MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception

Carlos Busso, *Senior Member, IEEE,* Srinivas Parthasarathy, *Student Member, IEEE,* Alec Burmania, *Student Member, IEEE,* Mohammed AbdelWahab, *Student Member, IEEE,* Najmeh Sadoughi, *Student Member, IEEE,* and Emily Mower Provost *Member, IEEE*

Abstract—We present the MSP-IMPROV corpus, a multimodal emotional database, where the goal is to have control over lexical content and emotion while also promoting naturalness in the recordings. Studies on emotion perception often require stimuli with fixed lexical content, but that convey different emotions. These stimuli can also serve as an instrument to understand how emotion modulates speech at the phoneme level, in a manner that controls for coarticulation. Such audiovisual data are not easily available from natural recordings. A common solution is to record actors reading sentences that portray different emotions, which may not produce natural behaviors. We propose an alternative approach in which we define hypothetical scenarios for each sentence that are carefully designed to elicit a particular emotion. Two actors improvise these emotion-specific situations, leading them to utter contextualized, non-read renditions of sentences that have fixed lexical content and convey different emotions. We describe the context in which this corpus was recorded, the key features of the corpus, the areas in which this corpus can be useful, and the emotional content of the recordings. The paper also provides the performance for speech and facial emotion classifiers. The analysis brings novel classification evaluations where we study the performance in terms of inter-evaluator agreement and naturalness perception, leveraging the large size of the audiovisual database.

Index Terms—emotion elicitation, audiovisual emotional dataset, emotional evaluation, emotion recognition

1 Introduction

Audiovisual data with fixed lexical content, but with different elicited emotions, is valuable in the field of affective computing. Data of this type have played a key role in studies addressing how emotion modulates facial expressions and speech at the phoneme level [1], [2]. By fixing the lexical content, the analysis can focus on differences in expressive behaviors associated with emotions. These stimuli are also important for perceptual evaluations that address audiovisual emotion integration [3]. Multimodal cue integration can also be studied by creating stimuli with emotionally inconsistent content [3]–[6]. The approach requires sentences with the same lexical content spoken with different emotions (sadness, happiness, anger and neutrality). By fixing the lexical content and using resampling and interpolation, it is possible to create videos with mismatched conditions (e.g., happy speech, sad face), without introducing inconsistencies between the actual speech and the facial appearance (especially the lips). These requirements necessitate the use of controlled recordings that can be only collected from actors.

Actors have played an important role in the study of emotions. Most of the early emotional corpora were recorded from actors (e.g., Berlin database of emotional speech [7], emotional prosody speech and transcripts [8], and the Danish Emotional Speech Database [9]). While desirable from the viewpoint of providing controlled recordings, the research community has shifted toward natural databases. However, as highlighted by Douglas-Cowie et al. [10], many of the problems associated with acted recordings are not the use of actors per se, but the methodology used to elicit emotions (e.g., lack of interaction dialogs, read instead of conversational speech, lack of context). Most of the acted databases were recorded by actors or naïve speakers reading isolated sentences, words or sounds portraying given emotions. These recordings usually provide prototypical behaviors that do not represent the emotional expressions that we expect to observe in real applications [10]–[12].

Our work [13], [14] and that of other researchers [15]–[18] have argued that using actors is still a valid approach to study emotions. The key is to explore better elicitation techniques rooted in different acting styles and theater theory which provide a viable research methodology for studying human emotions [13], [17]. In this paper, we explore this option through the MSP-IMPROV corpus. The elicitation approach for the MSP-IMPROV corpus is designed to satisfy the aforementioned requirements (fixed lexical con-

- *C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, and N. Sadoughi are in the Erik Jonsson School of Engineering & Computer Science, The University of Texas at Dallas, Richardson, TX 75080 (e-mail:busso@utdallas.edu, sxp120931@utdallas.edu, axb124530@utdallas.edu, mxa129730@utdallas.edu, nxs137130@utdallas.edu).*
- *E. Mower Provost is in the Computer Science and Engineering Department, University of Michigan, Ann Arbor, MI 48109 (e-mail:emilykmp@umich.edu*

tent across emotions), while keeping the recordings as natural and spontaneous as possible. To balance the tradeoff between natural expressive interactions and controlled conditions, we implement a novel recording paradigm that approaches the naturalness found in unsolicited human speech. The approach consists of creating different conversational scenarios (one per emotion) that two actors improvise in dyadic interactions. The novelty of the approach is that these scenarios are designed such that the actors can utter a sentence with fixed lexical content during the improvised dialog that expresses the target emotion (we refer to these utterances as target sentences). We capitalize on the context provided by the stories to elicit the target emotion, while maintaining the fixed lexical content required by our experimental framework. By recording the entire improvisation, not just target sentences, the actors have the flexibility to introduce emotional behaviors during their dyadic performance that are closer to "real" emotions observed in natural conversations. As the perceptual evaluation indicates, this approach provides recordings that are perceived as more natural than read renditions.

The focus of this paper is the introduction of the MSP-IMPROV corpus, describing the context in which it was recorded, the key features of the corpus, the areas where this corpus can be useful, and the emotional content of the recordings. We have recorded six sessions (three hours each) from 12 actors, collecting 652 target sentences containing lexical content that meets the aforementioned criteria. In addition to the target sentences, crucial for the project's goals, we also collect and segment all the turns during the improvisation that led to the target sentences (*Other - improvised*), plus all the interactions between the actors during breaks (*Natural interaction*), resulting in a unique emotional database with 7,818 conversational interaction turns (9 hours of audiovisual data). The emotional content of the turns was evaluated in terms of categorical labels and dimensional attributes. Each turn was annotated by at least five subjects using a crowdsourcing-based approach [19]. We present detailed analysis of the emotional content of the corpus.

While the motivation for the MSP-IMPROV corpus is emotion perception, the large size of the audiovisual database and the number of evaluations per sentence are important contributions in other areas of affective computing including emotion recognition. We evaluate the performance of speech emotion classifiers using the emotional labels as ground truth. The evaluation includes classification performance for all the data, and for each of the individual portions of the corpus (*Target - improvised*, *Target - read*, *Other - improvised*, *Natural interaction*). In addition, the analysis also brings novel classification evaluations where we study the performance in terms of inter-evaluator agreement and naturalness perception.

We organize the rest of the paper as follows. Section 2 gives an overview of existing databases. We describe the important role that the MSP-IMPROV database can have in the context of existing corpora. Section 3 explains the design, collection and post processing steps of the corpus. Section 4 explains the perceptual evaluation process to collect categorical and attribute-based annotations. This section also analyzes the emotional content of the corpus. Section 5 presents the results of automatic emotion classifiers using speech features. This section also explores classification performance in terms of agreement between evaluation and perceived naturalness. Section 6 concludes the study with discussion, potential uses of this corpus, and final remarks.

2 Related Work

Emotional databases play a key role in the study of affective computing. Reviews about emotional databases are given by Douglas Cowie et al. [10], El Ayadi et al. [20], and Ververidis and Kotropoulos [21]. This section overviews techniques to record emotional databases, emphasizing the role of recordings collected by actors.

2.1 Acted versus Natural Databases

Most recent efforts have focused on the recording of natural human interactions. The main benefit of these corpora is the genuine emotional behaviors captured by these approaches, which are difficult to elicit with acted recording. Devillers et al. [12] discussed the challenges of real-life emotions highlighting the differences from archetypal emotions from acted recordings. Batliner et al. [11] argued that the classification performances achieved with acted recordings are not representative of the accuracies that are obtained in real-applications. They argued that natural recordings are needed to improve performance in real applications. This conclusion is supported by Douglas-Cowie et al. [10], who emphasized that recordings in databases should reflect the emotional behaviors found in real human interactions.

Some natural databases exploit existing recordings from television, web-content or actual applications. An example is the *Vera Am Mittag* (VAM) audiovisual database, which consists of recordings from a German television program, during which couples discussed their problems [22]. Other similar approaches include databases consisting of videos from the web [23]–[25], and call center recordings [26], [27]. The main challenges in these recordings are privacy/copyright issues, and the lack of control over the settings (microphones, frontal views of the subjects, and background noise). An alternative approach to mitigate these limitations is to elicit emotional reactions from participants using more controlled settings. Examples include recordings of kids interacting with robots [28],

using *wizard of oz* (WOZ) methods during human-machine interaction [29]–[32], requesting the subjects to recall personal emotional experiences [33], inducing emotions with *sensitive artificial listener* (SAL) [34], having collaborative tasks between participants [35], and recording the reactions of individuals watching emotional videos [36]. Additionally, Tcherkassof et al. [37] and Sneddon et al. [38] used various computer tasks to elicit natural facial behaviors.

While natural corpora are key resources to the community, many interesting research questions require more controlled conditions. In particular, studies on emotion perception usually require stimuli with fixed lexical information. These studies can systematically rely on acted recordings to satisfy the requirements imposed by the given research questions. Since the goal is to understand how people perceive displays of emotion, it is less problematic that these displays are acted. Furthermore, we argue that even for emotion recognition, acted databases offer complementary information to natural recordings [39]. The context in which the natural recordings are collected dictates the spectrum of emotional behaviors on the corpora, which are often limited. For example, the VAM database is biased toward negative emotions. In contrast, acted recordings provide diversified and balanced emotional behaviors that can be difficult to achieve using the aforementioned natural elicitation approaches. Scherer [16] argued that display rules regulate the emotions that people are willing to express in human interaction. Therefore, even natural recordings may not provide a good representation of the 'true' felt emotion. In contrast, the study suggested that acted recordings offer the experimental control required to address important open questions [15].

2.2 Elicitation Approaches for Acted Databases

The use of acting styles and theater theory can mitigate the elicitation of exaggerated behaviors that do not represent the natural expressive behaviors observed in real human interactions. Acting methods such as the one proposed by Stanislavsky [40] suggest that actors should develop an action for a particular script. By developing the personality and elements to achieve their action, the actors are expected elicit believable behaviors [17]. This goal-oriented performance was explored in the collection of the USC-CreativeIT Database [41]. The actors interacted during dyadic interactions, using an approach inspired by active analysis methodology [40]. The approach is called "two-sentence-exercise", where each actor was allowed to use only one sentence ("marry me" and "I'll think about it") during their improvisation, which lasts a few minutes. The flow of the interaction was influenced by a specific verb given to the actors (e.g., confront versus deflect). Appraisal theory indicates that emotions are elicited as a result of an individual's assessment of his/her surrounding environment, events or situations [42]. Therefore, we can collect more realistic acted data by carefully designing protocols where the emotions are naturally elicited following the flow of the interaction - as suggested by these acting techniques. Enos and Hirschberg [17] argued that acted databases can provide natural realistic emotions provided they are recorded using suitable elicited scenarios. The IEMOCAP database is an example of such a corpus, where in each session two actors participated in scripted and improvised scenarios [14].

There are interesting studies that have used contextual information to record their corpora from actors. Lefter et al. [43] used hypothetical scenarios to record stressful conditions at a service desk. Two participants improvised the scenarios. In contrast to our work, there was no lexical constraint in the recording. Some studies have used elicitation methods similar to our proposed approach. Cao et al. [44] instructed the actor to express a target emotion while reading a sentence. They used a director who provided scenarios to help the actors to achieve their goal. Some of the actors used personal experience to elicit the emotions. Banse and Scherer [18] and Martin et al. [45] asked actors to utter target sentences, where a scenario was presented for each sentence to elicit different emotional reactions. In contrast to our approach, these studies included a single speaker performing only the target sentences. The proposed approach presents the following advantages: (a) the use of dyadic interaction to record target sentences with fixed lexical content (previous studies only used monologues), and (b) the recording of the entire scenario, not just target sentences (other studies have only collected the target sentence). Dyadic interactions allow actors to elicit natural emotions responding to the flow of hypothetical scenarios. By recording the entire improvisation, not just target sentences, the actors have the flexibility to introduce emotional behaviors during their performance that are closer to "real" emotions observed in natural conversations.

3 The MSP-IMPROV Database

The MSP-IMPROV is an acted audiovisual database that explores emotional behaviors during conversational dyadic improvisations. The corpus, which will be made available to the community, fills a gap in the study of emotion perception as few emotional corpora have been collected specifically for perceptual studies (important exceptions are the GEMEP [46] and CREMA-D [47] databases). This section describes the elicitation technique (Sec. 3.1), the setup used to collect the corpus (Sec. 3.2), and the post-processing steps (Sec. 3.3).

3.1 Collecting Target Sentences with Dyadic Improvisation

We design the MSP-IMPROV corpus to elicit emotional behaviors with fixed lexical content, but that convey different emotions, which are referred to as *target sentences*. Having stimuli where we have control over lexical content and emotion provides an opportunity to explore interesting questions. For example, the database will allow us to better understand the interplay between audiovisual modalities in the expression of emotions, the importance of which was highlighted in our previous studies [1], [2], [48], [49]. The comparison of sentences with fixed lexical content, but that convey different emotions, allows us to explore speech and facial emotion expression production strategies. This will enable the identification of primary and secondary multimodal cues that are perceptually important. One of the long-term goals of this project is to develop new insight into emotion perception by creating stimuli where we reduce the correlation that exists between facial and vocal expressions of emotion. This allows us to have greater insight into how each channel affects perception. We accomplish this goal through the creation of emotional stimuli with congruent and conflicting audio-visual emotional expressions (e.g., a video with happy facial expressions, but angry speech) [3]–[6]. By fixing the lexical content in the MSP-IMPROV corpus, we will be able to create conflicting stimuli by aligning sentences using resampling and interpolation techniques [6].

It is important to record the target sentences using elicitation techniques that produce behaviors that are as natural and spontaneous as possible, while meeting the requirements of fixed lexical content (as opposed to read renditions, which are likely to produce prototypical emotions that differ from natural expressive behaviors). We build upon the approach presented by Martin et al. [45], in which a scenario that is carefully designed to elicit a given emotion is provided to the actors. Note that for each target sentence, the approach requires different scenarios to elicit different emotions. In Martin et al. [45], the subjects were asked to act only target sentences in monologue recordings. Instead, we record 30 to 60 second dyadic improvisation dialogs. We instruct the actors to utter the target sentences during the improvisation. We hypothesize that the production of the target sentences using this approach is more natural and genuine than read, monologue recordings, since the emotions are elicited as dictated by the scenario's context during spontaneous improvisation between two actors. The study focuses on three basic emotions corresponding to happiness, sadness, and anger. We also include neutrality. These categories are the most common emotional classes used in related studies.

We designed 20 target sentences with one hypothetical scenario per emotion (20 sentences $\times$ 4 emotions = 80 scenarios). The sentences were selected from the same list used in our previous recordings [48], which included phonetically balanced sentences. We selected 20 sentences with the following criteria: the sentences should be generic enough to design appropriate scenarios to trigger the target emotions; and, the sentences should include short and long utterances for the perceptual study (perceiving emotional cues from short stimuli without context is a challenging task even without any mismatch).

TABLE 1
Target sentences used for the recordings.

#	Target Sentences
1	How can I not?
2	I'm quite sure that we will find some way or another
3	Ella Jorgenson made the pudding
4	The floor was completely covered
5	They are just going to go ahead regardless
6	It has all been scheduled since Wednesday
7	I am going shopping
8	A preliminary study shows rats to be more inquisitive than once thought
9	That's it, the meeting is finished
10	I don't know how she could miss this opportunity
11	It is raining outside
12	Your dog is insane
13	She told me what you did!
14	Your grandmother is on the phone
15	Only I joined her in the ceremony

After creating the 80 scenarios, four of the authors of this paper ranked these sentences according to the quality (i.e., do these scenarios elicit the target emotions?) and the acting complexity of the scenarios (i.e., can an actor provide realistic renditions of the situation?). Table 1 lists the top 15 sentences, which are the target sentences used in the recordings (the five sentences with the lowest scores were not used). Table 2 summarizes the four scenarios used for the target sentence "How can I not?" The actor playing *Person A* is always the one who utters the target sentence. We include the full list of scenarios for the 20 sentences as a Supplemental Material. Notice that the scenarios only describe the context. The actors were free to improvise the scenario using their own language. The only constraint was that they had to utter the target sentence word-for-word. If we noticed overlapped speech during the recording of the target sentences, we asked the actors to repeat the improvisation. After a few minutes of practice, the actors followed the described protocol. They did not note any cognitive load challenge associated with the task.

3.2 Collection of the Audiovisual Corpus

The use of skilled actors, as opposed to naïve subjects, increases the authenticity of the portrayed emotions [14]. Therefore, we carefully recruited 12 English speaking students (6 males, 6 females) from the theatre program of The School of Arts and Humanities at *The University of Texas at Dallas* (UT Dallas). Their

TABLE 2
Example of scenarios per emotion for the target sentence "*How can I not?*." Scenarios for other sentences are given as a Supplemental Material.

Anger
Person A: Your friend is lazy and rarely goes to class. You got up late and look tired, your friend suggests that you don't go to class. You are upset by this suggestion and reply "How can I not?"
Person B: Your friend looks tired and doesn't seem to want to go to class. You don't think it is a big deal to skip the class.
Happiness
Person A: You just got a phone call and were told that you were hired for the job that you really wanted. Your friend asks you if you are going to accept. You reply "How can I not?"
Person B: Your friend just got the job that he/she wanted. You ask him/her about the job and if he/she is going to take it.
Sadness
Person A: You are failing your classes in college, and realize that you will probably have to drop out and return home. Your friend is trying to calm you down and tells you that you don't have to go home. You reply "How can I not?"
Person B: You are consoling your friend by telling him/her that he/she can pass the class, so he/she won't have to go back home.
Neutral
Person A: You are out shopping with your friend and head to checkout. Your friend reminds you of the coupon you brought and asks you if you are going to use it. You reply "How can I not?"
Person B: You remind your friend about the coupon he/she brought. You are curious if he/she is going to use it this time.

ages ranged between 18 and 21 at the time of the data collection (M=19.2, SD=1.1). We recruited many of them by contacting actors from the casts of the main stage productions organized by UT Dallas. We gave a brief description of the database and the scenarios so they were familiar with the task. We recorded six dyadic sessions, by pairing an actor and an actress.

We collected the scenarios for the sentences listed in Table 1 as follows. First, we split the sentences into three groups of five, following the order listed in the table (i.e., we first collected the sentences that were more highly ranked – see Sec. 3.1). For a given emotion, we recorded the scenarios of sentences within a group of five target sentences. We repeat this process for each of the four target emotions. After the completion of a group of five sentences, we moved onto the next group of five sentences. This protocol prevents the actors from playing similar improvisations for a target sentence across emotions. This approach also balances the tradeoff between jumping too often between emotions, increasing the actor's cognitive load, and having repetitive behaviors by sequentially recording many scenarios associated with the same emotion. One actor played the role of *Person A*, the one uttering the target sentence, while the other played the role of *Person B*, the one supporting the improvisation (see Table 2). After recording the scenarios for the first 10 sentences, the actors switched roles and we collected these scenarios again. With this approach, we have recordings of the target sentences from the 12 actors. In four of the six dyadic sessions, we had enough time to record the scenarios for five extra sentences for both actors (i.e., four sessions with 15 target sentences, two sessions with 10 target sentences). The duration of the recordings was three hours including breaks, as defined in our *institutional review board* (IRB).

The MSP-IMPROV database was collected in a 13ft $\times$ 13ft ASHA certified single-walled sound booth (Fig. 1(a)). The actors sat two meters apart facing each other. The distance was needed to simultaneously record both actors. Two high resolution digital cameras were placed facing each actor to capture their faces (Figs. 1(c) and 1(d)). The resolution of the videos was set at 1,440 $\times$ 1,080 pixels at 29.97 frames per second. We used two chroma-key green screens behind the actors, which allowed us to achieve a uniform background. We used two professional LED light panels, which were placed behind the actors to provide uniform illumination. The audio was recorded with two collar microphones, one for each actor (48kHz and 32-bit PCM). The audio was simultaneously recorded using the PC audio interface Tascam US-1641 (Fig. 1(b)). We used a clapboard to synchronize the cameras with the audio channels. The description of the scenarios were displayed in a monitor screen on one side of the room (Fig. 1(b)). A buzzer ring is generated and separately recorded between scenarios to facilitate the segmentation of the corpus.

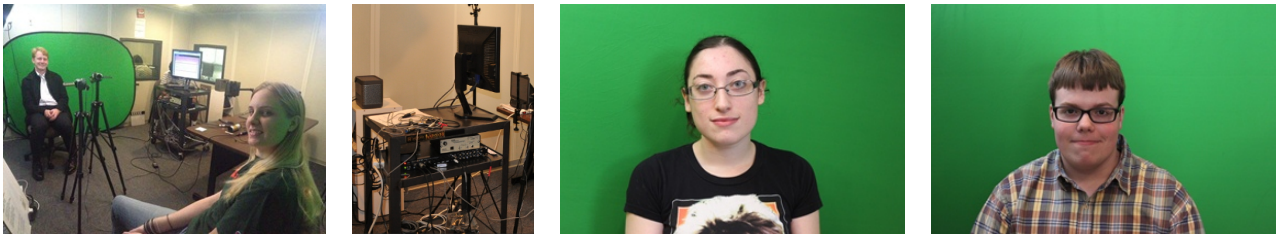
3.3 Post-processing of the Corpus

The corpus was manually segmented into dialog turns by one of the authors of this paper. We define a dialog turn as an uninterrupted utterance or a sentence, whichever is shorter. When possible, we extended the boundaries to include small silence segments at the beginning and ending of the turn. This approach allows us to explore anticipatory facial gestures that are not apparent in the speech signal. The corpus has four main datasets:

Target - improvised: We collected 652 target sentences which can be used to generate thousands of emotionally congruent and conflicting audiovisual stimuli [6]. We refer to these sentences as *Target - improvised*. The word "Target" refers to the fact that the lexical content of the sentences is fixed (i.e., one of the sentences in Table 1). The word "improvised" refers to the fact that the sentences were recorded during the improvisation of the corresponding scenarios.

Other - improvised: We are interested in all the actors' turns during the improvisation sessions, not just the target sentences. We collected 4,381 of these turns, which we refer to as *Other - improvised*.

Natural interaction: We followed the continuous recording approach used by McKeown et al. [50] consisting of recording natural interaction during the breaks (i.e., while the subjects were not acting). We



(a) Setup (b) Audio setup (c) Actress (d) Actor

Fig. 1. Data recording setup. Two actors face each other, while two cameras and two collar microphones simultaneously capture the dyadic interaction in a sound booth.

noticed that spontaneous conversation during breaks conveyed emotions, mostly positive, as the actors discussed the scenarios and reacted to mistakes. This was possible since the cameras and microphones were never stopped, recording the entire sessions including the breaks. We collected 2,785 of these turns, which we refer to as *Natural interaction*.

Target - read: We asked the actors to come back on a different day to record the target sentences again. This time, we asked them to read the sentences portraying the four target emotions (one actor at a time without context). Eight of the actors agreed to participate in these extra recordings. We collected 620 of these turns, which we refer to as *target - read*. We use these recordings to compare the proposed elicitation technique to the conventional approach used in previous acted databases.

Overall, the MSP-IMPROV database contains 7,818 non-read turns, including *Target - improvised*, *Other - improvised*, and *Natural interaction*, plus 620 *Target - read* sentences.

The last columns of Table 3 report the mean duration of the dialog turns. These values only include the duration of the speaking turns and do not include silence at the beginning and ending of the segments. To estimate the duration of the dialog turns, we use a state-of-the-art *speech activity detection* (SAD) system developed by Sadjadi and Hansen [51]. The average duration of the target sentences (*Target - improvised*, and *Target - read*) is shorter than the duration for the other dialog turns at 1.9 seconds and 3 seconds, respectively. The standard deviation of the utterance length is higher for the non-target sentences (Table 3).

4 Emotional Content of the Corpus

We annotate the emotional content of the MSP-IMPROV corpus using crowdsourcing. Perceptual evaluations are crucial to characterize the affective content of emotional databases. While early studies conducted perceptual evaluations in laboratory conditions, recently, researchers have explored crowdsourcing services to evaluate the emotional content of databases [6], [44], [52], [53]. We have three main motivations to use crowdsourcing. First, we aim to evaluate the corpus with many evaluators per turn. This is a challenging, resource-demanding task since the corpus has 8,438 turns. Crowdsourcing offers the opportunity to recruit evaluators at a fraction of the cost required to conduct the evaluation in the laboratory [54]. For example, the CREMA-D database was evaluated by ten annotators per sentence using crowdsourcing [44], which is more than the common number of raters used in other corpora [39]. More evaluations provide better characterization of the perceived emotions. Second, using crowdsourcing significantly reduces the time required for the evaluation [54]. This is important since we collected over 50,000 evaluations. Third, crowdsourcing gives access to a diverse pool of subjects, who would be difficult to reach with other conventional settings [55].

The evaluators were recruited from the United States and were paid between 4 and 8 cents per video. The evaluators could annotate multiple videos out of context from the four datasets, but could not evaluate a particular video more than once. This choice was made to reduce the effect of context. Contextual information affects perceptual judgment of the evaluators. Cauldwell [56] demonstrated important differences in emotional perceptions between evaluations conducted in sequential and random order, showing that knowledge from previous speaking turns influenced the judgment of the evaluators. Since our aim is to understand acoustic and facial cues that we use to decode emotional behaviors, contextual information introduces an extra dimension that is not easy to control. After briefly describing the emotional annotation process (Sec. 4.1), this section analyzes the emotional content of the recordings in detail (Sec. 4.2).

4.1 Perceptual Evaluation Using Crowdsourcing

We evaluate the emotional content of the corpus using crowdsourcing (Amazon Mechanical Turk). Monitoring the consistency and quality of annotations provided by crowdsourcing is crucial to collect useful data. Perceptual evaluations to annotate emotional corpora are usually long and tedious. An evaluator may tire and provide unreliable labels. To address this problem, we design an elegant approach to identify in

TABLE 3
General statistics per dataset. The table reports information about the evaluators including their gender distribution (*F*=Female, *M*=Male). It describes the number of evaluators per turn, and the average duration of the survey. It also reports the duration of segmented turns. (*M*=mean, *SD*=Standard deviation)

	Evaluators				Evaluations				Dialog Turns	
	Total #	Gender		Age	Evaluators per Turn		Duration per turn [s]		Duration per turn [s]	
		F [%]	M[%]		M	SD	M	SD	M	SD
Target - improvised	1236	56	44	36.3	28.2	4.6	60.0	125.1	1.9	0.7
Target - read	229	61	39	38.7	5.4	1.2	66.0	148.4	1.9	0.8
Other - improvised	917	60	40	36.9	5.4	1.1	61.0	132.5	3.0	2.1
Natural interactions	867	58	42	37.0	5.3	1.0	64.5	160.7	3.0	2.4

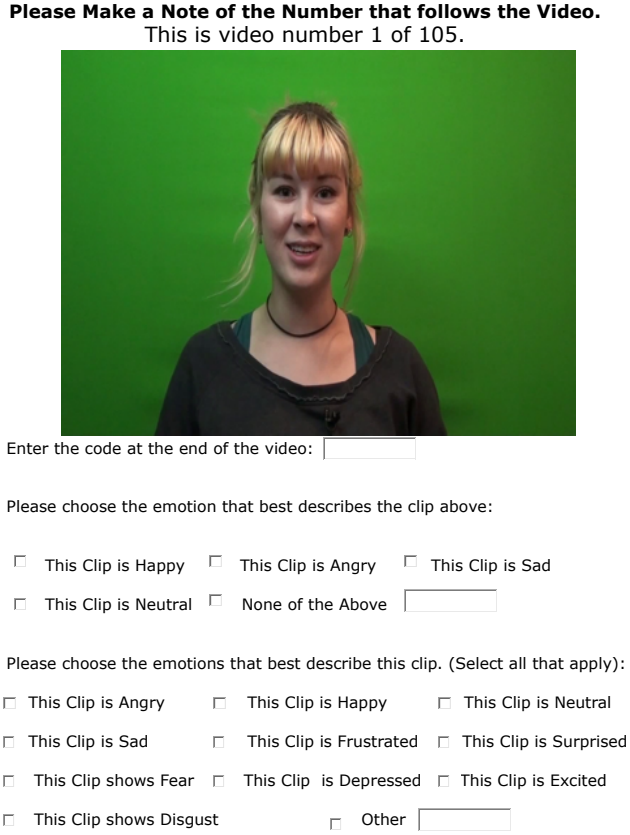


Fig. 2. First part of the perceptual evaluation, including a gold standard question, primary emotional categories (one selection) and secondary emotional categories (multiple selections). The evaluators can propose other labels.

real time the performance of the evaluators. We describe the details of the novel annotation approach in the study of Burmania et al. [19]. Here, we summarize the most important aspects of the approach.

First, we annotate the emotional content of a reference set, which is used to track the performance of the evaluators. We use the *Target - improvised* sentences to create this reference set (652 sentences). Then, we combine utterances whose labels are unknown (7786 turns) with sentences from the reference set (e.g., sequences with five sentences from reference set followed by 20 sentences from the data to be evaluated). We measure whether the inter-evaluator agreement increases or decreases when the assessments of the evaluator are added (categorical emotions). The quality assessment is implemented using only the reference set, where each sentence was previously annotated by other evaluators. By checking the performance over the reference set, we aim to identify unreliable/tired evaluators. The scheme stops the evaluation when the inter-evaluator agreement, measured over the reference set, decreases due to poor performance of the rater, or when his/her consistency drops over time. This process is implemented at different checkpoints during the evaluation. As a result, we increase the inter-evaluator agreement in the data to be evaluated. By measuring changes in inter-evaluator agreement, and not the actual values of the inter-evaluator agreement, we take into consideration that certain sentences are harder to evaluate than others.

The approach significantly improves the inter-evaluator agreement, achieving a Fleiss' Kappa statistic of κ=0.487 for a five class emotion problem (happy, angry, sad, neutral and other). The inter-evaluator agreement is consistent across datasets (*Target - improvised* κ=0.497; *Target - read* κ=0.479; *Other - improvised* κ=0.458; *Natural interaction* κ=0.487). Without the online quality assessment approach, the inter evaluator agreement is only κ=0.4 [19]. The agreement for the corpus (κ=0.487) is similar, or even higher, than the agreement levels reported in controlled perceptual evaluations of spontaneous recordings for similar tasks [12], [14], [57], [58]. The improvement in inter-evaluator agreement justifies the longer time required to collect the evaluations (evaluators in crowdsourcing prefer to complete shorter rather than longer evaluations, even if the payment per video is higher), and the overhead associated with the evaluation (extra evaluations of the videos in the reference set).

Notice that we do not have to throw away annotations (and resources) to improve the agreement. Instead, we just stop the survey when the evaluator is viewed as unreliable. This is a key difference between our method and post-processing frameworks where all the data is collected and then unreliable

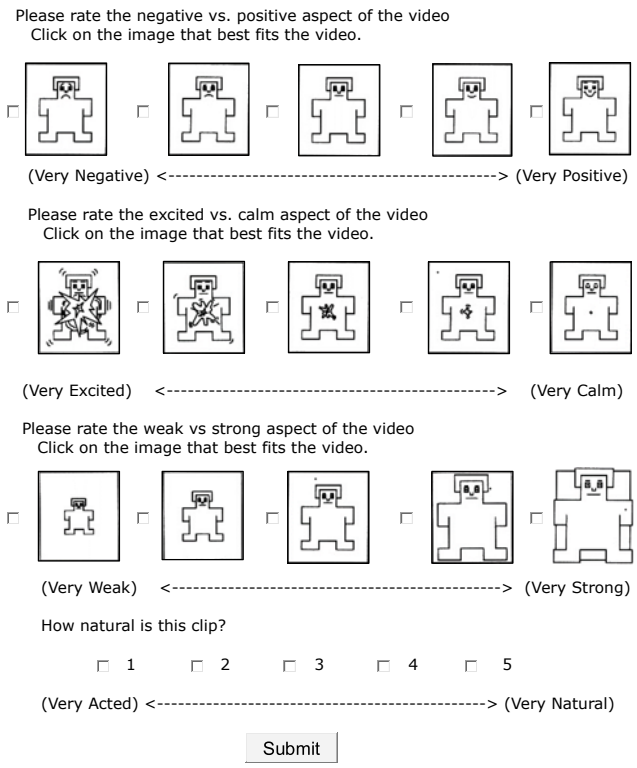


Fig. 3. Second part of the perceptual evaluation, including the SAM for valence (1-negative versus 5-positive), activation (1-excited versus 5-calm) and dominance (1-weak versus 5-strong). The evaluation also includes a five Likert-like scale for naturalness perception (1-very acted, 5-very natural).

evaluators are not considered (wasting resources used to pay these evaluations that are discarded). We use *Target - improvised* sentences as the reference set in the evaluation, since these sentences are emotionally annotated by more evaluators, due to the overhead associated with the annotation process. Sentences for this dataset, which is the most relevant set for our emotional human perception study, are evaluated by 28.2 evaluators on average. The emotional content of the *Other - improvised*, *Natural interaction*, and *Target - read* sentences are annotated by at least five evaluators.The perceptual evaluation tasks were identical across datasets, so the emotional labels can be directly compared.

Figures 2 and 3 describe the surveys used to annotate the emotional content of the corpus, which appear as a single questionnaire. We present the video to the evaluators so they can annotate the emotions after perceiving cues from speech and facial expressions. First, we ask the evaluators to assess the emotional content in terms of 5 classes: happy, angry, sad, neutral and other (see Fig. 2). The evaluator has to choose one of these labels. We expect ambiguous expressive behaviors with mixed emotions, since the corpus has spontaneous interactions [12], [59]. Therefore, after selecting the primary emotion class, we ask the evaluators to select all the emotional classes perceived in the video. We expand the emotional categories including the remaining three basic emotions (surprise, fear, and disgust). We augment this list with emotional states that have been commonly used in other studies such as frustrated, depressed, and excited. We also include the category "other" in case none of these emotions characterize the expressive behaviors. In contrast to the primary emotional category, the evaluators can select all the relevant emotional categories perceived in the clip. This approach is inspired by the work of Vidrascu and Devillers [60]. In their work, they asked the evaluators to annotate secondary emotions, in addition to primary emotions, referred to as "minor".

The second part of the survey includes an emotional annotation using a five-point Likert-like scale in terms of the following attributes: valence (1-negative versus 5-positive), activation (1-excited versus 5-calm), and dominance (1-weak versus 5-strong). These emotional primitives provide complementary information, allowing us to explore the emotional spectrum associated with each primary categorical emotion (e.g., different degree of happiness). These emotional attributes are annotated with *self-assessment manikins* (SAMs) [61] (see Fig. 3). These "manikins" simplify the understanding of the meaning of the attributes, improving the reliability of the perceptual task [62]. Finally, we ask the evaluators to annotate the perceived naturalness of the video using a five-point Likert-like scale (1-very acted, 5-very natural). We use these labels to evaluate the benefits of the proposed elicitation technique.

TABLE 4
Number of sentences per emotional class using majority vote (*A*: Anger, *S*: Sadness, *H*: Happiness, *N*: neutral, *O*: Other, *WA*: without agreement).

Dataset	# sent.	A	S	H	N	O	WA
Target - improvised	652	115	106	136	283	1	11
Target - read	620	169	80	88	241	3	39
Other - improvised	4,381	470	633	1,048	1,789	70	371
Natural interaction	2,785	38	66	1,372	1,164	11	134
Total	8,438	792	885	2,644	3,477	85	555

4.2 Emotional Content Analysis

This section analyzes the emotional content of the corpus for each of the sets (*Target - improvised*, *Target - read*, *Other - improvised*, and *Natural interaction*).

Table 4 reports the number of sentences assigned to each of the four primary emotional classes (anger, sadness, happiness, and neutrality). The classes are assigned using majority vote. The column *WA* (i.e., without agreement) lists the number of sentences in which their evaluations did not reach agreement under this criterion. For *Target - improvised* and *Target - read*, the number of sentences are similarly distributed among

TABLE 5
Confusion matrices for intended versus perceived emotions for target sentences using spontaneous and read elicitation methods. We report in parentheses the percentage of cases where the perceived and intended emotions matched (*WA*: without agreement).

Target - improvised (73.3%)

		Perceived Emotion					
		Angry	**Sad**	**Happy**	**Neutral**	**Others**	**WA**
Intended	**Ang**	107	0	3	52	0	3
	Sad	3	100	0	60	0	3
	Hap	1	1	127	27	1	4
	Neu	4	5	6	144	0	1

Target - read (69.4%)

		Perceived Emotion					
		Angry	**Sad**	**Happy**	**Neutral**	**Others**	**WA**
Intended	**Ang**	128	0	0	20	0	2
	Sad	19	77	0	47	3	24
	Hap	12	0	88	37	0	8
	Neu	10	3	0	137	0	5

the four emotional classes. The emotion distribution for *Natural interaction* sentences is unbalanced. The interaction of the actors during the breaks was mostly positive or neutral given their colloquial conversation. The actors relaxed and engaged in natural, friendly interactions. Given the number of evaluations per video, sentences without agreement include only 6.6% of the corpus, a relatively small proportion compared with other databases.

We evaluate the effectiveness of the proposed elicitation technique by studying the confusion matrix of the target sentences. Table 5 reports the results which show that in 73.3% of the cases, the intended emotional class was actually perceived by the evaluators. The most challenging class to elicit is sadness, which is often confused with neutral speech. Notice that similar confusion trends have been reported in previous studies [63]. We also observe that some sentences that were intended to be angry were perceived as neutral. We compare these results with the confusion matrix for the read sentences. The lexical content of these sentences is identical to that of the target sentences. The only difference is the elicitation approach. Table 5 reports the results which show that the perceived emotion of 69.4% of the sentences matches the intended emotion. We compare the confusion matrices associated with the read and improvised data using the difference of proportions test between the two populations (one-tailed). The test does not reject the null hypothesis that the proportions are equal (p-value = 0.0618, z-value = 1.6449). From the confusion matrices, we estimated the Wagner's unbiased hit rate for each emotional class [64]. The metric compensates for the bias introduced when the number of samples across classes differs. Table 6 shows the results. While we observe differences for each emotion, the average values for *Target - improvised* and *Target - read* are

TABLE 6
Wagner's unbiased hit rate for each emotional class.

	Angry [%]	**Sad** [%]	**Happy** [%]	**Neutral** [%]	**Mean** [%]
Target - improvised	61.5	57.8	76.0	46.1	60.4
Target - read	65.5	51.8	64.2	51.9	58.4

similar suggesting that the proposed approach does not negatively affect our ability to elicit the target emotion, as compared to read renditions. Providing context helps the actors to effectively elicit the target emotion. As a reference, Yildirim et al. [63] reported 68.3% agreement between intended and perceived emotions in controlled perceptual evaluations (instead of crowdsourcing) over acted read sentences.

For a given turn, we derive consensus labels for valence, activation, dominance, and naturalness by averaging the corresponding scores assigned by different evaluators. Figure 4 shows the distribution, mean, and standard deviation for the emotional attributes valence, activation, and dominance. The figure gives the statistics for the *Target - improvised*, *Target - read*, *Other - improvised* and *Natural interaction*. The sentences for *Target - improvised* and *Other - improvised* present similar means across valence, activation and dominance. As expected, the valence values for *Natural interaction* are mostly positive given the colloquial discussion during the breaks. Figure 4 shows that the distribution of the corpus includes samples over the entire valence, activation and dominance space.

Figure 4 shows the results for naturalness perception (1-very acted, 5-very natural). Across the MSP-IMPROV corpus, 78.8% of the turns have average naturalness scores above 3. These percentages vary across datasets: *Target - improvised* 79.0%; *Target - read* 57.6%; *Other - improvised* 76.7%; and, *Natural interaction* 86.8%. While most of the samples from the *Target - improvised* sentences have average naturalness scores above 3, only 57.6% of the *Target - read* sentences satisfy this condition. We use a one-way ANOVA to assess differences in naturalness scores across datasets. The results reveal significant differences, [$F(3, 8432) = 398.27$, $p < 10^{-8}$]. We compute pair-wise comparisons of the means between datasets using one-tailed large-sample population mean test (standard normal z-test statistics), asserting significance when $p < 0.01$. The difference in the average of the naturalness scores across datasets are all statistically significant. Recording the target sentences with the proposed elicitation process increases the naturalness perception from 3.09 (*Target - read*) to 3.33 (*Target - improvised*) ($p = 7.25 * 10^{-17}$). Hence, we achieve recordings of sentences with fixed lexical content that are perceived as more natural than read renditions. The *Other - improvised* sentences are perceived as more natural than the *Target - improvised* sentences (3.56 versus 3.33) ($p = 7.59 * 10^{-38}$). The *Natural interaction*

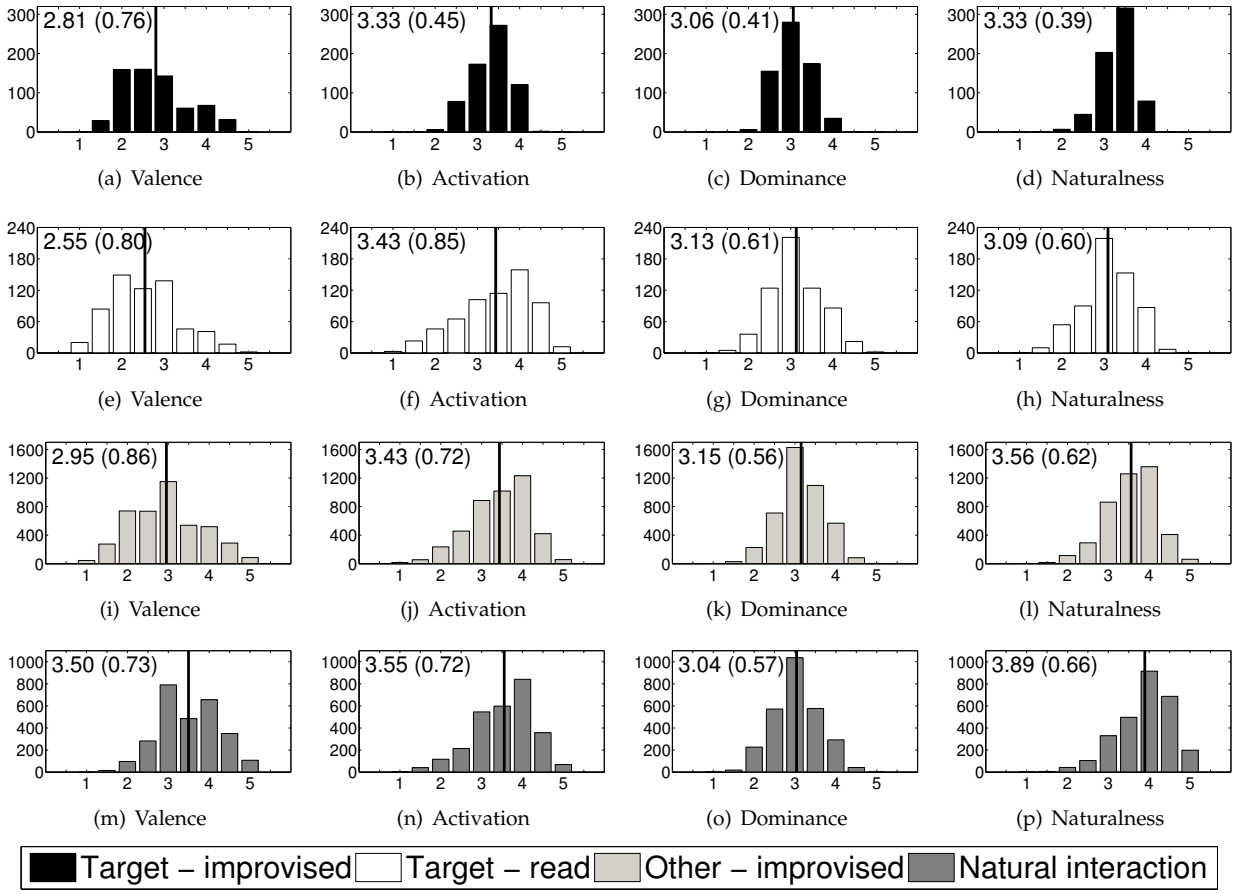


Fig. 4. Distribution for valence (1-negative versus 5-positive), activation (1-excited versus 5-calm), dominance (1-weak versus 5-strong), and naturalness (1-very acted, 5-very natural) for each of the datasets. The figures provide the mean value (also visualized as a vertical line) with the standard deviation between parentheses.

turns are perceived with the highest scores (i.e., more natural), as expected ($p < 10^{-100}$ in all comparisons).

Expressive behaviors during spontaneous interactions convey a mixture of emotions [59]. Even when one sentence is labeled as "happiness", for example, we expect to observe other secondary emotional traits such as excitement or surprise. We can study this aspect with the annotations of secondary emotions. The means (standard deviations in parentheses) of the number of secondary emotions selected by the evaluators are 0.78 (0.87) for *Target - improvised*; 0.84 (0.97) for *Target - read*; 0.82 (0.86) for *Other - improvised*; and 0.54 (0.73) for *Natural interaction*.

The emotional annotation protocol used in this corpus allows us to study the emotion spectrum associated with each of the primary emotions. To address this question, we group all the individual evaluations into five groups according to the emotional label of the primary emotional class (anger, sadness, happiness, neutral and other). For each group, we estimate the distribution of the secondary emotions included in these annotations. Notice that the evaluators were requested to label all the perceived emotions including the primary emotional classes. However, this analysis only considers secondary emotions different from the primary emotions. Figure 5 shows the results. From the evaluations in which the primary emotion was "anger", Figure 5(a) shows that evaluators also perceived traits of frustration (46%) and disgust (21%). Figure 5(d) shows the close relationship between sadness and depressed behaviors. The primary class "other" represents expressive behaviors that are not well characterized by anger, sadness, happiness or neutral state. Figure 5(e) shows that these samples are better described by frustration (18%), surprise (15%), excitement (10%) and disgust (8%). Interestingly, 25% of these turns convey emotional behaviors that are not well represented by the extended set of secondary emotions. Common emotional labels provided by the evaluators are *annoyed*, *concerned*, *confused*, *disappointed*, and *worried* (not shown in Fig. 5). These emotional labels are candidate classes that may be included in future emotional perceptual evaluations (the supplemental material gives a list with the most common terms suggested by evaluators). While the analysis considers the entire MSP-IMPROV, it is interesting that the general trends in Figure 5 are consistent across datasets (*Target - improvised*, *Other*

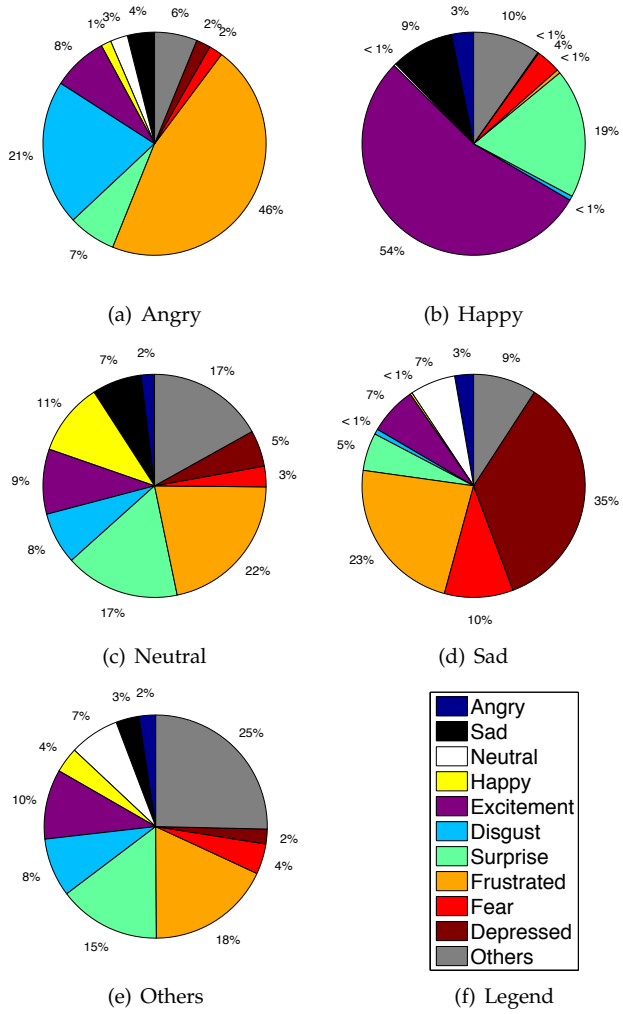


Fig. 5. Distribution of the secondary emotions assigned to each primary emotion.

Improvised, *Target - read*, and *Natural Interaction*). A detailed analysis on the relationship between secondary and primary emotions per dataset is presented as supplemental material.

5 Evaluation of Emotion Recognition

While the MSP-IMPROV database was collected for studying human emotion perception, the corpus can play an important role in emotion recognition. This section provides overall performance scores for speech and facial emotion recognition (Sec. 5.1). We also explore the performance of emotion recognition systems in terms of inter-evaluator agreement (Sec. 5.2), and naturalness (Sec. 5.3).

The primary emotional classes annotated in the perceptual evaluation (with the exception of "other") are used as ground truth for the experiments. We define the consensus labels using the majority vote rule, where sentences without agreement are not considered for the evaluations (555 sentences representing 6.6% of the corpus – see Table 4). All the classification evaluations consist of four-class problems (anger, sadness, happiness and neutrality), implemented with *support vector machine* (SVM) with *sequential minimal optimization* (SMO). We use *radial basis function* (RBF) kernel, setting the soft margin parameter c equal to 1 across evaluations. We use the WEKA toolkit for the evaluation. We implement the classification experiments with *leave-one-speaker-out* (LOSO) 12-fold cross-validation. For each fold, the samples of one of the 12 subjects are used for testing, and the sentences from the remaining 11 subjects are used for training the classifiers.

We estimate acoustic features with OpenSmile [65]. We consider the set provided for the 2009 Interspeech Emotion Challenge [66]. The set includes a set of 384 prosodic, spectral and voice quality features estimated at the sentence level (see [66] for a detailed description of the features). We use this set since it has been commonly used in related studies. Unlike the feature sets proposed for latter editions of the Interspeech challenges (e.g., over 4,000 features), the feature dimension is significantly lower, so we do not need to use feature selection. Therefore, the results can be easily replicated by other groups.

For emotion recognition with facial features, we follow the approach used in our previous studies [67], [68]. We estimated 20 *action units* (AUs) and the three head pose angles using the *computer expression recognition toolbox* (CERT) [69] (AUs: 1-2, 4-7, 9-10, 12, 14-15, 17-18, 20, 23-26, 28, 45; head pose: yaw, pitch, and roll). While CERT is robust to reasonable degree of head rotation, 16 turns were discarded since CERT was not able to track the face. For each dialog turn, we estimate seven global statistics across AUs and head pose angles (minimum, maximum, standard deviation, mean, median, lower quartile and upper quartile). This approach generates a 161-dimensional feature vector, which is used for classification.

We measure the performance of the classifiers in terms of accuracy. Since the data is not emotionally balanced, we also report performance with average precision, $\bar{P}$, and average recall, $\bar{R}$ (i.e., we estimate the precision and recall rate for each of the four classes and we estimate their average values). From these metrics, we compute the F-score defined in equation 1.

$$F = \frac{2\bar{P}\bar{R}}{\bar{P} + \bar{R}} \tag{1}$$

We evaluate the classification results using pairwise comparisons using the large-sample test of population proportion, asserting significance if $p < 0.05$.

5.1 General Performance

We implement emotion recognition experiments using all the sentences from the MSP-IMPROV corpus

TABLE 7
Performance of speech and facial emotion classifiers in terms of accuracy (*Acc.*), average precision (*Pre.*), average recall (*Rec.*) and F-score.

	Set	Acc. [%]	Pre. [%]	Rec. [%]	F-score [%]
Speech	MSP-IMPROV	54.4	46.9	41.4	44.0
	Target - improvised	51.4	47.8	46.9	47.3
	Target - read	51.6	47.1	46.6	46.9
	Other - improvised	49.5	45.0	40.8	42.8
	Natural interaction	64.2	38.9	39.5	39.2
Face	MSP-IMPROV	65.4	58.7	51.0	54.5
	Target - improvised	57.3	54.1	52.4	53.3
	Target - read	58.1	56.4	59.1	57.7
	Other - improvised	62.5	56.7	53.0	54.8
	Natural interaction	73.9	42.4	43.0	42.7

TABLE 8
Classification performance in terms of inter-evaluator agreement. The classifiers are trained and tested with sentences in which 60%, 80% or 100% of the evaluators agreed on one emotional class.

	Agreement	# sent.	Accuracy [%]	Precision [%]	Recall [%]	F-score [%]
Speech	60%	6,918	55.5	48.3	42.4	45.2
	80%	4,676	57.4	50.1	44.7	47.2
	100%	2,372	60.7	51.9	46.6	49.1
Face	60%	6,904	67.0	59.8	52.0	55.6
	80%	4,668	71.2	61.9	56.0	58.8
	100%	2,367	75.5	63.9	59.0	61.4

(first and sixth rows of Table 7). The classifiers have accuracies of 54.4% for acoustic features, and 65.4% for facial features over the 4 classes. The corresponding F-scores are 44.0% and 54.5%, respectively. In general, the classifiers trained with facial features provide better performance than the ones trained with acoustic features. The low classification performance is consistent with the values achieved in other acted spontaneous databases such as the IEMOCAP corpus [70]. We also implement emotion recognition experiments over each of the four types of sentences in the database (*Target - improvised*, *Target - read*, *Other - improvised*, and *Natural interaction*). The results show that the classifier trained with the *Target - improvised* sentences has the best F-scores for acoustic features. For facial features, the classifier trained with *Target - read* sentences achieves the best F-scores. For *Natural interaction*, the accuracy is higher than the accuracy on other portions of the data (both acoustic and facial features). However, this dataset provides the lowest F-scores. These results are due to unbalanced emotional classes, where there are very few sentences for anger (we do not use over-sampling or under-sampling for training and testing the classifiers).

5.2 Performance and Inter-Evaluator Agreement

The large size of the corpus provides an opportunity to evaluate the performance as function of the inter-evaluator agreement. Instead of using majority vote to consolidate the emotional labels, we define three criteria to re-label the emotional class of the sentences. The three criteria require that at least 60%, 80%, or 100% of the evaluators agree on a given emotional class, respectively. Notice that sentences that satisfy the more restrictive thresholds (e.g., 100%) also satisfy the weaker thresholds (e.g., 60%). Table 8 gives the number of files that we consider for each criterion. The classifiers are trained and tested with these sentences using acoustic or facial features, using the same approach described before (SVM classifiers, 12 fold LOSO cross-validation)

Table 8 shows the results. We observe consistent results for classifiers trained with acoustic and facial features. As expected, the classification performance improves as the inter-agreement increases. When we consider sentences with perfect agreement (2,372 sentences), the accuracy and F-score for speech-based classifiers are 60.7% and 49.1%, respectively. For face-based classifiers, the the accuracy and F-score are 75.5% and 61.4%, respectively. These values are between 5% and 8% better than the performance achieved when we use the 60% agreement criterion (first row of Table 8), or majority vote (first row of Table 7). In these cases, the differences are significantly different. Sentences with ambiguous emotional content are excluded from the classification evaluation when we use the 100% agreement criterion. Therefore, the classification performance increases.

5.3 Performance and Naturalness

Finally, we evaluate the classification performance in terms of the naturalness score provided to the samples (i.e., multi-class recognition of anger, happiness, sadness and neutrality). The labels are assigned based on majority vote. We consider four conditions. The first condition includes only the sentences in which the average naturalness score is more than 4 ($t_{(1)} \geq 4$). This set corresponds to the most natural sentences perceived by the evaluators (1-very acted, 5-very natural). For the second, third, and fourth conditions, we use less restricted thresholds set to $t_{(2)} \geq 3$, $t_{(3)} \geq 2$, and $t_{(4)} \geq 1$, respectively. Notice that the last threshold includes all the sentences (i.e., first row of Table 7). Table 9 reports the number of sentences for each of these conditions. Notice that 6,651 samples (78.8% of corpus) have an average naturalness score greater than 3, revealing that the elicitation technique effectively produced naturalistic behaviors.

Table 9 shows the classification performance in terms of their naturalness scores for acoustic and facial features. There are not significant differences in the classification accuracies for $t_{(3)}$ and $t_{(4)}$. Contrary to what we may expect, the differences in the perceived naturalness of these samples do not affect the performance of the classifiers. For $t_{(1)}$, the accuracies are significantly higher than the accuracies for $t_{(3)}$ and

TABLE 9
Classification performance in terms of naturalness scores. The classifiers are trained and tested with sentences in which their average natural scores were less than a threshold.

	Naturalness Scores	# sent.	Accuracy [%]	Precision [%]	Recall [%]	F-score [%]
Speech	$t_{(1)} \geq 4$	2,725	58.0	43.7	42.4	43.1
	$t_{(2)} \geq 3$	6,651	56.0	48.3	42.1	45.0
	$t_{(3)} \geq 2$	7,717	54.6	47.1	41.6	44.2
	$t_{(4)} \geq 1$	7,796	54.4	46.9	41.4	44.0
Face	$t_{(1)} \geq 4$	2,720	70.1	52.4	50.6	51.5
	$t_{(2)} \geq 3$	6,635	66.3	56.6	50.2	53.2
	$t_{(3)} \geq 2$	7,701	65.6	58.4	51.0	54.4
	$t_{(4)} \geq 1$	7,780	65.4	58.7	51.0	54.5

$t_{(4)}$. However, the F-score is lower. Since 53.7% of the samples from $t_{(1)}$ correspond to *Natural interaction*, these classification results present similar trends to the performance of the *Natural interaction* dataset (higher accuracy, lower F-score – see last row in Table 7).

6 Conclusions and Discussion

This study introduced the MSP-IMPROV corpus, a multimodal emotional database comprised of spontaneous dyadic interactions, designed to study audiovisual perception of expressive behaviors. The corpus relied on a novel elicitation scheme, where two actors improvise scenarios that lead one of them to utter target sentences. The context of the emotion-specific scenarios evokes emotional reactions driven by the spontaneous interaction that are perceived as more natural than the read renditions. With this elicitation approach, we recorded spontaneous sentences with the same lexical content, conveying different emotions. These stimuli are ideal for the study of the integration of audio and video cues in emotion perception. In addition to the target sentences, the corpus includes all the turns during the improvisation, and the natural interactions between the actors during the breaks. The *Natural interaction* turns have many sentences with positive valence, given the friendly interaction between the actors. Very few of these sentences are labeled as *anger*. For other portions of the corpus (i.e., *Target - improvised*, and *Target - read*), the corpus provides a large number of sentences for each of the primary emotions. Overall, the corpus consists of 8,438 turns (over 9 hours) of emotional sentences.

The emotional labels for the corpus were collected through crowd-sourced perceptual evaluations. For the *Target - improvised* dataset, the analysis of the emotional content revealed that the proposed elicitation technique was effective in eliciting the target emotion in 73.3% of the cases. The approach is as effective as the approach where actors read sentences portraying target emotions. More importantly, the perceived naturalness for 79% of the *Target - improvised* sentences was over 3 (1-very acted, 5-very natural), where only 57.6% of the *Target - read* sentences satisfied this condition. These results confirm that using context plays an important role in eliciting spontaneous renditions of target emotions. The study uses majority voting to merge the assessments of multiple evaluators. Given the size of the corpus and the number of evaluators per speaking turn, this corpus provides a perfect resource to explore more sophisticated methods to fuse multiple assessments, deriving robust emotional labels [71].

We are currently working on creating the congruent and incongruent audiovisual emotional stimuli for our perceptual evaluation [6]. By studying the emotional perception of these stimuli, we expect to identify primary cues that are important to infer emotions. These studies will provide evidence for the mechanisms underlying audiovisual emotion perception. The studies will also provide insight about machine learning solutions for affective computing.

The features of this multimodal corpus make this database a valuable resource for studies on emotion recognition. The annotation of attribute-based descriptors and primary and secondary emotional categories provides an opportunity to understand the relationship between emotional categories and dimensional attributes (activation, valence and dominance scores). Studies following this direction can lead to practical solutions to address ambiguous emotional behaviors [59] (i.e., creating emotional profiles instead of forcing the system to make a hard decision on emotional behaviors [72]).

Given the potential of this corpus in the field of affective computing, the MSP-IMPROV corpus will be released to the research community through our website (http://ecs.utdallas.edu/research/researchlabs/msp-lab/MSP-Improv.html).

Acknowledgments

This study was funded by National Science Foundation (NSF) grants IIS-1217104 and IIS-1217183.

References

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

[2] ——, "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, October 2007, pp. 43–47.

[3] E. Mower Provost, I. Zhu, and S. Narayanan, "Using emotional noise to uncloud audio-visual emotion perceptual evaluation," in *IEEE International Conference on Multimedia and Expo (ICME 2013)*, San Jose, CA, USA, July 2013.

[4] E. Mower, S. Lee, M. Matarić, and S. Narayanan, "Joint-processing of audio-visual signals in human perception of conflicting synthetic character emotions," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 961–964.

[5] E. Mower, M. Matarić, and S.Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 843–855, August 2009.

[6] E. Mower Provost, Y. Shangguan, and C. Busso, "UMEME: University of Michigan emotional McGurk effect dataset," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 395–409, October-December 2015.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[8] M. Liberman, K. Davis, M. Grossman, N. Martey, and J. Bell, "Emotional prosody speech and transcripts," Philadelphia, PA, USA, 2002, Linguistic Data Consortium.

[9] I. Engberg and A. Hansen, "Documentation of the Danish emotional speech database (DES)," Center for Person Kommunikation, Aalborg University, Aalborg, Denmark, Tech. Rep., September 1996.

[10] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: Towards a new generation of databases," *Speech Communication*, vol. 40, no. 1-2, pp. 33–60, April 2003.

[11] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth, "Desperately seeking emotions or: actors, wizards and human beings," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle, Northern Ireland, UK, September 2000, pp. 195–200.

[12] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, no. 4, pp. 407–422, May 2005.

[13] C. Busso and S. Narayanan, "Recording audio-visual emotional databases from actors: a closer look," in *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 2008, pp. 17–22.

[14] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[15] K. R. Scherer and T. Bänziger, "On the use of actor portrayals in research on emotional expression," in *Blueprint for affective computing: A sourcebook*, K. R. Scherer, T. Bänziger, and E. B. Roesch, Eds. Oxford, England: Oxford university Press., November 2010, pp. 166–176.

[16] K. Scherer, "Vocal markers of emotion: Comparing induction and acting elicitation," *Computer Speech & Language*, vol. 27, no. 1, pp. 40–58, January 2013.

[17] F. Enos and J. Hirschberg, "A framework for eliciting emotional speech: Capitalizing on the actor's 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, May 2006, pp. 6–10.

[18] R. Banse and K. Scherer, "Acoustic profiles in vocal emotion expression." *Journal of Personality and Social Psychology*, vol. 70, no. 3, pp. 614–636, March 1996.

[19] A. Burmania, S. Parthasarathy, and C. Busso, "Increasing the reliability of crowdsourcing evaluations using online quality assessment," *IEEE Transactions on Affective Computing*, vol. To Appear, 2015.

[20] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, March 2011.

[21] D. Ververidis and C. Kotropoulos, "A state of the art review on emotional speech databases," in *First International Workshop on Interactive Rich Media Content Production (RichMedia-2003)*, Lausanne, Switzerland, October 2003, pp. 109–119.

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

[23] C. Busso, S. Mariooryad, A. Metallinou, and S. Narayanan, "Iterative feature normalization scheme for automatic emotion detection from speech," *IEEE Transactions on Affective Computing*, vol. 4, no. 4, pp. 386–397, October-December 2013.

[24] T. Rahman and C. Busso, "A personalized emotion recognition system using an unsupervised feature adaptation scheme," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012)*, Kyoto, Japan, March 2012, pp. 5117–5120.

[25] L. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in *Proceedings of the 13th international conference on multimodal interfaces (ICMI 2011)*, Alicante, Spain, November 2011, pp. 169–176.

[26] C. Lee and S. Narayanan, "Toward detecting emotions in spoken dialogs," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, March 2005.

[27] L. Devillers and L. Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Interspeech - International Conference on Spoken Language (ICSLP)*, Pittsburgh, PA, USA, September 2006, pp. 801–804.

[28] S. Steidl, "Automatic classification of emotion-related user states in spontaneous children's speech," Ph.D. dissertation, Universität Erlangen-Nürnberg, Erlangen, Germany, January 2009.

[29] D. Rösner, J. Frommer, R. Friesen, M. Haase, J. Lange, and M. Otto, "LAST MINUTE: a multimodal corpus of speech-based user-companion interactions," in *International conference on Language Resources and Evaluation (LREC 2012)*, Istanbul, Turkey, May 2012, pp. 2559–2566.

[30] F. Schiel, S. Steininger, and U. Türk, "The SmartKom multimodal corpus at BAS," in *Language Resources and Evaluation (LREC 2002)*, Las Palmas, Spain, May 2002.

[31] E. Mower, M. Black, E. Flores, M. Williams, and S. Narayanan, "Rachel: Design of an emotionally targeted interactive agent for children with autism," in *IEEE International Conference on Multimedia and Expo (ICME 2011)*, Barcelona, Spain, July 2011.

[32] G. Le Chenadec, V. Maffiolo, N. Chateau, and J. Colletta, "Creation of a corpus of multimodal spontaneous expressions of emotions in human-machine interaction," in *International conference on Language Resources and Evaluation (LREC 2006)*, Genoa,Italy, May 2006, pp. 24–26.

[33] N. Amir, S. Ron, and N. Laor, "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, September 2000, pp. 29–33.

[34] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, January-March 2012.

[35] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE 2013)*, Shanghai, China, April 2013.

[36] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, January-March 2012.

[37] A. Tcherkassof, D. Dupré, B. Meillon, N. Mandran, M. Dubois, and J.-M. Adam, "DynEmo: A video database of natural facial expressions of emotions," *The International Journal of Multimedia & Its Applications (IJMA)*, vol. 5, no. 5, pp. 61–80, October 2013.

[38] I. Sneddon, M. McRorie, G. McKeown, and J. Hanratty, "The Belfast induced natural emotion database," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 32–41, January-March 2012.

[39] C. Busso, M. Bulut, and S. Narayanan, "Toward effective automatic recognition systems of emotion in speech," in *Social emotions in nature and artifact: emotions in human and human-computer interaction*, J. Gratch and S. Marsella, Eds. New York, NY, USA: Oxford University Press, November 2013, pp. 110–127.

[40] S. Carnicke, *Stanislavsky in Focus: An Acting Master for the Twenty-First Century*. Abingdon Oxon, UK: Routledge, Taylor & Francis Group, September 2008.

[41] A. Metallinou, Z. Yang, C.-C. Lee, C. Busso, S. Carnicke, and S. Narayanan, "The USC CreativeIT database of multimodal dyadic interactions: From speech and full body motion capture to continuous emotional annotations," *Journal of Language Resources and Evaluation*, vol. accepted, 2015.

[42] K. Scherer, "Appraisal theory," in *Handbook of cognition and emotion*, T. Dalgleish and J. Power, Eds. New York, NY, USA: John Wiley & Sons Ltd, March 1999, pp. 637–663.

[43] I. Lefter, G. Burghouts, and L. Rothkrantz, "An audio-visual dataset of human-human interactions in stressful situations," *Journal on Multimodal User Interfaces*, pp. 1–13, April 2014.

[44] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Transactions on Affective Computing*, 2014.

[45] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The eNTERFACE'05 audio-visual emotion database," in *Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW 2006)*, Atlanta, GA, USA, April 2006.

[46] T. Bänziger, H. Pirker, and K. Scherer, "GEMEP - Geneva multimodal emotion portrayals: a corpus for the study of multimodal emotional expressions," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, Genoa,Italy, May 2006, pp. 15–19.

[47] M. Shah, D. Cooper, H. Cao, R. Gur, A. Nenkova, and R. Verma, "Action unit models of facial expression of emotion in the presence of speech," in *Affective Computing and Intelligent Interaction (ACII 2013)*, Geneva, Switzerland, September 2013, pp. 49–54.

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

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

[50] G. McKeown, W. Curran, C. McLoughlin, H. Griffin, and N. Bianchi-Berthouze, "Laughter induction techniques suitable for generating motion capture data of laughter associated body movements," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2013)*, Shanghai, China, April 2013.

[51] S. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, March 2013.

[52] A. Tarasov, S. Delany, and C. Cullen, "Using crowdsourcing for labelling emotional speech assets," in *W3C workshop on Emotion ML*, Paris, France, October 2010.

[53] S. Mariooryad, R. Lotfian, and C. Busso, "Building a naturalistic emotional speech corpus by retrieving expressive behaviors from existing speech corpora," in *Interspeech 2014*, Singapore, September 2014, pp. 238–242.

[54] R. Snow, B. O'Connor, D. Jurafsky, and A. Ng, "Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks," in *Conference on empirical methods in natural language processing (EMNLP 2008)*, Honolulu, HI, USA, October 2008, pp. 254–263.

[55] J. Ross, L. Irani, M. Silberman, A. Zaldivar, and B. Tomlinson, "Who are the crowdworkers?: shifting demographics in Mechanical Turk," in *Proceedings of the 2011 annual conference extended abstracts on Human factors in computing systems*, ser. CHI EA '10, April 2010, pp. 2863–2872.

[56] R. Cauldwell, "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, September 2000, pp. 127–131.

[57] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, no. 10-11, pp. 787–800, October-November 2007.

[58] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann, ""Of all things the measure is man" automatic classification of emotions and inter-labeler consistency," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2005)*, vol. 1, Philadelphia, PA, USA, March 2005, pp. 317–320.

[59] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *International Conference on Affective Computing and Intelligent Interaction (ACII 2009)*, Amsterdam, The Netherlands, September 2009.

[60] L. Vidrascu and L. Devillers, "Annotation and detection of blended emotions in real human-human dialogs recorded in a call center," in *IEEE International Conference on Multimedia and Expo (ICME 2005)*, Amsterdam, The Netherlands, July 2005.

[61] P. Lang, "Behavioral treatment and bio-behavioral assessment: computer applications," in *Technology in mental health care delivery systems*, J. B. Sidowski, J. H. Johnson, and T. A. Williams, Eds. Norwood, NJ, USA: Ablex Pub, January 1980, pp. 119–137.

[62] M. Grimm and K. Kroschel, "Evaluation of natural emotions using self assessment manikins," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2005)*, San Juan, Puerto Rico, December 2005, pp. 381–385.

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

[64] H. Wagner, "On measuring performance in category judgment studies of nonverbal behavior," *Journal of Nonverbal Behavior*, vol. 17, no. 1, pp. 3–28, March 1993.

[65] F. Eyben, M. Wöllmer, and B. Schuller, "OpenSMILE: the Munich versatile and fast open-source audio feature extractor," in *ACM International conference on Multimedia (MM 2010)*, Florence, Italy, October 2010, pp. 1459–1462.

[66] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 emotion challenge," in *Interspeech 2009 - Eurospeech*, Brighton, UK, September 2009, pp. 312–315.

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

[68] ——, "Facial expression recognition in the presence of speech using blind lexical compensation," *IEEE Transactions on Affective Computing*, vol. to appear, 2015.

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

[70] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, February 2014.

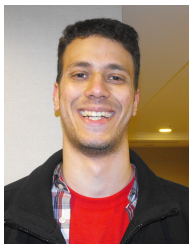
[71] K. Audhkhasi and S. Narayanan, "Data-dependent evaluator modeling and its application to emotional valence classification from speech," in *Interspeech 2010*, Makuhari, Japan, September 2010, pp. 2366–2369.

[72] E. Mower, M. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotional profiles," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1057–1070, May 2011.

Carlos Busso (S'02-M'09-SM'13) received the BS and MS degrees with high honors in electrical engineering from the University of Chile, Santiago, Chile, in 2000 and 2003, respectively, and the PhD degree (2008) in electrical engineering from the University of Southern California (USC), Los Angeles, in 2008. He is an associate professor at the Electrical Engineering Department of The University of Texas at Dallas (UTD). He was selected by the School of Engineering of Chile as the best electrical engineer graduated in 2003 across Chilean universities. At USC, he received a provost doctoral fellowship from 2003 to 2005 and a fellowship in Digital Scholarship from 2007 to 2008. At UTD, he leads the Multimodal Signal Processing (MSP) laboratory [http://msp.utdallas.edu]. He is a recipient of an NSF CAREER Award. In 2014, he received the ICMI Ten-Year Technical Impact Award. He also received the Hewlett Packard Best Paper Award at the IEEE ICME 2011 (with J. Jain). He is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge. His research interests include digital signal processing, speech and video processing, and multimodal interfaces. His current research includes the broad areas of affective computing, multimodal human-machine interfaces, modeling and synthesis of verbal and nonverbal behaviors, sensing human inter- action, in-vehicle active safety system, and machine learning methods for multimodal processing. He is a member of ISCA, AAAC, and ACM, and a senior member of the IEEE.

Srinivas Parthasarathy received his BS degree in degree in Electronics and Communication Engineering from College of Engineering Guindy, Anna University, Chennai, India (2012) and MS degree in Electrical Engineering from the University of Texas at Dallas - UT Dallas (2014). During the academic year 2011-2012, he attended as an exchange student The Royal Institute of Technology (KTH), Sweden. He is currently pursuing his Ph.D in Electrical Engineering at UT Dallas. At UT Dallas, he received the Ericsson Graduate Fellowship during 2013-2014. He joined the Multimodal Signal Processing (MSP) laboratory in 2012. In summer and fall 2014 he interned at Bosch Research and Training Center working on Audio Summarization. His research interest includes the area of affective computing, human machine interaction, and machine learning.

Alec Burmania (S'12) is a senior at the University of Texas at Dallas (UTD) majoring in Electrical Engineering. He works as an undergraduate researcher in the Multimodal Signal Processing (MSP) laboratory. He attended the Texas Academy of Mathematics and Science at the University of North Texas (UNT). He is the Technical Project Chair for the IEEE student branch at UTD for the 2014-2015 school year, and served as the secretary of the branch for the 2013-2014 school year. His student branch was awarded Outstanding Large Student Branch for region 5 two years in a row (2012-2013 & 2013-2014). During the fall of 2014 he received undergraduate research awards from both UTD and the Erik Jonsson School of Engineering and Computer Science. His current research projects and interests include crowdsourcing and machine learning with a focus on emotion.



Mohammed AbdelWahab (S'14) received his B.Sc. degree in electrical and electronic engineering at Ain Shams University, Cairo, Egypt in 2010, and his M.S degree in Electrical engineering from Nile university, Cairo, Egypt 2012. He is currently pursuing his Ph.D. degree in electrical engineering at the University of Texas at Dallas. His current research interest includes speech signal processing, emotion recognition, artificial intelligence and machine learning.

Najmeh Sadoughi (S'14) received her BSC and MS in Biomedical Engineering-Bioelectric from Amirkabir University of Technology (AUT), Tehran, Iran. She started her PHD in Electrical Engineering at the University of Texas at Dallas (UTD) in 2013. She is a member of Multimodal Signal Processing (MSP) Laboratory as a research assistant. She received a Jonsson School Graduate Scholarship (2013-2014). Her research interest includes analysis of the relationship between speech and nonverbal behaviors, synthesis of meaningful nonverbal behaviors for Conversational Agents, and machine learning applications in video and audio processing.

Emily Mower Provost (S'07-M'11) is an Assistant Professor in Computer Science and Engineering at the University of Michigan. She received her B.S. in Electrical Engineering (summa cum laude and with thesis honors) from Tufts University, Boston, MA in 2004 and her M.S. and Ph.D. in Electrical Engineering from the University of Southern California (USC), Los Angeles, CA in 2007 and 2010, respectively. She is a member of Tau-Beta-Pi, Eta-Kappa-Nu, and a member of IEEE and ISCA. She has been awarded the National Science Foundation Graduate Research Fellowship (2004-2007), the Herbert Kunzel Engineering Fellowship from USC (2007-2008, 2010-2011), the Intel Research Fellowship (2008-2010), and the Achievement Rewards For College Scientists (ARCS) Award (2009 2010). She is the co-author of the winner paper of the Classifier Sub-Challenge event at the Interspeech 2009 emotion challenge (with C. Busso). Her research interests are in human-centered speech and video processing, multimodal interfaces design, and speech-based assistive technology. The goals of her research are motivated by the complexities of human emotion generation and perception.