach's α coefficient of global activation, valence and dominance ratings from all different annotators (no annotator selection is performed here). The Cronbach's α coefficient measures internal consistency, i.e., how closely related a group of raters are, and is widely used in the cognitive sciences. The values of α varies from 0 to 1. A higher value indicates a higher level of agreement. Those coefficients are also presented in Table 3. Overall, we notice that annotator consistency is at an acceptable level (around and over 0.7), except from dominance which is slightly lower.

In addition to examining interrater agreement, we also investigate the relationship between emotion annotations and the verbs selected by the actors (in Table 2) to induce different affective and expressive interactions. In general, we find that the global emotion annotations are congruent with the predefined goals. For example, the actor with the verb *shut him out* has been annotated as high activation (rating 3.5) and low valence (rating 2), while another actor with the verb *seduce* has been labeled as high activation (rating 4) and high valence (rating 4.5). Such consistency between human perception and predefined emotion goals further validates the effectiveness of the improvised technique of Active Analysis for data elicitation as used by CreativeIT.

5.4.2 *Comparing continuous and discrete annotations*

The availability of both global discrete and detailed continuous ratings from the same annotator, and for the same actor-recording, allows us to examine how annotators summarize the local perceived events of continuous information to produce an overall judgement about an interaction. Understanding the perception mechanism of annotators is essential for behavioral science where human assessment is the main approach for various research analyses (Lindahl 2001; Humphrey 1924). We applied several functionals to summarize each continuous rating and examined how close the resulting functional is to the global rating given by the annotator. The functions include mean, median, maximum, minimum, first and third quantile ($q1$ and $q3$) of various derivations from the continuous ratings of the recording. The discrete ratings were first shifted and re-scaled to match the range of the Feeltrace annotations.

Table 4 shows the mean squared error (MSE) between the discrete ratings and different functionals of continuous ratings over all actor-recordings. The last line is the MSE when we choose the function closest to the discrete rating between $q1$ and

Table 4 Mean squared error (MSE) between the discrete ratings and different functions of continuous ratings over all actor-recordings

Functional	Activation	Valence	Dominance
Mean	0.13	0.10	0.06
Median	0.14	0.10	0.07
Max	0.31	0.37	0.24
Min	0.71	0.26	0.40
$q1$	0.22	0.12	0.12
$q3$	0.14	0.13	0.08
Either $q1$ or $q3$	0.07	0.06	0.03

$q3$. We notice that the discrete rating is generally closer to either $q1$ or $q3$ compared to the other metrics, although which of the two functionals is closer to the discrete annotation varies per rating. Hence, the global rating is more influenced by either the highest or the lowest values of a continuous rating during a recording. Specifically, for 66 % of the activation ratings the discrete rating is closer to $q3$, for 59 % of the valence ratings the discrete rating is closer to $q1$, while for dominance there is an almost equal split. This suggests that global judgements of activation tend to be more influenced by the higher activated events of a recording, while global judgments of valence tend to be more influenced by the more negatively valenced events.

It also seems that different raters weight the same recording differently when making an overall decision. For example, for the recordings that were rated by three people (which is the large majority), only about 40 % of all annotators were consistent as to the quantile that they weighted more, either that was $q1$ or $q3$. These findings illustrate the complexity of the human perceptual and cognitive processing when summarizing emotional content; this processing is influenced by the emotional aspect to be evaluated, the events that are being observed, as well as person-specific characteristics. Some of these details can be explicitly modeled as proposed in Audhkhasi and Narayanan (2011, 2013).

6 Database analysis

In this section we describe some illustrative uses of the CreativeIT database.

6.1 Body language features

The availability of full body MoCap information, as shown in Fig. 1, enables us to extract detailed descriptions of each actor's body language expressed during an interaction. The derived body language features are motivated from the psychology literature which indicates that body language behaviors, such as looking at the other, turning away, approaching, touching or hand gesturing, are informative of a subject's attitude and emotion towards his/her interlocutor (Harrigan et al. 2005).

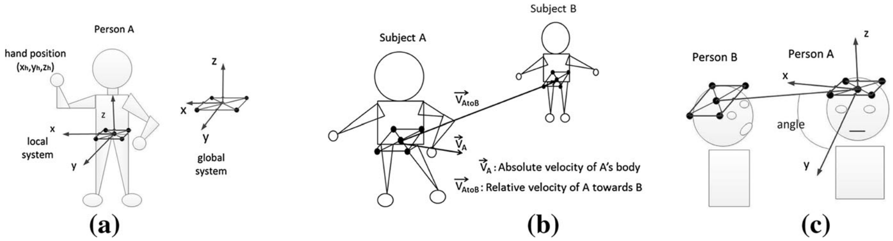


Fig. 4 The global and local coordinate systems for body language feature extraction; examples of extracted features. **a** Global and local reference systems, **b** body Velocities, **c** face orientation angle

While some of our features carry information about an individual's posture and motion, such as hand or body positions, others describe behaviors relative to the interlocutor, e.g., body and head orientation, approaching and moving away. The features are computed in a geometrical manner by defining global and local coordinate systems illustrated in Fig. 4a and by computing Euclidean distances, relative positions, angles and velocities. Similar movement features have also been applied for analyzing the link between emotion and gesture dynamics (Kapur et al. 2005; Metallinou et al. 2013). The full body mocap data available as a part of CreativeIT affords detailed quantitative investigation of gestural dynamics.

Some extracted body language features are summarized in Table 5. These are the main features used for describing behaviors of looking, touching, approaching, etc. The example features of *aVbody*, *rVbody* and *cosFace* are illustrated in Fig. 4b, c. The absolute velocity of a subject's body is computed from the movement in one's local coordinate system, while the relative velocity towards one's interlocutor is obtained by projecting the velocity vector in the direction between the two partners, as shown in Fig. 4b. Similarly in Fig. 4c, the angle of a subject's face towards the other, i.e., *cosFace*, is the angle between the head orientation in one's local system and the direction between the two participants. If *cosFace* is close to 1, the actor is looking towards the interlocutor, while $cosFace < 0$ indicates looking away.

Table 5 Examples of extracted body language features

cosFace	Cosine of subject's face angle towards the other
HAngle	Angle of head looking up and down
cosBody	Cosine of one's body orientation towards the other
cosLean	Cosine of one's body leaning angle towards the other
aVbody	Absolute velocity of subject's body
aVarmr,l	Absolute velocity of subject's right/left arms
rVbody	Relative velocity of one's body towards the other
rVhandr,l	Relative velocity of one's hand towards the other
rhandx, y, z	Subject's right hand (x, y, z) coord position
lhandx, y, z	Subject's left hand (x, y, z) coord position
dHands	Distance between subject's right and left hands

6.2 Example analysis: comparison of paraphrase and two-sentence exercise

As described in Sect. 3, the theatrical techniques of paraphrase and two-sentence exercise can elicit different body language expressions from actors. To better understand the body language behavior patterns, we investigate how actors behave in the two types of interactions. Based on the extracted body language features described in Sect. 6.1, we derive behavior descriptions of walking, moving away or towards the interlocutor, looking down or up, as well as facing away or towards the interlocutor. For example, if the absolute velocity of one subject's body is greater than a certain defined threshold, the corresponding behavior is tagged as walking. Otherwise, it is labeled as non-walking. For each type of behavior, we further develop two behavior metrics in each actor-recording: the frequency that each type of behavior occurs and the frequency that each type of behavior changes. The first one is computed by taking the ratio of frames when such type of behavior occurs and the total frames of the recording. The second one is derived by counting the behavior changes, e.g., changing from walking to non-walking and vice versa, and by normalizing the count with the recording length.

Table 6 presents the mean values of the two metrics for each type of behavior in paraphrases and two-sentence exercises, along with the statistical *t* test comparison results (*p* value) between the two types of interactions. As can be observed, all the derived behaviors except facing towards occur more frequently in the two-sentence interactions than in the paraphrase ones. Moreover, actors tend to change their behaviors more often when performing the two-sentence exercises. These may result from the restricted lexical content in the two-sentence exercise which helps to encourage richer body language expressions. These results reinforce the finding in Beattie (2004) that body language is an important nonverbal behavior for conveying communication intentions, especially when verbal behavior is constrained. From the observations, we may conjecture that the "two-sentence" theatrical approach could provide richer nonverbal behavior than when lexical content is also unrestricted, and

Table 6 Comparison of body language behavior in paraphrase and two-sentence exercise

	Two-sentence	Paraphrase	*p* value
Behavior occurring frequency (%)			
Walking	34.51	18.12	0.000
Moving towards	13.37	6.36	0.000
Moving away	13.06	6.17	0.000
Looking up	3.22	2.95	0.125
Looking down	0.97	0.23	0.018
Facing towards	75.27	83.39	0.025
Facing away	14.41	8.90	0.023
Behavior changing frequency (%)			
Walking	0.59	0.36	0.000
Moving	0.28	0.16	0.000
Looking	0.10	0.05	0.038
Facing	0.34	0.22	0.002

Table 7 Comparison of nonverbal vocalization in paraphrase and two-sentence exercise

	Two-sentence	Paraphrase	*p* value
Laughter	1.81	2.13	0.295
Sigh	1.67	0.77	0.021

that such richer body movements would be a strong cue to the authenticity of the elicited emotions. Our database enables us to study these issues more thoroughly.

As described in Sect. 4, each audio recording has been annotated with nonverbal vocal behavior of laughter and sigh. These nonverbal vocalizations occur spontaneously in interpersonal interactions, and are strongly linked to higher level human behavior related to internal cognitive and emotional state. Studies have shown that they often carry out significant social functions and affective state information (Hayworth 1928; Sauter et al. 2010). Studies on laughter have already attracted researchers' attention. Owren et al. investigated the acoustic cues of laughter in Bachorowski et al. (2001). Szameitat et al. (2009) further examined such cues with respect to different emotions. An interactive system has been developed in Niewiadomski et al. (2013) to automatically detect human laughs. It has been found that such laughter-aware system has a positive impact on user amusement. Our database provides the research community a good resource to study nonverbal vocalizations of laughter and sigh in conversations.

In addition to body language behavior, we also examine the nonverbal vocalization in paraphrase and two-sentence interactions. Table 7 presents the average number of times that laughter or sigh occurs in an actor-recording, associated with *p* values of *t* tests. It is interesting to observe that the vocal behavior of sigh occurs more frequently in the two-sentence exercise. The analyses of body language and nonverbal vocalizations support the effectiveness of the theatrical technique of two-sentence exercise for eliciting expressive communicative manifestations.

7 Ongoing and future research directions

The USC CreativeIT database is a novel, multimodal corpus collected in an interdisciplinary setting which represents a unique integration of engineering methods with the theory and practice of acting. It provides a rich resource for studying human emotional states, investigating expressive behaviors, and modeling human communication and interaction. In this section, some ongoing and possible future research directions with this corpus are discussed.

The following example case studies are used to highlight some of the possible uses with the CreativeIT data.

7.1 Continuous emotion tracking

The affective state of a participant evolves continuously over the course of an interaction. The work in Metallinou et al. (2013) used the behavior information of a participant, i.e., speech and body language, as features to continuously track changes in activation, valence and dominance during the active dyadic interactions of the CreativeIT database. Promising results demonstrated that the trends of participants' activation and dominance can be well captured. Future possibilities include computational modeling of affective state dynamics, and validating the models against observed data.

7.2 Dyadic behavior coordination

During dyadic interactions, participants adjust their behavior and give feedback continuously in response to the behavior of their interlocutors and the interaction context. In Yang et al. (2013), we studied how a participant in a dyadic interaction adapts his/her body language, such as body motion, posture and orientation, to the behavior (body language and speech) of the interlocutor, conditioned on the interaction goals and attitude/stance assumed e.g., of friendliness and conflict. Experimental studies revealed intricate dyadic behavior coordination as well as the predictability of one's body language from their interlocutor cues in these goal-driven interactions.

7.3 Attitude-related hand gesture dynamics

Hand gesture is one of the most expressive, natural and common types of body language for conveying attitudes and emotions in human interactions. We have explored affective content, such as underlying attitudes of friendliness or conflict, conveyed in the hand gesture dynamics during interactions in Yang et al. (2014a). The gesture dynamics are derived based on data-driven gesture primitives and have shown effectiveness in discriminating underlying attitude types.

All these studies involve the use of full-body Motion Capture data as well as the dyadic interaction settings of CreativeIT. Only a few existing corpora possess at most one of these features. Such uniqueness makes CreativeIT stand out and enables the possibility of detailed quantitative multimodal studies of human interaction and affect expression than hitherto possible.

Below we also list a few possible future research directions with this database. Although some initial research efforts have been pursued along some of these avenues, the rich resources provided by CreativeIT enable us to explore these, and other, directions with greater depth and breadth. In contrast to videos which are the central modality in most existing databases, the availability of detailed full body Motion Capture data allows us to describe body movements from head, hands, body to legs more accurately, and in conjunction with the vocal modality. The carefully-collected emotion annotations (both continuous and discrete) can facilitate developing proper emotion representations for various applications. In addition to the multimodal resources and emotion annotations, the theatrical interaction settings lend a good opportunity to study interaction dynamics evolution computationally and empirically, by considering aspects of emotional states, mutual influence of interlocutors, and expressive details of speech, posture and gestures.

7.4 Speech–gesture interplay

The speech and nonverbal gesture channels are internally and intricately coordinated toward conveying communicative intentions and reflect the underlying emotional state. Significant research progress has been made in studying aspects of speech-gesture interplay (Yang and Narayanan 2014). For example, researchers have investigated the interplay of neural activities between speech and hand gesture comprehension and they found a possible integration of hand gesture and speech at the early and late stages of language processing (Kelly et al. 2004). In addition to descriptive analysis inspired by such results to explicate the nature of speech-gesture interplay, this multimodal database is suitable for further modeling the interplay between speech and gestures and studying how such multimodal coupling is influenced by the internal emotion state or communication goal and stance (which in the database is specified by the improvisation verb in the active analysis methodology adopted for data elicitation).

7.5 Multimodal modeling

The rich multimodal resources provided by CreativeIT, including speech, gestures, continuous and discrete emotion annotations, as well as the communication verbs, allow us to design more robust and advanced emotion/attitude (continuous or discrete) recognition systems by modeling the interaction between the multimodal channels. Example previous work include where deep neural networks have been applied for cross-modality feature learning (Ngiam et al. 2011). In addition, the audio-video emotion challenge (AVEC) annually exhibits research work on multimodal integration for emotion recognition. However, the visual modality in these works mainly refers to videos which require more advanced computer vision techniques for gesture description. Availability of multimodal data like CreativeIT allows for feature engineering experiments beyond what is possible with conventional audio-visual data.

7.6 Full-body animation

The availability of full body MoCap data can also enable expressive body language animation. An animation framework for full body langugage synthesis driven by speech prosody has been proposed in Levine et al. (2009). The richness of CreativeIT enables us to better model the expressiveness, and predictability of animated body language in interactions. For example, we could derive fine-grained emotion categories from annotations and design animation functions for each derived category, in order to create intelligent virtual agents that displays appropriate body language with specific emotions in response to the human user's audio-visual behavior.

7.7 Interpersonal mutual influence

This database can be used to study the dynamic evolution of interpersonal mutual influence along aspects of speech prosody, body language, and emotional states, over a dyadic interaction. Lee et al. have made an initial attempt at modeling the mutual influence of partners' emotional states at turn level in a dyadic interaction for emotion tracking (Lee et al. 2009). With CreativeIT, it will be possible to further analyze how the dynamics of dyadic emotion coordination are related to the dyadic synchronization dynamics in terms of the available multimodal information.

7.8 Engineering-theater studies link

This database is designed based on theatrical techniques. Analysis of the relation between the theatrical performance ratings and the presentation of body language and speech prosody could facilitate quantitative research on theatrical performance.

7.9 Applications

The above research directions on this database can bring insights for developing naturalistic Human Computer Interfaces (HCI). An intelligent HCI can continuously process a variety of unfolding multimodal information from the human user(s) by monitoring their user's internal state and responding appropriately verbally and nonverbally. Pelachaud et al. (2002) have designed a conversational agent which can exhibit coherent speech and facial expressions and can display its emotions when communicating with human users; such efforts could be further advanced with these new data. One could also imagine educational technology applications with audio-visual sensing capabilities that would continuously assess a user's engagement and frustration levels and accordingly modify the educational material (Yildirim et al. 2011). Similarly, health-related applications could continuously monitor an individual's stress and anxiety levels and potentially give useful feedback to the user. Such applications are part of the emerging Behavioral Signal Processing domain that explores the role of engineering in developing health-oriented methods and tools (Narayanan and Georgiou 2013). Moreover, affect-sensitive virtual agents in gaming applications could continuously sense and interpret verbal and non-verbal cues of the user in order to estimate enjoyment and satisfaction. Such technologies would bring HCI closer to producing a human-like experience, and could have large impact in domains such as entertainment, education, security and healthcare.

8 Conclusion

In this paper, we presented the USC CreativeIT database which is designed based on the well-established improvisation technique of Active Analysis developed by Stanislavsky in order to provide naturally induced affective and expressive, goal-driven interactions. In total, it contains 50 dyadic theatrical improvisations performed by 16 actors, providing detailed full body Motion Capture data and audio and video data of each participant in an interaction. This database provides a novel bridge between the study of theatrical improvisation and human expressive behavior (speech and body language) in dyadic interactions. The carefully engineered data collection, the improvisation design to elicit natural emotions and expressive speech and body language, as well as the well-developed annotation processes provide a gateway to study and model various aspects of theatrical performance, expressive behaviors as well as human communication and interaction. The database is being made freely available to the research community (http://sail.usc.edu/data.php).

Acknowledgments This material is based upon work supported by DARPA and Space and Naval Warfare Systems Center Pacific under Contract Number N66001-11-C-4006 and the NSF.

References

Anolli, L., Mantovani, F., Mortillaro, M., Vescovo, A., Agliati, A., Confalonieri, L., Realdon, O., Zurloni, V., & Sacchi, A. (2005). A multimodal database as a background for emotional synthesis, recognition and training in e-learning systems. In *Affective computing and intelligent interaction*, pp. 566–573. Berlin: Springer.

Audhkhasi, K., & Narayanan, S. S. (2011). Emotion classification from speech using evaluator reliability-weighted combination of ranked lists. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4956–4959.

Audhkhasi, K., & Narayanan, S. (2013). A globally-variant locally-constant model for fusion of labels from multiple diverse experts without using reference labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(4), 769–783.

Bachorowski, J. A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, *110*(3), 1581–1597.

Bänziger, T., & Scherer, K. R. (2007). Using actor portrayals to systematically study multimodal emotion expression: The GEMEP corpus. In *Affective computing and intelligent interaction*, pp. 476–487.

Beattie, G. (2004). *Visible thought: The new psychology of body language*. New York: Psychology Press.

Busso, C., & Narayanan, S. (2008). 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*, pp. 17–22.

Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, *42*(4), 335–359.

Carnicke, S. M. (2009). *Stanislavsky in focus: An acting master for the twenty-first century*. London: Taylor & Francis.

Cowie, R., & Sawey, M. (2011). GTrace-General trace program from Queen's, Belfast.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.

Cowie, R., McKeown, G., & Douglas-Cowie, E. (2012). Tracing emotion: An overview. *International Journal of Synthetic Emotions (IJSE)*, *3*(1), 1–17.

Crane, E., & Gross, M. (2007). Motion capture and emotion: Affect detection in whole body movement. In *Affective computing and intelligent interaction*, pp. 95–101. Berlin: Springer.

Devillers, L., Cowie, R., Martin, J., Douglas-Cowie, E., Abrilian, S., & McRorie, M. (2006). Real life emotions in French and English tv video clips: An integrated annotation protocol combining continuous and discrete approaches. In *5th international conference on language resources and evaluation (LREC 2006)*, Genoa, Italy.

Dhall, A., Member, S., Lucey, S., & Gedeon, T. (2012). Collecting large, richly annotated facial-expression databases from movies. *IEEE Multimedia*, *19*(3), 34–41.

Douglas-Cowie, E., Cowie, R., Sneddon, I., Cox, C., Lowry, O., Mcrorie, M., Martin, J. C., Devillers, L., Abrilian, S., Batliner, A., Amir, N., & Karpouzis, K. (2007). The humaine database: Addressing the collection and annotation of naturalistic and induced emotional data. In *Affective computing and intelligent interaction*, pp. 488–500. Berlin: Springer.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, *40*(1), 33–60.

Enos, F., & Hirschberg, J. (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))*, pp. 6–10.

Grafsgaard, J. F., Fulton, R. M., Boyer, K. E., Wiebe, E. N., & Lester, J. C. (2012). Multimodal analysis of the implicit affective channel in computer-mediated textual communication. In Proceedings of the 14th ACM international conference on multimodal interaction, pp. 145–152. New York: ACM.

Grimm, M., Kroschel, K., & Narayanan, S. (2008). The Vera am Mittag German audio-visual emotional speech database. In *2008 IEEE international conference on multimedia and expo*, pp. 865–868. New York: IEEE.

Harrigan, J., Rosenthal, R., & Scherer, K. (2005). *The new handbook of methods in nonverbal behavior research*. Oxford: Oxford University Press.

Hayworth, D. (1928). The social origin and function of laughter. *Psychological Review*, *35*(5), 367.

Humphrey, G. (1924). The psychology of the gestalt. *Journal of Educational Psychology*, *15*(7), 401.

Johnstone, K. (1981). *Impro: Improvisation and the theatre*. London: Routledge.

Kanluan, I., Grimm, M., & Kroschel, K. (2008). Audio-visual emotion recognition using an emotion space concept. In *16th European signal processing conference*, Lausanne, Switzerland.

Kapur, A., Kapur, A., Virji-Babul, N., Tzanetakis, G., & Driessen, P. F. (2005). Gesture-based affective computing on motion capture data. In *Affective computing and intelligent interaction*, pp. 1–7. Berlin:Springer.

Kelly, S. D., Kravitz, C., & Hopkins, M. (2004). Neural correlates of bimodal speech and gesture comprehension. *Brain and Language*, *89*(1), 253–260.

Koelstra, S., Muhl, C., Soleymani, M., Lee, J. S., Yazdani, A., Ebrahimi, T., et al. (2012). Deap: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing*, *3*(1), 18–31.

Lee, C. C., Busso, C., Lee, S., & Narayanan, S. S. (2009). Modeling mutual influence of interlocutor emotion states in dyadic spoken interactions. In *INTERSPEECH*, pp. 1983–1986.

Levine, S., Theobalt, C., & Koltun, V. (2009). Real-time prosody-driven synthesis of body language. *ACM Transactions on Graphics (TOG)*, *28*(5), 172.

Lindahl, K. M. (2001). Methodological issues in family observational research. In: P. K. Kerig & K. M. Lindahl (Eds.), *Family observational coding systems: Resources for systemic research* (pp. 23–32). Mahwah, NJ:Lawrence Erlbaum Associates.

Malandrakis, N., Potamianos, A., Evangelopoulos, G., & Zlatintsi, A. (2011). A supervised approach to movie emotion tracking. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2376–2379. New York: IEEE.

McKeown, G., Curran, W., McLoughlin, C., Griffin, H. J., & Bianchi-Berthouze, N. (2013). Laughter induction techniques suitable for generating motion capture data of laughter associated body movements. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–5. New York: IEEE.

McKeown, G., Valstar, M., Cowie, R., Pantic, M., & Schroder, M. (2012). The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, *3*(1), 5–17.

Mendonca, D. J., & Wallace, W. A. (2007). A cognitive model of improvisation in emergency management. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, *37*(4), 547–561.

Metallinou, A., & Narayanan, S. (2013). Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, pp. 1–8. New York: IEEE.

Metallinou, A., Katsamanis, A., & Narayanan, S. (2013). Tracking continuous emotional trends of participants during affective dyadic interactions using body language and speech information. *Image and Vision Computing, Special Issue on Continuous Affect Analysis*, *31*(2), 137–152.

Metallinou, A., Katsamanis, A., Wang, Y., & Narayanan, S. (2011). Tracking changes in continuous emotion states using body language and prosodic cues. In *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 2288–2291. New Yok: IEEE.

Metallinou, A., Lee, C. C., Busso, C., Carnicke, S., Narayanan, S., & Tx, D. (2010). The USC CreativeIT database: A multimodal database of theatrical improvisation. In *Workshop on Multimodal Corpora, LREC*.

Narayanan, S., & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE*, *101*(5), 1203–1233.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pp. 689–696.

Niewiadomski, R., Hofmann, J., Urbain, J., Platt, T., Wagner, J., Piot, B., Cakmak, H., Pammi, S., Baur, T., & Dupont, S., et al. (2013). Laugh-aware virtual agent and its impact on user amusement. In *Proceedings of the 2013 international conference on autonomous agents and multi-agent systems*, pp. 619–626. International Foundation for Autonomous Agents and Multiagent Systems.

Pelachaud, C., Carofiglio, V., De Carolis, B., de Rosis, F., & Poggi, I. (2002). Embodied contextual agent in information delivering application. In *Proceedings of the first international joint conference on autonomous agents and multiagent systems: Part 2*, pp. 758–765. New York: ACM.

Perlin, K., & Goldberg, A. (1996). Improv: A system for scripting interactive actors in virtual worlds. In *Proceedings of the 23rd annual conference on computer graphics and interactive techniques*, pp. 205–216. New York: ACM.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408–2412.

Scherer, K. R., Bänziger, T., & Roesch, E. (2010). *A blueprint for affective computing: A sourcebook and manual*. Oxford: Oxford University Press.

Sneddon, I., McRorie, M., McKeown, G., & Hanratty, J. (2012). The belfast induced natural emotion database. *IEEE Transactions on Affective Computing*, *3*(1), 32–41.

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, *3*(1), 42–55.

Szameitat, D. P., Alter, K., Szameitat, A. J., Wildgruber, D., Sterr, A., & Darwin, C. J. (2009). Acoustic profiles of distinct emotional expressions in laughter. *The Journal of the Acoustical Society of America*, *126*(1), 354–366.

Wallbott, H. G., & Scherer, K. R. (1986). Cues and channels in emotion recognition. *Journal of Personality and Social Psychology*, *51*(4), 690.

Wu, S., Falk, T. H., & Chan, W. Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication*, *53*(5), 768–785.

Yang, Z., & Narayanan, S. (2014). Analysis of emotional effect on speech-body gesture interplay. In *Proceedings of Interspeech*.

Yang, Z., Metallinou, A., & Narayanan, S. (2013). Towards body language generation in dyadic interaction settings from interlocutor multimodal cues. In *Proceedings of ICASSP*.

Yang, Z., Metallinou, A., Erzin, E., & Narayanan, S. (2014a). Analysis of interaction attitudes using data-driven hand gesture phrases. In *Proceedings of ICASSP*.

Yang, Z., Metallinou, A., & Narayanan, S. (2014b). Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues. *IEEE Transactions on Multimedia*, *16*, 1766–1778.

Yang, Z., Ortega, A., & Narayanan, S. (2014c). Gesture dynamics modeling for attitude analysis using graph based transform. In *Proceedings of IEEE international conference on image processing*.

Yildirim, S., Narayanan, S., & Potamianos, A. (2011). Detecting emotional state of a child in a conversational computer game. *Computer, Speech, and Language*, *25*, 29–44.