Published in final edited form as:
IEEE Trans Affect Comput. 2014 ; 5(4): 377–390. doi:10.1109/TAFFC.2014.2336244.

CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset

Houwei Cao,
Radiology Department at the University of Pennsylvania, 3600 Market Street, Suite 380, Philadelphia, PA 19104. Houwei.Cao@uphs.upnne.edu

David G. Cooper,
Math and Computer Science Department at Ursinus College, 601 E. Main Street, Collegeville, PA, 19426. dgc@alumni.cmu.edu

Michael K. Keutmann,
Department of Psychology at the University of Illinois at Chicago, 1007 West Harrison Street, M/C 285, Chicago, IL, 60607. michaelk@alumni.upenn.edu.

Ruben C. Gur,
Neuropsychiatry section of the Psychiatry Department of the University of Pennsylvania 3400 Spruce Street, 10th Floor, Gates Bldg.,and the Philadelphia Veterans Administration Medical Center, Philadelphia, PA 19104. gur@mail.med.upenn.edu

Ani Nenkova, and
Department of Computer and Information Science, University of Pennsylvania, 3330 Walnut Street, Philadelphia, PA 19104. nenkova@seas.upenn.edu.

Ragini Verma
Radiology Department at the University of Pennsylvania, 3600 Market Street, Suite 380, Philadelphia, PA 19104.

Abstract

People convey their emotional state in their face and voice. We present an audio-visual data set uniquely suited for the study of multi-modal emotion expression and perception. The data set consists of facial and vocal emotional expressions in sentences spoken in a range of basic emotional states (happy, sad, anger, fear, disgust, and neutral). 7,442 clips of 91 actors with diverse ethnic backgrounds were rated by multiple raters in three modalities: audio, visual, and audio-visual. Categorical emotion labels and real-value intensity values for the perceived emotion were collected using crowd-sourcing from 2,443 raters. The human recognition of intended emotion for the audio-only, visual-only, and audio-visual data are 40.9%, 58.2% and 63.6% respectively. Recognition rates are highest for neutral, followed by happy, anger, disgust, fear, and sad. Average intensity levels of emotion are rated highest for visual-only perception. The accurate recognition of disgust and fear requires simultaneous audio-visual cues, while anger and happiness can be well recognized based on evidence from a single modality. The large dataset we introduce can be used to probe other questions concerning the audio-visual perception of emotion.

Keywords

Emotional corpora; facial expression; multi-modal recognition; voice expression

1 Introduction

Emotion is conveyed by both facial and vocal expressions. Numerous prior studies followed Ekman's basic discrete emotion theory and concentrated on emotion perception from facial cues. They have established that prototypical basic emotions can be universally recognized by different groups of people based on the activation of specific facial expressions [1, 2, 3]. Parallel research exists in vocal expression [4, 5]. Many prosodic features, e.g., pitch, duration, loudness, voice quality, etc. contribute to the transmission of emotional content in voice. Although acoustic correlates are subject to large inter-speaker variability, it has been commonly found that pitch is the most important feature in emotion communication, followed by duration and loudness [6, 7, 8]. In addition, recent computational emotion recognition systems use spectral features to improve recognition [9, 10, 11]. Research on bimodal perception of emotional expressions in faces and voices together has received considerable attention only in recent years. It is well-established that facial expressions convey more information about a subject's emotional state than changes in voice, which typically convey arousal [12, 13]. Some studies investigate how audio and video information is integrated and focus particularly on the issues arising when conflicting interpretations of the facial and vocal expression are possible [14, 15]. However, we still do not have a full understanding of the interplay between the two modalities. Specifically, it is of interest to know how often and for which emotion the audio and visual modalities exhibit complementarity (i.e. when the combination of modalities create an impression different than either of the individual modalities create), dominance (when the two modalities create impressions of different emotions, one of which matches the impression from multi-modal perception) and redundancy (when the two separate modalities create impressions of the same emotion that match the impression from multi-modal perception).

Datasets consisting of emotion expressions are available that contain rated visual displays [16, 17, 18, 19], auditory stimuli [20, 21, 22, 23, 24] and audio-visual clips [25, 26, 27, 28, 29], but these are relatively small, with perceptual rating only from a few raters or do not contain independent ratings for each modality of the same recording. As a result, researchers have been limited in their ability to tease apart the contribution of visual and audio modalities to the perception of intended emotion.

We present the Crowd-sourced Emotional Multi-modal Actors Dataset, CREMA-D, a labeled data set for the study of multimodal expression and perception of basic acted emotions. The large set of expressions was collected as part of an effort to generate standard emotional stimuli for neuroimaging studies, and these require varying degrees of length and intensity and separation of visual and auditory presentation modalities. Actors were used to generate these stimuli because they have training in expressing emotions clearly at varying levels of intensity. They were coached by professional theatre directors. CREMA-D includes 7,442 clips from 91 actors and actresses, with diverse age and ethnicity, expressing six of

the seven “universal emotions” [30]: happy, sad, anger, fear, disgust, and neutral (surprise was not considered by the acting directors to be sufficiently specific, as it could relate to any of the other emotions with rapid onset). The data set contains a collection of multiple perceptual ratings for each clip in three different modalities (audio-only, visual-only, or audio-visual) that were submitted for validation by 2,443 raters using crowd sourcing. Each clip from each modality received two types of emotion perception ratings--categorical emotion label and intensity. Multiple ratings for the same clip are summarized in group perceived emotion labels, i.e. the emotion selected most frequently by the raters; intensity for the group perceived emotion is the average intensity specified by the raters for that emotion. In addition, from each of the three perceptual perspectives, each clip is categorized into one of three subsets depending on the combination of intended and group perceived emotions. Matching clips are those where the most common label from the perceptual ratings is assigned by the majority of raters and it matches the intended emotion. Non-matching clips are those where the majority of raters identified an emotion that is different from the intended emotion. Finally, the ambiguous clips are those without a majority group perceptual agreement.

Our work offers the first data set to address the question of audio-visual perception, meeting the following 5 criteria: (1) over 50 actors (2) over 5,000 clips (3) 6 emotional categories (4) at least 6 ratings per clip (5) 3 modalities. At most three of these five criteria are satisfied for older datasets, as illustrated in Table 1. The GEMEP database [31] includes ratings with 15 categories, with at least 23 raters per clip, using all three rating modalities. However, the clips come from only 10 actors with 1,260 rated clips. De Silva et al. [32] investigated the interactions between audio and visual perception with 6 categories, 18 raters per clip, and used all three modalities, however there were only 2 actors and 72 clips rated. Mower Provost et al. [15] discussed an audio-visual dataset created using the McGurk effect paradigm with matching and mismatching vocal and facial information. The dataset was collected by crowd sourcing 72 original and 216 re-synthesized clips from only 1 actress. Collignon et al.'s AV Integration data set combines silent video clips (the subject is not talking while expressing emotion on the face) and speech audio clips from separate sources in order to create the rated emotional displays as well as mismatched emotions. That dataset only compares fear and disgust [33]. The AV Synthetic Character study uses animated video clips that are synchronized with human audio clips [34] to make the audio-visual stimuli. Some audio-visual databases have a set of audio data with audio ratings and video data with video ratings, but no mixed audio-visual data with corresponding ratings [35, 36]. IEMOCAP [37] and the Chen bimodal databases [38] have an extensive collection protocol, but only include ratings when both the audio and visual information is present. Similarly, the HUMAINE [39] and RECOLA [40] databases have many different types of labels to get at different aspects of audio-visual emotional expression, but all annotations are made for portions of the recording where both vocal and facial information is present. The CHAD database has over 5000 audio-visual clips with 7 emotional categories and 120 raters per clip, but only the audio is rated [41]. The MAHNOB-HCI [42] database is a recent audio-visual database of participants watching emotional videos that has self-reported emotion labels. GEMEP database includes 1,260 audio-visual expressions with 15 emotions

expressed by 10 actors. Each expression is evaluated by 18 -20 raters [31]. Further details on other data-sets of emotion expression can be found in [31, 43].

Our dataset has a large number of actors and raters, which allows us to study the variation in successful emotion communication. It also provides crowd-sourced ratings in all three modalities, allowing for comparisons between single mode and multimodal perception.

Crowd sourcing gives us a large number of total raters, which increases the ecological validity of the ratings. In addition, crowd sourcing allows us to create tasks that do not require too many ratings from any single rater. This protects against bias to the judgments of specific raters. Crowd sourcing also allows us to collect a larger number of ratings per clip and enables us to study the variation of emotion perception of the intended emotions.

CREMA-D explicitly distinguished between matching, non-matching and ambiguous clips. The dataset will be released with classes of emotional expressions and levels of ambiguity. These characteristics permit the development of applications needing subtle and un-prototypical emotion expressions. In some studies, such as the Berlin dataset [8], the matching clips are the only ones released as part of the dataset. Some datasets, such as IEMOCAP [37] and FAU Aibo [24], do not characterize stimuli ambiguity at all. In our dataset however we may retrieve prototypical/non-prototypical emotion expressions by further analyzing the evaluation file distributed as part of the dataset. In most studies there is no mention of such a distinction.

The rest of the paper presents the full details of our data acquisition effort and a first analysis of the interplay between modalities in emotion expression. The aim of the presented analysis is to highlight the potential of the dataset and the various applications in which the corpus can be used. Stimuli preparation is described in Section 2; the acquisition of emotion expressions in Section 2.1, and the ratings protocol and crowd-sourcing method in Section 2.2. Section 3 presents the cleaning of spurious responses from the raw crowd sourcing data and a discussion of how we defined the group response for each emotion expression clip. In Section 4, we further investigate the variability of our dataset and discuss the advantages of the wide range of emotion expressions involved in our collected dataset. In Section 5, we explore how people perceive emotions in terms of audio, video, and audio-visual modalities and further discuss the interaction between modalities in the perception of different emotions.

2 Data Preparation

2.1 Audio-Visual Stimuli

The stimuli for our study include video recordings of professional actors working under the supervision of professional theatre directors. The task for the actors was to convey that they are experiencing a target emotion while uttering a given sentence. The director guided the actors by describing scenarios specially designed to evoke the target emotion. An example scenario for happy is: “Ask the actor what their favorite travel destination is. Tell them they’ve just won an all-expenses paid 10-day trip to the destination.” Some actors preferred to work with personal experiences rather than the scripted scenarios to evoke the emotion

and were allowed to do so. The actors acted out a given sentence in a specific target emotion until the director approved the performance. Two directors were required due to the large number of participants who needed to be scheduled.

There are 91 actors, 48 male and 43 female (51 actors worked with one director, 40 with the another). The actors were between the ages of 20 and 74 with a mean age of 36. Table 2 provides detailed age information. Several racial and ethnic backgrounds were represented in the actor group: Caucasian, African American, Hispanic, and Asian. Table 3 provides a detailed breakdown of the racial and ethnic groups.

Recording sessions typically lasted about three hours and the full sessions were captured on video. Recording sessions took place in a sound attenuated environment with professional light boxes. Actors were seated against a green screen for ease of post-production editing. Videos were recorded on a Panasonic AG-HPX170 at a resolution of 960x720 in the DVCPRO HD format. A directional, far-field microphone is used to collect the audio signal. The final emotional video clips were manually extracted from the full recording of raw digital video files and converted to MP4 using H.264/MPEG-4 Part 10 compression for video and AAC compression at 48 kHz for audio. Videos were further converted to Adobe Flash video, cropped from widescreen to a full screen aspect ratio (4:3).

The target emotions were happy, sad, anger, fear, disgust, as well as neutral. There are 12 sentences, each rendered in all of the emotional states. The actors were directed to express the first sentence in three levels of intensity: low, medium, and high. For the remaining 11 sentences the intensity level was unspecified. It would have been prohibitively expensive to record all sentences for three levels of emotion. The expressions of the one sentence that we did collect can be used in pilot studies for feature analysis of features related to the expression of emotion or as tiered test data to quantify the change of detection capabilities at different intensities.

The semantic content of all 12 sentences was rated as emotionally neutral in a prior study [44]. The 12 sentences were:

- It's eleven o'clock.
- That is exactly what happened.
- I'm on my way to the meeting.
- I wonder what this is about.
- The airplane is almost full.
- Maybe tomorrow it will be cold.
- I would like a new alarm clock
- I think I have a doctor's appointment.
- Don't forget a jacket.
- I think I've seen this before.

- The surface is slick.
- We'll stop in a couple of minutes.

According to the data acquisition design, the data should consist of 7,462 clips, where a clip is the rendition from actor A of a target emotion E while speaking sentence S. However, technical issues prevented 20 clips from being extracted from the original videos. Three of the sentences were missing all clips from one actor, accounting for 18 of the missing clips. In addition, two sentences were missing one neutral clip from one actor.

2.2 Perceptual Ratings From Crowd Sourcing

We collected perceptual ratings for three versions of the clips in the actors' database: the original audio-visual video clip, audio-only, and visual-only. Each version corresponds to a modality of perception. With 7,442 original clips, there are a total of 22,326 items for annotation. The goal was to collect 10 ratings for each item, for a total of 223,260 individual ratings.

We created a perceptual survey using Adobe Flash and utilized crowd sourcing to obtain ratings. We hired raters through Survey Sampling International (SSI), which specializes in providing support for survey research and recruits subjects for this purpose. SSI advertised the task to their pool of participants and recruited the 2,443 raters who completed our task.

The raters were between the ages of 18 and 89 with a mean age of 43. Of them 40.5% were male and 59.5% female. They were mostly Caucasian, but some African American, Hispanic, and Asian raters participated as well. Table 4 shows the distribution of race and ethnicity for the raters next to the actors for comparison.

Each rater was allowed to do only a single session of annotation. At the beginning of the session, participants were asked to use headphones, and a two-part listening task was given before the rating started in order to make sure that quiet speech was audible and loud actor speech was tolerable while using headphones. The listening task consisted of samples of words recorded at low, medium, and high intensity. The low intensity approximated the quietest speech from the actors' clips. The high intensity approximated the level and quality of the loudest speech from the actors' clips. The medium intensity was recorded at a level in between the high and the low intensity.

First, raters were asked to complete a sound calibration task. Three word samples are used as examples of low, medium, and high volume clips as shown in Fig. 1 on the left. The participant is asked to listen to the word samples as many times as needed and to adjust the volume in order to hear the low volume sound without the high volume sound being too harsh. Once they are satisfied with their sound settings, each participant moves from the calibration task to the sound test. The sound test allows up to three trials to recognize three new word samples at low, medium, and high levels. Each trial starts with a single audio presentation of the three samples. Then raters have to select the words they heard from a list of twelve words displayed on a grid, shown on the right panel of Fig. 1. When a participant correctly selects the three target words in the trial, they pass the sound test and can move on

to the perceptual rating task. Participants who fail three sound trials are not permitted to continue with the perceptual rating.

12 words were used as target words so that there would be no repetition of words in either the sound calibration task or any of the trials in the sound test. In addition, 27 foil words were chosen so that each trial would have 9 unique alternate choices on the multiple-choice grid. The target words and the foil words were chosen from WRAT4 reading lists [45] so that all participants would know the words. For each rater, the 12 target words and the 27 foil words are pseudo-randomly ordered, and each word is used only once per rater.

Upon completion of the sound test, raters were given a description of the rating task, which is to specify two ratings of emotion expressed in each clip presented to them. The first rating is the emotional category. Raters can select an emotion from the target emotions (anger, disgust, fear, happy, sad) or select “No Emotion” if they do not perceive any specific emotions. The second rating is a continuous value, which corresponds to the intensity of the emotional state if an emotion is recognized or the confidence level that there is no perceivable emotion if “No Emotion” is selected.

Raters were then presented with three sections consisting of clips with a single modality: first audio-only, then visual-only, and finally audio-visual (the original clip). Each section began with a short description of the respective modality, followed by two practice questions so that the raters could get used to the form of the presentation and the interface for selecting emotions and intensities. Once the practice questions were completed, a screen would indicate the beginning of the real questions for the section. Each clip was presented, then the video display became black for visual stimuli, and raters were asked to select an emotion. Once an emotion was selected, an option for intensity rating is displayed below the selected emotion. This sequence is shown in Fig. 2. The intensity rating is provided on a continuous sliding scale that is scored from 0 to 100. If someone wanted to change their selection, they could press the reset button to go back to the emotion selection dialog. After the ‘Continue’ button was pressed, the next clip would be presented. This would continue until the end of each section.

Each rater annotated one list of assignments with 105 items each. Each list of assignments had three parts, corresponding to each modality, which were always presented in the same order: 35 audio-only, 35 visual-only, and 35 audio-visual videos. In each modality, there were two practice items presented at the beginning, 30 items for the CREMA-D dataset, and three items that were repeated in order to test the consistency of ratings. The repeated items were randomly interspersed among those for the final dataset; among these, the first rating provided was used in the dataset, the second one was only used to check consistency. Across the modalities in each list of assignments, there were no items associated with the same original clip. The same actor is however occasionally seen in items from different modalities.

We created 2,490 assignment lists to guarantee that each item is rated at least 10 times. Each sentence was assigned a number from 1 to 7,442. Each rating session was assigned the next available list of 90 unique sentences (30 audio-only, 30 visual-only, and 30 audio-visual).

Each of the three sub-lists for each section of the survey was pseudo-randomly permuted. Then 9 file numbers (3 for each section) were selected at pre-specified locations for duplication, and the duplicated values were inserted at pre-specified positions. These duplicated values are used for the consistency test.

Some raters did not complete the session. They either stopped before rating all of their clips, had technical difficulties, or failed the sound test. Since incomplete sessions suggest lack of interest in the task, ratings from incomplete sessions were discarded and the assigned list was added back to the available lists. This required manual intervention and left us with 47 incomplete lists.

We lost about 1% of ratings (or 733 ratings per mode) in transit from the flash program to our server. Despite of this, only about 1,700 stimuli have fewer than 10 ratings in each mode. This is due a combination of the data lost in transit and the 47 lists that were not completed. With this data loss, there were still more than 95% of clips with 9 ratings or more before the data was cleaned, and 8 ratings or more after the data was cleaned. The data cleaning is described in the next section.

3 Data Cleaning

3.1 Producing a Reliable Dataset

A concern in crowd-sourcing data acquisition is “cheating” or “poor” responses [46, 47, 48, 49, 50, 51]. Our study was designed to be less attractive to cheaters as suggested by Eickhoff and de Vries [48]. Cheating on our task does not gain the participant too much time and each participant is given a unique order of presentation, so raters cannot copy from each other.

To eliminate poor responses due to distraction, we remove responses that exceed a threshold of 10 seconds for the initial response. The average response time for all responses is 3 seconds and median response time is 1.7 seconds. Taking the median and mean as reference points, it may be safe to assume that participants who took more than 10 seconds to respond were most likely distracted when answering the question. As a result, 7,687 responses are removed as distracted responses, accounting for 3.6% of the full set of 219,687 responses from the 2,443 participants. There was no good way to assess if questions were answered too quickly, since a good answer could be formulated during the presentation of the stimulus. The rater could click as soon as the stimulus stops if they position the mouse adeptly.

We chose to keep all other annotations. In traditional data annotation efforts it is considered essential to quantify annotator consistency and inter-annotator agreement. Assessing annotation reliability for our data set, however, poses a number of challenges. First, even before starting the annotation we expected that some of the stimuli are ambiguous, expressing nuanced or mixed emotions. Our aim was to identify the ambiguous stimuli as those where the annotators did not agree. This aspect of the problem we address complicates the study of inter-rater agreement because disagreement on these stimuli is not a sign of an annotation problem.

The other difficulty for the application of standard reliability measures is that there is very little overlap of raters between clips---most clips were annotated by different groups of raters. This was done by design, to increase the chance that each clip is rated by a set of good raters and minimize the possibility that a the same group of poor raters annotates multiple stimuli.

Despite these challenges we provide reliability analysis in two ways. First we analyze the reliability of individual raters in terms of real-valued annotation of emotion intensity. We investigate the reliability of raters in terms of the correlation between the individual rater's annotation and the “reference” label calculated as the averaged intensity rating of all raters. These calculations are per annotator, with correlations computed for all the stimuli rated by this particular annotator. Our raters show reasonable reliability. Most of them achieve moderate or high correlation with the reference labels. Specifically, when we consider the average correlation score of all 6 emotions, only 20 raters out of 2443 raters (~1% of the raters) show week correlation (correlation < 0.4); 888 raters (~36% of the raters) achieve moderate correlation (0.4 < correlation < 0.7); and the rest of 1535 raters (~63% of the raters) obtain strong or high correlation (correlation > 0.7). There is a stark difference in rater agreement with the reference dependeing on the modality of the stimuli. The highest average correlation of 0.75 with the reference label is on audio-visual clips, dropping to 0.72 on video-only clips. The lowest average correlation of 0.62 is on audio-only clips. We provide the detailed histogram of rater's correlation on each modality and emotion in the paper supplement.

Next we investigate the overall inter-annotator agreement. Here we analyze groups of annotators. The calculations are perfromed for each unique group of raters that annotated at least one clip together, on all clips for which at least two of the annotators of the group provided judgements. We analyze the agreement on the nominal emotion rating on the entire accepted dataset, as well as on the subset for which we find that 80% of the raters agree on the emotion label. We consider the latter subset of emotions to be clear depictions of emotional states. We consider the stimuli where the majority emotion label was given by fewer than 80% of the raters as examples of ambiguous portrayal of emotion as we discuss in later sections so for these agreement between the annotators in not expected. We also analyze the agreement on the scale intensity rating.

We use Krippendorff's alpha as our reliability metric [53,54] because it can handle categorical responses (selected emotion label) as well as ratio responses (where 0 is the lowest value and means absence, like our real-value intensity rating). It can also handle missing responses which is common for our dataset as discussed at the beginning of this section. Our methods of analysis are fully defined in the supplement. Results are summarized in Table 5, listing the average number of clips over which the scores are computed, as well as the average number of raters from the group in these clips, along with the alpha values.

The average alpha reliability score for the full set is 0.42. When assessing the reliability only for the subset of data for which at least 80% of the annotators identified the same emotion, the average alpha reliability score is 0.79. Given this group-wise analysis of inter-annotator

agreement, each clip is characterized by the reliability of annotators that rated the clip. In studies with specific requirements on realiability researchers could select a subset of stimuli rated by raters with high agreement. We provide the detailed analysis of Krippendorff's alpha in the supplement.

In addition to reliability analysis, we measure rater self-consistency based on how raters annotate the duplicated clips. In our study, 9 clips were selected (3 for each modality) for duplication. The consistency of an annotator is defined as the fraction of all repeated clips that were assigned the same label among all repeated clips. The detailed consistency information of every individual rater will be provided in the CREMA-D dataset, such that the database users can decide how to use the data based on their own applications. In general, the consistency of individual raters is high, with overall consistency close to 70%. The expected accuracy of randomly selecting the same emotion of a clip as that given in the first rating is 16%. In addition, we observe that there is less consistency on the low intensity stimuli, and annotators are more likely to change their annotation on those clips. Specifically, for the clips with consistent rating, we notice the higher intensity (mean 65.5, std 23.6), while the intensity for the clips where annotators were not consistent are much lower (mean 54.5, std 24.9). We discuss the consistency of raters on the three modalities in section 5.1.

3.2 Grouping the Responses

After the data were cleaned, we tabulated the responses by sentence and modality. The tabulated responses include the number of times a clip has been recognized as a given emotion, the total number of responses, the mean intensity value for each emotion label, the mean confidence value for neutral responses, and normalized versions of intensity and confidence values. The intensity and confidence values are normalized per rater to span the full range of 0 to 100 matching the scale of the original ratings. Specifically, the mean intensities for audio-only are around 50 for all emotions, except for anger, which is nearly 60. The mean intensities for visual-only is around 60 for all emotions except for happy, which is almost 70. For audio-visual responses, the mean intensity is in between the visual-only and the audio-only intensity except for anger, where the intensity is slightly higher than visual-only.

We define unambiguous clips as those with a group perceived emotion identified by the majority of raters. Ambiguous clips are those with a group perceived emotion with no majority. Some unambiguous clips had rater agreement on the emotion, but the group perceived emotion did not match the intended emotion of the actor despite a director validating the emotion at the time of acquisition. We therefore further split the unambiguous clips into a matching subset, in which the group perceived emotion matches the intended emotion, and a non-matching subset, in which the group perceived emotion differs from the intended emotion.

The resulting three subsets are matching, non-matching and ambiguous. The matching subset corresponds to group recognition and is typically treated as the gold-standard, however, there are a number of uses for clips where the intended emotion does not match the majority rating. We include all three subsets in the final release of the dataset so that

researchers can utilize data that best fits their research objectives in terms of ambiguity and prototypical expressions.

Binomial majority is used to define majority recognition. Unlike traditional majority, which is defined as more than 50% of raters having selected the specific emotion, binomial majority is achieved when a binomial test would reject at the 95% confidence level the null hypothesis that the most commonly chosen label is selected randomly from the six possible labels.

Table 6 shows the number of clips that have between 4 and 12 ratings for the cleaned data. More than 99.9% of the clips have more than 6 rateings. More than 95% of the clips have more than 8 ratings. These values are shaded in the table. We also list the minimum number of votes necessary to select an emotional label using the binomial majority. Compared with the strict majority, we need fewer votes with the binomial majority for the clips with 8 ratings and above. In such case we will put more clips in one of the unambiguous categories. Table 7 gives the proportion of three subsets of matching, non-matching, and ambiguous, for each modality. The large number of annotated clips in CREMA-D ensures that each of the 9 subsets has a sufficient number of clips to be useful. The smallest group, ambiguous audio-visual, has over 590 rated clips.

On the other hand, we will also specify a subset of clips (MAIN set) with 10 or more rating. This MAIN set is expected can be reliably used in future perception studies from a statistical/psychological point of view.

4 The Variability and Ambiguity of Dataset

Based on the perceptual surveys, we first explore the variability and ambiguity of the CREMA-D database.

4.1 Perception and Modality

First of all, our proposed CREMA-D dataset shows its variability in terms of large collection of recording from various sources and separate human ratings on each corresponding source channels.

The multi-modal recordings in terms of audio-only, video-only and audio-visual emotion expression allows for future studies of emotion communications either focusing on any of the single channels or on the more complex problem of multi-modal expression. Unlike many other multi-modal datasets of emotion expression that contain only one overall label reflecting the impression created by the multi-modal stimulus, our collection provides separate human ratings in each modality. The availability of ratings in all three modalities allows us to further investigate the relationship and difference among single- and multi-modal perception.

4.2 Emotion expression with varying intensity

The stimuli in the CREMA-D dataset also demonstrate variability in terms of wide range of emotion expression intensity in all three modalities.

Table 8 summarizes the distribution of different intensity levels for each modality with corresponding recognition rates. As shown in the table, intensities were split into three levels, where mild expressions correspond to the lower quartile of intensity ratings, extreme fall in the upper quartile, and medium is everything in between. In order to better understand the relationship between expression intensity and corresponding recognition rate, we list the corresponding group perception rate in the same table. The group perception rate is the percentage of clips that the majority of the group labels with the same emotion that the actor expressed.

A large proportion of emotion expressions in our datasets are of moderate intensity, in all three modalities. This finding attests to the fact that in general the actors, guided by the directors, in many instances expressed subtle emotional states rather than extreme exaggerated expressions. The visual and audio-visual expressions had the largest proportion of portrayals with extreme intensity, these account for only about a third of all stimuli in a given modality. In contrast, for audio a large fraction of the acquired expressions are in fact perceived as mild, with less than 15% of the stimuli rated to be of extreme intensity.

Most importantly, each level of intensity for each modality contains at least 1,000 clips. Thus, if researchers want to use only a portion of the dataset, for example only the mild, more ambiguous expressions, or only the prototypical, easily recognizable expressions, the subset will still be larger than many of the currently existing corpora of emotion expression. Models built on the former subset may be more appropriate for application on natural conversational data, while those on the latter may be more useful for emotion generation applications.

The intensity of emotion expression is strongly correlated with human recognition rate. The recognition rates increase as intensity increases, in each modality. Moreover, simialr as Busso *et. al* [52] found, the recognition rate at each intensity level is worst for audio-only and best for audio-visual. All of these observations illustrate that human can perceive emotion more correctly with more explicit information, e.g., higher intensity level and more complete channels of expression.

4.3 Mixture of emotion expression

Natural emotion expression is often mixed or ambiguous. Accordingly, human raters are expected to have high agreement for clear and prototypical expressions but for ambiguous expressions, different raters may disagree when asked to pick a single emotion. We tabulate the distribution of raters' responses for each clip as an emotion profile. The emotion profile of a clip shows the mixture of emotions perceived by the raters.

Fig. 3 depicts the average emotion profiles of stimuli in our dataset, showing the distributions of responses when anger, disgust, fear, happy, neutral, and sad are the primary perceived emotion in each modality.

The clarity of emotion expressions varies in terms of distinct expression channels on different emotions. In general, emotion expression is most ambiguous in terms of vocal expression and most clear based on multimodal audio-visual expressions. In terms of

emotions, anger is the clearest emotion. Disgust is its clear secondary perceived emotion, in all three modalities. On the other hand, sad is the most ambiguous emotion with the lowest rate on the primary component and it does not show any clear preference for secondary emotion. We also notice that various modalities show advantages in representing different emotions. For example, facial expression conveys happiness quite unambiguously. About 90% of the raters agree with each other in perception of happy in terms of facial or audio-visual expressions.

Our inclusion of clips with ambiguous emotional profiles is motivated by applications where the emotions expressed may not be as clear. For example, sometimes ambiguous expressions may be more close to natural emotion expression in real life. Different from many emotional datasets which only include recording with clear and prototypical emotion expressions, the partitioning of three subgroups of matching, non-matching, and ambiguous in our dataset can support a wider range of applications in future studies.

5. Human Perception of Emotion Expression

Our analysis so far has demonstrated the wide variety of expressions in our datasets. Here we are particularly interested in the difference and correlation among various modalities in emotion perception. We first discuss how people perceive emotion differently across the three modalities of audio-only, video-only, and audio-visual in section 5.1. Then in section 5.2 we further discuss the interaction of modalities in perception of different emotions.

5.1 Emotion perception in various modalities

In order to better understand how people perceive emotion expressions in different modalities, here we examine the response time of perception, recognition rate and intensity, rater consistency, and individual rater difference, across different modalities.

Fig. 4 compares three histograms, one for each modality, with the distribution of response time for the task of selecting an emotion category for each clip. The difference in the histogram shapes shows that people have different speed of emotion perception across modalities. People need longer time to perceive emotion via vocal expression in general, while they have similar faster response speed for facial and audio-visual expressions on average. There are more samples with very quick response of less than 1 second on audio-visual multimodal compared with the video-only modality. This suggests that involving of vocal expression may help to increase the speed of perception in some cases.

Next, we analyze how individual human raters differ across modalities in terms of 1) the accuracy with which the rater was able to recognize the intended emotions (i.e. the percentage of stimuli correctly recognized as the intended emotion), and 2) levels of intensity or confidence once an emotion is selected by the rater.

We find that there is a significant difference between recognition rates in each of the modalities. On average, audio-only, visual-only, and audio-visual ratings from individual raters matched the intended emotion at 40.9%, 58.2% and 63.6% respectively, with $p < 0.001$ for ANOVA test for difference between the groups. This suggests that more emotion-

related information is portrayed in the visual-only than in the audio-only expressions, and than the combined audio-visual presentation further improves the recognition rate of individual raters.

In addition, there is a significant difference between each emotion. The overall recognition rates, across modalities, from highest to lowest by emotion are neutral, happy, anger, disgust, fear, and sad. Fig. 5 shows the comparison of individual recognition rates between modalities and emotions.

It is evident that different modalities are better suited for expressing particular emotions. Video-only and audio-visual modalities show similar trend but they are markedly different from the audio-only modality. For example, facial expression conveys happy the clearest, while vocal expression conveys anger much better than other emotions.

We also test the effect of modality on mean emotional intensity and show the results in Fig. 6. There is a significant difference between modalities for both emotional intensity and neutral confidence. The mean intensities per modality ranked from highest to lowest are visual-only, audio-visual, and audio-only. In contrast the highest confidence in assigning a neutral label was for audio-visual, then for visual-only. Confidence was lowest for audio-only stimuli.

We also examine the consistency of rater responses for testing clips when a stimulus is repeated. In our database, three stimuli were repeated per modality for each rater. Table 9 shows the consistency of individual raters on three modalities. The overall consistency is quite high, nearly 70%, indicating that the same expressions trigger the same interpretation of emotional content. Moreover, the consistency is highest for audio-visual stimuli, 76%, and lowest for audio-only.

In addition to the analysis of general trends based on average recognition and intensity from individual raters, we were also interested in the variations across individual raters. Fig. 8 shows the boxplot of recognition rates across all human raters, in each modality. The central mark on the boxplot is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually.

Overall, considerable variations are observed across different raters, in all three modalities. For example, in terms of vocal expression, the best recognition rate is higher than 80% for one rater, while the lowest performance is worse than 10% for another rater. Video-only modality shows the smallest difference across all raters, followed by audio-visual one, while audio-only exhibit the largest variations.

The boxplot analysis shows that audio-visual expression more clearly conveys emotional states than facial and vocal expressions across all human raters. However this general trend does not always hold true for individual raters. In order to discount the impression of a potential learning effect by having a fixed order of presentation of audio, then visual, then audio-visual, two tests were run. The first test is to group the individual ratings by the modality of presentation, and check the ordering of recognition. Since there are three

modalities, there are three ordinals, best, middle, and worst. These individual recognition rate orders are tabulated in Table 10, which summarizes the distribution individual rater recognition order by modality. (e.g. An individual whose recognition order from best to worst was Visual, Audio, Audio-visual would increase the count of Best Visual, Middle Audio, Worst Audio-visual.) The audio-visual multimodal expressions were ordered as Best at the highest rate, 58.7%. However, in order to state that there is a learning effect, audio-visual should be best, visual should be middle, and audio should be worst for almost all of the time. Instead, each bin has at leat 5% of the raters. In addition, audio is not always the modality with worst recognition when individual raters are considered: almost 20% of the raters show worse recognition rate on video or multimodal expressions. More evidence against the presence of a learning effect is that the recognition of the audio stimuli was in fact the highest for 5% of the raters. Similarly, 36% of the individual raters recognized the visual-only stimuli with the highest accuracy. These findings suggest that the overall trend that audio-visual stimuli are most accurately recognized while audio-only stimuli are least accurately recognized is not due to learning effects; individual raters do not follow the same trend in their individual performance. Some individuals are best at the clips that come first or second, and some are worst at the clips that come second or last.

To complete our analysis of individual rater recognition rates and possible learning effects due to presentation order, we examine the average recognition rate of all raters per stimulus number (stimuli number corresponds to the order in which raters saw them) and plot the results in Fig. 9. There does not appear to be a direct relationship between recognition rates and presentation order of the stimuli. The recognition rates are comparable for different stimuli for the video-only modality, which was always presented second and audio-visual modality, which was always presented last. Audio stimuli were always presented first, and there we do observe steady improvement on the first 5 responses, however the performance tends to become stable for the remaining 30 responses.

5.2 Interaction of various modalities

Now we turn to investigate the interaction among modalities in the perception of different emotions. We perform the analysis in terms of group perceived ratings.

First, we examine the confusion between intended emotion and the group perceived emotion that has binomial majority across all modalities together. Fig. 10 shows the confusion matrix between intended and group perceived emotion for all unambiguous clips. There is a clear bias of the raters to select neutral. Apart from the intended emotions in the diagonal, neutral was the most often perceived emotion, in all of the five emotion classes. Also more than 95% neutral utterances are correctly recognized as intended with only a handful of exceptions. The most common confusion between actual emotion classes were intended anger perceived as disgust (11%); intended disgust perceived respectively as sad (9%), fear (6%), and anger (5%) with decreasing rate. Intended fear is perceived as sad (9%) or vice versa (8%). Happy is mostly recognized correctly.

Next, in order to tease apart the contribution of audio and video modalities to the perception of intended emotion, we examine the agreement in perception based on individual modality (audio or visual) compared to the audio-visual ratings. In order to do this we count the

number of clips that have the same group perceived label in each combination of modalities. We show these counts in the Venn diagrams of Fig. 11.

For each emotion there is a Venn diagram in Fig. 11. The sum of the numbers inside the each circle is the total count of clips perceived as the emotion of the diagram's title for the specified mode. For example, happy is perceived as the expressed emotion by the audio, visual, and audio-visual modalities in 318, 952, and 1206 clips, respectively. In addition, of the 1235 perceived happy clips, merely 45 were identified in only one modality.

Based on the diagrams of Fig. 11, we first notice that a visual signal is crucial for the accurate perception of a happy expression. Next, neutral and anger both have the majority of clips detected in all modalities (the center number of each diagram), suggesting that much of the information for perception of neutral and anger is redundant across modalities. On the other hand, disgust and fear benefit from the multimodal perception the most, since they have a much smaller proportion of clips detected in all modalities and relatively larger portion of clips detected via only audio-visual modality. Finally, each single modality is important for sad because sad has almost the same number of clips detected by more than one modality as are detected by a single mode.

6. Discussion And Conclusions

CREMA-D is a new data set for multimodal research on emotion. We have described the creation of this data set containing ratings of intended and perceived emotional expression. We used crowd sourcing as a way to collect the perceptual ratings. CREMA-D has over 7 ratings in each modality (audio, visual, and audio-visual), for more than 95% of 7,442 clips that cover 12 sentences spoken with 6 different emotion categories (happy, sad, anger, fear, disgust, and neutral). This makes CREMA-D an excellent resource when considering questions of audio-visual perception of emotions. Group perceived emotion is provided on each clip and we further created three subgroups (matching, non-matching, and ambiguous) by applying the binomial majority test and comparing the intended emotion to the group perceived emotion.

The dataset contains a wide range of emotion expression from ambiguous to prototypical emotion, subtle to extreme expression, in all three. We also provided a detailed perception analysis in terms of the individual ratings, the different modalities, and the group ratings.

The recognition of emotion in each modality was tested in multiple ways. Audio is the most difficult to recognize, visual is in the middle, and audio-visual is the easiest to recognize. This holds both when we considered the overall performance of all individual raters and group perceived emotion recognition. However, there are also some exceptions when we consider specific emotions or raters. For example, audio and visual show similar performance on anger, while visual and audio-visual are not significantly different for happy and sad. On the other hand, in terms of emotion recognition performance on different raters, we also notice that about 20% of the raters had their best recognition rate for audio stimuli.

When looking at consistency of individual emotion perception, the same order of audio, visual, and audio-visual is seen with regards to the proportion of consistent responses. There

also appears to be a relationship between recognition performance and perceived intensity of expression. Clips rated as expressing emotions of extreme intensity were recognized over 25% more accurately than those perceived to be of mild intensity in each modality. Similarly, the order of recognition (audio, visual, and audio-visual) holds in each intensity bin.

In terms of response time, we observed significantly faster response on audio-visual and video modalities, than on audio. Although video and audio-visual show comparable response speed on average, we notice a larger portion of very quick response (<1s) on audio-visual perception. Finally, for the perception within each modality, we do not observe significant difference depending on the order of presentation of the stimuli.

We further investigate the interaction of various modalities. We observe that certain modalities best convey particular emotions. For happy, the group perception of emotion is dominated by the visual channel for over 70% clips, and the rest of the clips are primarily having redundant audio and visual information. For neutral, most of the clips have redundant information (the center number of the neutral diagram in Fig. 11), and there are a small number of clips where either the visual modality dominates (the visual and audio-visual overlap) or the audio modality dominates (the audio and audio-visual overlap). For anger, disgust, fear, and sad, all modalities make some contributions and we do not observe obvious preferred channel of expression. There are cases of redundancy, visual dominance, audio dominance and complementarity shown to different degrees. A deeper examination of the data set, where individual clips are examined across modalities, could yield further insight into these questions. This is reserved for the future, when the analysis can be carried out as part of more emotion specific hypothesis driven studies.

Other highly relevant potential studies that can be carried out on the newly developed dataset can quantify the effects of gender, age, race, etc. for the production and perception of emotion expression. We performed some pilot studies and observed that women are better than man both in clearly expressing emotions and in accurately recognizing the intended emotion. This agrees with an earlier study examining gender differences in emotion detection using the voice channel [4]. More detailed studies can be performed in the future.

In conclusion, we presented a sampling of the analysis that can be performed on CREMA-D, and found differences related to emotional expression in each modality. These differences allow a user of the dataset to separate the data as appropriate. Labels include both intended and group perceived emotions as well as group perceived emotional intensities. The variety of labels and modalities is intended to make this dataset accessible and useful for a variety of purposes. This analysis is by no means comprehensive and has been carried out with the aim of demonstrating the scope of the dataset and the plethora of analyses that can be based on it. We plan to release the database soon to the academic community for research purpose. The researchers are free to apply the best methods of their choice to the dataset based on the needs of the study.

7. Acknowledgment

The authors wish to thank the directors, Amy Dugas Brown and David M. O'Connor, as well as Brian Chaffinch, who assisted with the recording sessions and video editing. This work was supported in part by the following grants: NIH R01-MH060722 and NIH R01 MH084856.

Biography

Howei Cao earned her B.E. degree from Shenzhen University, China, in 2004, the M.S. degree from University of Surrey, U.K., in 2005, and the Ph.D. degree in 2011 from The Chinese University of Hong Kong, Hong Kong. She is currently a Postdoctoral Fellow at the University of Pennsylvania. Her research interests are in speech and language processing, include multi-lingual and cross-lingual speech recognition, emotion and affect analysis and recognition.

David G. Cooper earned a B.S. in Cognitive Science from Carnegie Mellon in 2000, and M.S. and Ph.D. in Computer Science from the University of Massachusetts Amherst in 2009 and 2011 respectively. He worked at Lockheed Martin Advanced Technology Labs from 2000-2006. He held a position as postdoctoral researcher in the Section of Biomedical Image Analysis in the Radiology department of the University of Pennsylvania from 2011 to 2013. He is currently a lecturer at Ursinus College. His research interests include emotional and cognitive models of human interaction, visual and auditory emotion recognition, sensor integration for computer awareness, and biologically inspired computation.

Michael K. Keutmann earned a B.A. in Biology (2005) and an M.S.Ed. in Counseling and Psychological Services (2009) from the University of Pennsylvania. He is a former Research Coordinator in the Brain Behavior Laboratory at the University of Pennsylvania and is currently a graduate student in Clinical Psychology at the University of Illinois at Chicago.

Ruben C. Gur received his B.A. in Psychology and Philosophy from the Hebrew University of Jerusalem, Israel, in 1970 and his M.A. and Ph.D. in Psychology (Clinical) from Michigan State University in 1971 and 1973, respectively. He did Postdoctoral training with E.R. Hilgard at Stanford University and began at the University of Pennsylvania as Assistant Professor in 1974. He is currently a Professor of Psychology in Psychiatry and the Director of the Brain Behavior Laboratory at the University of Pennsylvania. His research has been in the study of brain and behavior in healthy people and patients with brain disorders, with a special emphasis on exploiting neuroimaging as experimental probes. His work has documented sex differences, aging effects, and abnormalities in regional brain function associated with schizophrenia, affective disorders, stroke, epilepsy, movement disorders and dementia. His work has been supported by grants from the NSF, NIH, NIMH, NIA, NINDS, NSBRI, private foundations (Spencer, MacArthur, EJLB) and industry (Pfizer, AstraZeneca).

Ani Nenkova obtained her BS degree in computer science at Sofia University in 2000, and MS and PhD degrees from Columbia University in 2004 and 2006 respectively. Prior to joining Penn she was a postdoctoral fellow at Stanford University. Her main areas of research are summarisation, text quality and affffect recognition. Ani and her collaborators were recipients of the best student paper award at SIGDial in 2010 and best paper award at

EMNLP in 2012. Ani was a member of the editorial board of Computational Linguistics (2009--2011) and has served as an area chair/senior program committee member for ACL, NAACL, AAAI and IJCAI.

Ragini Verma earned her M.S. in Mathematics and Computer Applications followed by a Ph.D. in computer vision and mathematics, from IIT Delhi (India). She did two years of postdoc at INRIA, Rhone-Alpes, with the MOVI project (currently LEARS and PERCEPTION). She then did two years of post doc in medical imaging at SBIA, prior to taking up her current position. She is currently an Associate Professor of Radiology in Section of Biomedical Image Analysis, Department of Radiology at the University of Pennsylvania. Ragini's research interests span the area of diffusion tensor imaging, multi-modality statistics and facial expression analysis. She is actively involved in several clinical studies in schizophrenia, aging, tumors and multiple sclerosis as well as projects in animal imaging. Ragini works in the broad area of multi-parametric image analysis which aims at integrating several channels of information (MRI, genetic and clinical scores) to solve a clinical or biological problems.

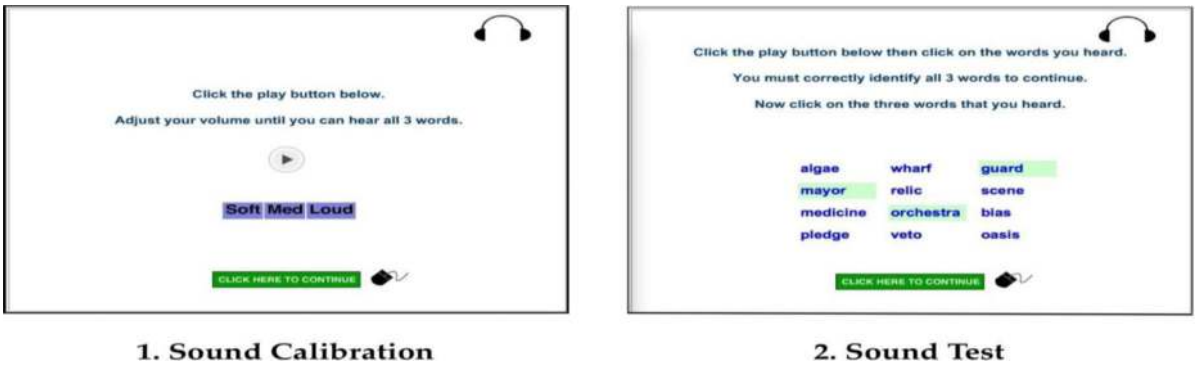
References

1. Ekman P. Facial expressions of emotion: New findings, new questions. Psychological Science. 1992; 3(1):34–38.
2. Ekman P. An argument for basic emotions. Cognition & Emotion. 1992; 6(3-4):169–200.
3. Beaupre MG, Hess U. Cross-cultural emotion recognition among canadian ethnic groups. Journal of Cross-Cultural Psychology. 2005; 36(3):355–370.
4. Scherer K, Banse R, Wallbott H. Emotion inferences from vocal expression correlate across languages and cultures. Journal of Cross-cultural psychology. 2001; 32(1):76–92.
5. Sauter DA, Eisner F, Ekman P, Scott SK. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. Proceedings of the National Academy of Sciences. 2010; 107(6):2408–2412.
6. Cummings KE, Clements MA. Analysis of the glottal excitation of emotionally styled and stressed speech. The Journal of the Acoustical Society of America. 1995; 98(1):88–98. [PubMed: 7608410]
7. Murray I, Arnott J. Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. The Journal of the Acoustical Society of America. 1993; 93
8. Protopapas A, Lieberman P. Fundamental frequency of phonation and perceived emotional stress. The Journal of the Acoustical Society of America. 1997; 101(4):2267–2277. [PubMed: 9104028]
9. Hu H, Xu M-X, Wu W. Gmm supervector based svm with spectral features for speech emotion recognition. Proc. ICASSP. 2007; 4
10. Rong J, Li G, Chen Y-PP. Acoustic feature selection for automatic emotion recognition from speech. Information processing & management. 2009; 45(3):315–328.
11. Neiberg D, Elenius K, Laskowski K. Emotion recognition in spontaneous speech using gmms. Proc. INTERSPEECH. 2006
12. Bachorowski J-A, Owren MJ. Vocal expressions of emotion. Handbook of emotions. 2008; 3:196–210.
13. Matsumoto D, Keltner D, Shiota MN, O'Sullivan M, Frank M. Facial expressions of emotion. Handbook of emotions. 2008; 3:211–234.
14. De Gelder B, Vroomen J. The perception of emotions by ear and by eye. Cognition & Emotion. 2000; 14(3):289–311.
15. Provost EM, Zhu I, Narayanan S. Using emotional noise to uncloud audio-visual emotion perceptual evaluation. Proc. ICME. 2013

16. Kanade T, Cohn JF, Tian Y. Comprehensive database for facial expression analysis. Proc. FG 2000. 2000:46–53.
17. Pantic M, Valstar M, Rademaker R, Maat L. Web-based database for facial expression analysis. Proc. ICME. 2005
18. O Toole AJ, Harms J, Snow SL, Hurst DR, Pappas MR, Ayyad JH, Abdi H. A video database of moving faces and people. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2005; 27(5):812–816. [PubMed: 15875802]
19. Sneddon I, McRorie M, McKeown G, Hanratty J. The belfast induced natural emotion database. IEEE Transactions on Affective Computing. 2012; 3(1):32–41.
20. Banse R, Scherer KR, et al. Acoustic profiles in vocal emotion expression. Journal of personality and social psychology. 1996; 70:614–636. [PubMed: 8851745]
21. Batliner A, Steidl S, Hacker C, Nöth E. Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. User Modeling and User-Adapted Interaction. 2008; 18(1-2):175–206.
22. Burger S, MacLaren V, Yu H. The ISL meeting corpus: The impact of meeting type on speech style. Proc. ICSLP 2002. 2002; 2:301–304.
23. Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. A database of German emotional speech. Proc. Interspeech 2005. 2005
24. Steidl, S. Automatic Classification of Emotion Related User States in Spontaneous Children's Speech. Logos-Verlag: 2009.
25. Douglas-Cowie E, Campbell N, Cowie R, Roach P. Emotional speech: Towards a new generation of databases. Speech communication. 2003; 40(1):33–60.
26. Klasen M, Kenworthy CA, Mathiak KA, Kircher TT, Mathiak K. Supramodal representation of emotions. The Journal of Neuroscience. 2011; 31(38):13635–13643. [PubMed: 21940454]
27. Klasen M, Chen Y-H, Mathiak K. Multisensory emotions: perception, combination and underlying neural processes. Reviews in the neurosciences. 2012; 23:381–392. [PubMed: 23089604]
28. Muller VI, Habel U, Derntl B, Schneider F, Zilles K, Turetsky BI, Eickhoff SB. Incongruence effects in crossmodal emotional integration. NeuroImage. 2011; 54(3):2257–2266. [PubMed: 20974266]
29. Muller VI, Cieslik EC, Turetsky BI, Eickhoff SB. Crossmodal interactions in audiovisual emotion processing. NeuroImage. 2012; 60(1):553–561. [PubMed: 22182770]
30. Ekman P. The argument and evidence about universals in facial expressions of emotion. Handbook of social psychophysiology. 1989; 58:342–353.
31. Bänziger T, Scherer KR. Introducing the geneva multimodal emotion portrayal (GEMEP) corpus. Blueprint for affective computing: A sourcebook. 2010:271–294.
32. De Silva L, Miyasato T, Nakatsu R. Facial emotion recognition using multi-modal information. Proc. ICICS 1997. 1997; 1:397–401.
33. Collignon O, Girard S, Gosselin F, Roy S, Saint-Amour D, Lassonde M, Lepore F, et al. Audio-visual integration of emotion expression. Brain research. 2008; 1242
34. Mower E, Mataric MJ, Narayanan S. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. Trans. Multi. Aug. 2009 11(5):843–855.
35. López J, Cearreta I, Fajardo I, Garay N. Validating a multilingual and multimodal affective database. Usability and Internationalization. 2007; 4560:422–431.
36. Grimm M, Kroschel K, Narayanan S. The vera am mittag german audio-visual emotional speech database. Proc. ICME 2008. 2008:865–868.
37. Busso C, Bulut M, Lee C-C, Kazemzadeh A, Mower E, Kim S, Chang JN, Lee S, Narayanan SS. Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation. 2008; 42(4):335–359.
38. Chen, L. Ph.D. thesis. 2000. Joint processing of audio-visual information for the recognition of emotional expressions in human-computer interaction.
39. 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. The humaine database: Addressing the collection

and annotation of naturalistic and induced emotional data. Affective Computing and Intelligent Interaction. 2007; 4738:488–500.

40. Ringeval F, Sonderegger A, Sauer J, Lalanne D. Introducing the recola multimodal corpus of remote collaborative and affective interactions. Proc. FG 2013. 2013:1–8.
41. You, M.; Chen, C.; Bu, J. Chad: A chinese affective database. In: Tao, J.; Tan, T.; Picard, RW., editors. Affective Computing and Intelligent Interaction, ser. Lecture Notes in Computer Science. Vol. 3784. Springer; Berlin Heidelberg: 2005. p. 542-549.
42. Soleymani M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Transactions on Affective Computing. 2012; 3(1):42–55.
43. Cowie R, Douglas-Cowie E, Cox C. Beyond emotion archetypes: Databases for emotion modelling using neural networks. Neural networks. 2005; 18(4):371–388. [PubMed: 15961273]
44. Russ JB, Gur RC, Bilker WB. Validation of affective and neutral sentence content for prosodic testing. Behav Res Methods. 2008; 40(4):935–939. [PubMed: 19001384]
45. Wilkinson, GS.; Robertson, G. Psychological Assessment Resources. Lutz: 2006. Wide range achievement test (wrat4).
46. Buchholz S, Latorre J. Crowdsourcing preference tests, and how to detect cheating. Proc. INTERSPEECH 2011. 2011
47. Difallah D, Demartini G, Cudre-Mauroux P. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. Proc. of the First International Workshop on Crowdsourcing Web Search. 2012
48. Eickhoff C, de Vries A. Increasing cheat robustness of crowdsourcing tasks. Information Retrieval. 2012:1–17.
49. Hirth M, Hoßfeld T, Tran-Gia P. Cheat-detection mechanisms for crowdsourcing. University of Wurzburg, Tech. Rep. 2010; 474
50. Hirth M, Hossfeld T, Tran-Gia P. Cost-optimal validation mechanisms and cheat-detection for crowdsourcing platforms. Proc. IMIS 2011. 2011:316–321.
51. Zhu D, Carterette B. An analysis of assessor behavior in crowdsourced preference judgments. SIGIR Workshop on Crowdsourcing for Search Evaluation. 2010:21–26.
52. Busso, Carlos; Deng, Zhigang; Yildirim, Serdar; Bulut, Murtaza; Lee, Chul Min; Kazemzadeh, Abe; Lee, Sungbok; Neumann, Ulrich; Narayanan, Shrikanth. Analysis of emotion recognition using facial expressions, speech and multi-modal information. Proc. ICMI 2004. 2004:205–211.
53. Krippendorff, Klaus. Estimating the Reliability, Systematic Error and Random Error of Interval Data. Educational and Psychological Measurement. 1970; 30:61–70.
54. Krippendorff, Klaus. Reliability in content analysis. Human Communication Research. 2004; 30(3):411–433.



1. Sound Calibration 2. Sound Test

Fig 1.

Left: Sound Calibration, Soft, Med, and Loud light up when the corresponding sound is played. Right: Sound Check, the 3 target words have been selected and are shown in green.

Fig. 2.

The ratings task involves 3 steps. Left: the visual only task is shown for step 1, viewing. Center: Step 2, selecting emotion category, is shown for the visual only practice task. Right: Step 3, selecting the emotion level is shown for the audio only practice task after anger was selected.

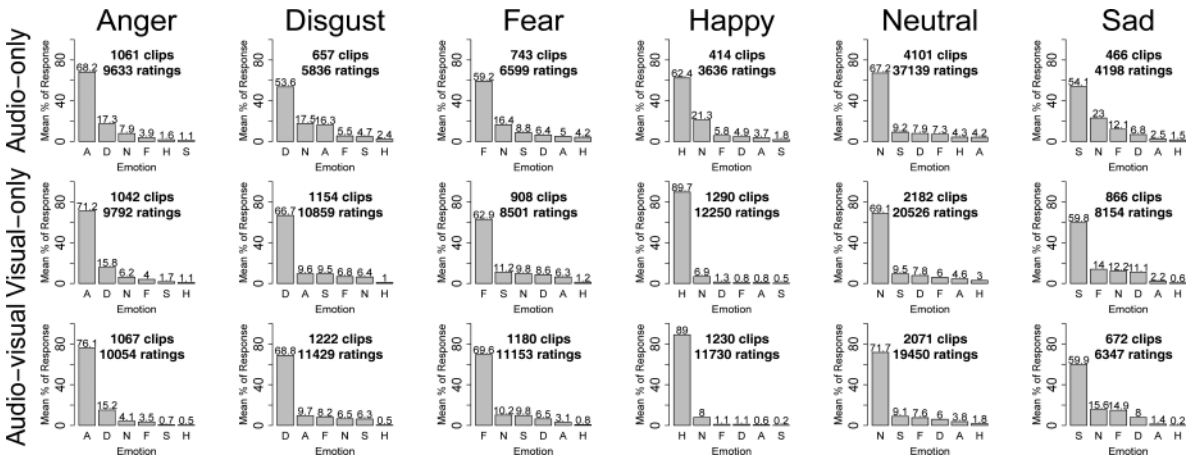


Fig. 3.

The distributions of response when anger, disgust, fear, happy, neutral, and sad are the primary perceived emotion in three modalities of audio-only, visual-only, and the audio-visual multimodal The percentage of emotion is shown per emotion in order of ranking, from most number of clips to least number of clips. A-anger, D-disgust, F-fear, H-happy, N-neutral, S-sad.

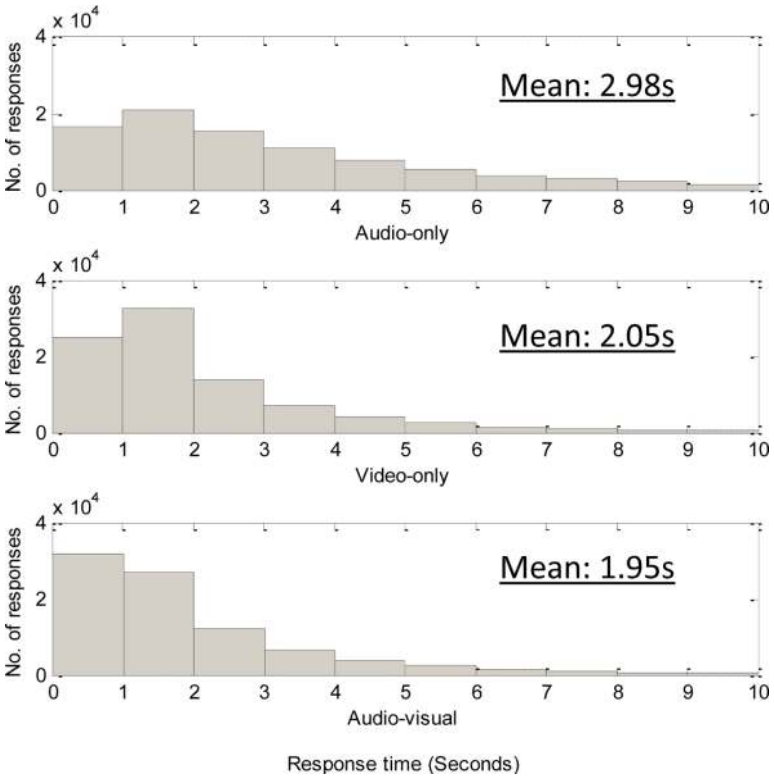


Fig. 4.
The distribution of response time for selecting an emotion category for each modality

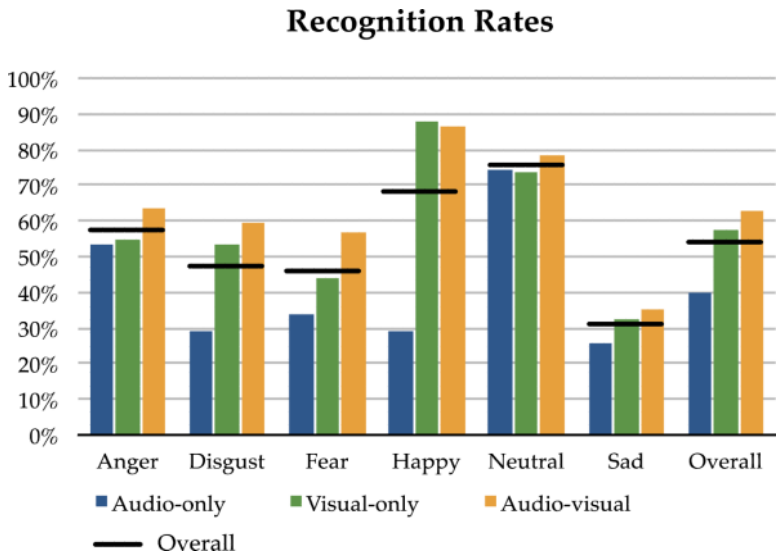


Fig. 5.

There are significant differences in recognition rates between each emotion, and between each modality. When looking at a particular emotion, all differences between modality are significant ($p < 0.05$) except for Face vs. Audio-visual when either happy or sad is the intended emotion, and Voice vs. Face when either anger or neutral is the intended emotion.

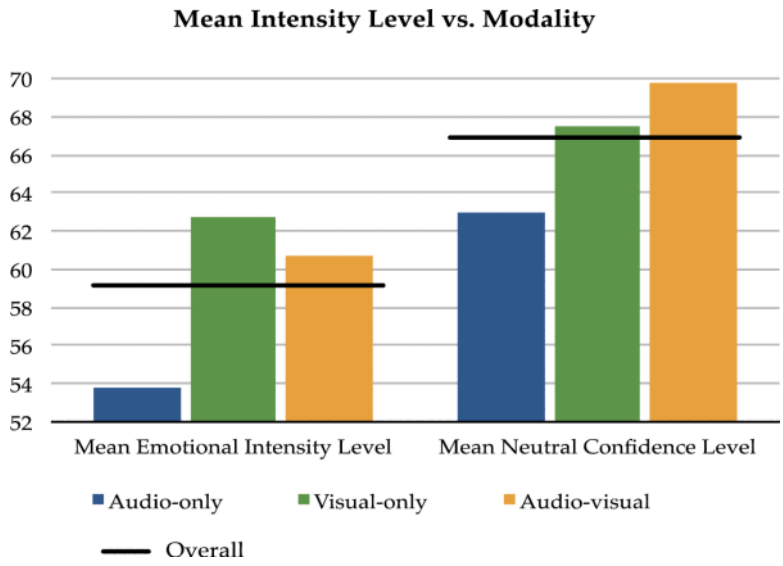


Fig. 6.

Mean emotion intensity (left) and mean neutral confidence (right) are significantly different per modality ($p < 0.05$).

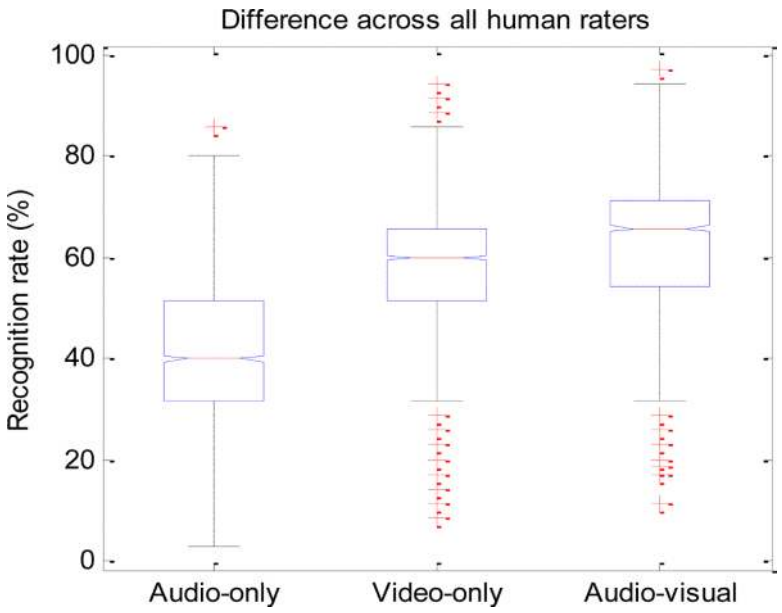


Fig. 8.

Boxplot of recognition rates of all human raters, in audio-only, video-only, and audio-visual respectively.

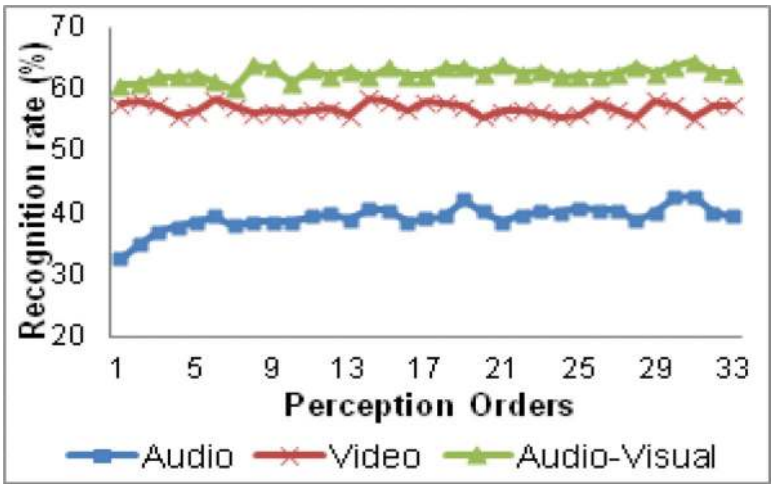


Fig. 9.

The recognition rate over time for audio-only, video-only, and audio-visual respectively. The first five clips of audio appear to have a slight learning effect, but after that no effect is present, and this initial effect does not explain the

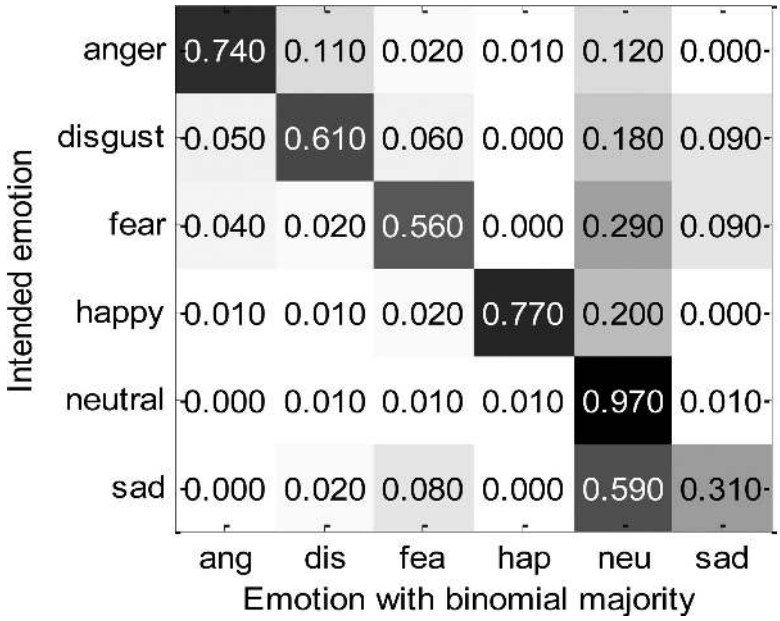


Fig. 10.

The confusion matrix between intended emotion and group perceived emotion indicates that the group perception matches the intended emotion more than other emotions except for sad, which tends to be recognized as neutral.

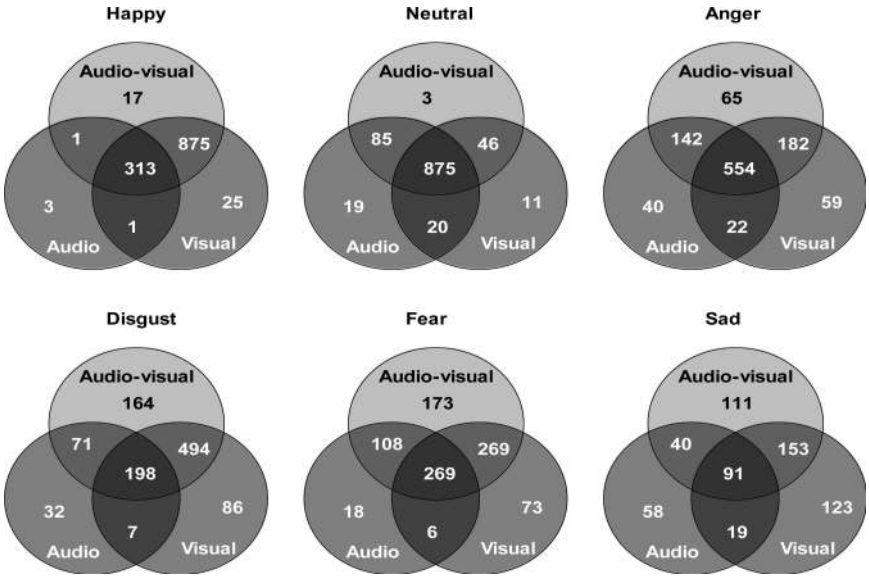


Fig. 11.

Examination of overlapping group perceived emotional responses: The center number of each diagram is the number of clips that are recognized in all conditions. The outer numbers indicate the number of clips that are only recognized by one modality. If a diagram has a high outer top number, with lower combined outer bottom numbers (as seen in the fear Venn diagram with 173 vs. 18 + 73), this suggests that the bimodal perception is advantagious over single mode perception.

TABLE 1

Comparison To Prior Work Using Data Set Criteria

Data Set	Greater than 50 people recorded (# people)	Greater than 5,000 Clips (# of clips)	At least 6 emotion categories (# categories)	At least 8 ratersper clip for over 95% of clips (# raters)	All 3 rating modalities (which modalities)
CREMA-D (this work)	✓ **(91)**	✓ **(7,442)**	✓ (6)	✓ **(4-12,mean 9.8)**	✓ **(audio, visual, audio-visual)**
GEMEP [31]	x(10)	x (1,260)	✓ **(18)**	✓ **(Audio 23, Visual 25, AV 23)**	✓ **(audio, visual, audio-visual)**
De Silva Multimodal [32]	x(2)	x(72)	✓ (6)	✓ **(18)**	✓ **(audio, visual, audio-visual)**
Mower Provost [15]	x(1)	x(72)	x(5)	✓ **(117)**	✓ **(audio, visual, audio-visual, AV mismatch)**
AV Integration [33]	x(6)	x(60)	x(2)	✓ **(8)**	✓ **(audio, visual, audio-visual, AV mismatch)**
AV Synthetic Character [34]	x (1 **female voice,**1 **animated face)**	x (210)	x(4)	x (3 **to** 4 **for AV,** 6 **to** 7 **for A or V)**	✓ **(audio, visual, audio-visual)**
RekEmozio [35]	x(17)	x (2,720)	x(0)	x (3 **to** 4)	x **(audio for oral, visual for faces)**
Vera Am Mittag German Audio- Visual Database [36]	✓ **(104)**	x (1,421)	x (7, **for faces)**	✓**(Audio :** 6 **or 17 Face: 8-34,mean 14)**	x **(audio, visual)**
IEMOCAP [37]	x(10)	✓ **(10,039)**	✓ **(9)**	x(3)	x **(audio-visual)**
Chen Bimodal [38]	✓ **(100)**	✓ **(9,900)**	✓ **(11)**	x **(None)**	x **(audio-visual)**
HUMAINE [39]	x (≤48, **unspecified)**	x(48)	✓ **(48)**	x(6)	x **(audio-visual)**
RECOLA [40]	x(46)	x(46)	x(2)	x(6)	x **(audio-visual)**
CHAD [41]	x(42)	✓ (6,**228)**	✓ **(7)**	✓ **(120)**	x **(audio)**
MAHNOB-HCI [42]	x(27)	x (1,296)	✓ **(9)**	x(1)	x **(self-report)**

✓ *indicates the criterion is met.* x *-indicates criterion is not met. Each highlighted cell indicates that the criterion for the column was met by the data set.*

TABLE 2

Actors' Age Distribution

Age	# actors
20-29 YRS	34
30-39 YRS	23
40-49 YRS	16
50-59 YRS	12
60-69 YRS	5
OVER 70 YRS	1

TABLE 3

Actors' Race/Ethnicity Breakdown

Ethnicity / Race	Not Hispanic	Hispanic	Total
Caucasian	53	8	61
African American	21	1	22
Asian	7	0	7
Unspecified	0	1	1
Total	81	10	91

TABLE 4

Race/Ethnicity Distribution

Race / Ethnicity	Raters	Actors
Caucasian	73.60%	58.24%
Hispanic	10.80%	10.99%
African American	8.10%	23.08%
Asian	4.50%	7.69%
Other/No Answer	3.00%	0.00%

TABLE 5

Krippendorff's alpha statistics on different groups

Rating	Average Alpha (min.,max.)	Average Num Clips (min., max.)	Average Num Raters (min., max.)
Emotion label	0.42 [0.25, 0.57]	134.3 [105,177]	9.5 [4, 12]
Emotion intensity	0.47 [0.18, 0.71]	134.3 [105,177]	9.5 [4, 12]
Emotion label (agreement > 0.8)	0.79 [0.58, 0.96]	42.9 [16, 78]	9.6 [5, 12]

Average, min and max reliability score for emotion class and intensity for the entire dataset, as well as for emotion class on clips unambiguous clips.

TABLE 6

Ratings Distribution

# Ratings	4	5	6	7	8	9	10	11	12
# Clips	2	10	71	460	2,046	6,951	11,296	1,471	19
Min. Majority Votes	3	3	4	4	4	4	5	5	5

The highlighted area shows over 95% of the clips with 8 or more ratings. The number of votes needed is fewer for binomial majority thanstrict majority when there are 8 or more ratings.

TABLE 7

Proportion of the three subsets

	Three subsets		
	Matching	**Non- matching**	**Ambiguous**
Audio-only	41%	46%	13%
Video-only	64%	25%	11%
Audio-visual	72%	21%	8%

TABLE 8

Emotion Expression at Different Intensities

		Mild (≤48)	Medium	Extreme (>65)
Audio	Rate	34.82%	44.35%	68.01%
	Count	2,685	3,711	1,016
Visual	Rate	55.31%	64%	80.38%
	Count	1,159	3,653	2,600
AV	Rate	61.04%	69.96%	87.61%
	Count	1,422	3,545	2,445

Recognition rate (the % of clips labeled as the intended emotion) increases with intensity for all modalities, and with modality for all intensities.

TABLE 9

Mean Percent Consistency vs. Modality

	Audio- only	Video- only	Audio-visual	Overall
Individual rater consistency	62.6%	69.1%	76.4%	69.4%

TABLE 10

Recognition Rate Ordering by Modality

		Best	Middle	Worst
Audio	Rate	5.2%	13.7%	81.1%
	Count	126	335	1,981
Visual	Rate	36.1%	53.4%	10.5%
	Count	882	1304	257
Audio-visual	Rate	58.7%	32.9%	8.4%
	Count	1,435	804	205

Each count bin represents the number of raters whose recognition rate order was best, middle, or worst for the modalities of audio, visual and audio-visual. The rate is the percentage of raters for in the column or row.