Recording audio-visual emotional databases from actors: a closer look

Carlos Busso and Shrikanth S. Narayanan

Signal Analysis and Interpretation Laboratory (SAIL)
USC Viterbi School of Engineering
University of Southern California

(This research was supported in part by funds from the NSF, and the Department of the Army)

Outline

- Introduction
- Guidelines to record databases from actors
- The USC IEMOCAP corpus
- Conclusions

Introduction

Motivation

- Emotions are crucial for understanding and modeling human behavior
- Availability of appropriate emotional databases is a major limitation for scientific research and technology development
  - Genuine realizations
  - Integrated information from relevant modalities
  - Models that generalize across domains/applications
- Acting and actors have played a key role in the study of emotions
- Current techniques to record databases from actors have limitations
  - Use of naïve or inexperienced subjects
  - Lack of contextualization
  - Emotional descriptors ("*read this sentence portraying anger*")
  - Unfamiliar tasks to the actors

Introduction

A variety of sources for spontaneously elicited material

- Natural emotional corpora
  - Broadcasted television programs (VAN [Grimm, 2007], EmoTV [Abrilian, 2005], Belfast [Douglas-Cowie, 2003])
  - Recording in Situ (Lost luggage [Scherer, 1997])
  - Recalling emotions ([Amir, 2000])
  - Wizard of Oz (SmartKom [Schiel, 2002])
  - Games (EmoTaboo [Zara, 2007])
  - Carefully design human-machine interaction (SAL [Cowie, 2005])
- Core limitations
  - Ethical issues (i.e., inducing emotions)
  - Copyright problems
  - Constrained to specific domains
  - Lack of control over the microphones and camera locations
  - Noisy visual and/or acoustic background
  - Incomplete information from modalities

We consider the role of acting as a viable research methodology for studying human emotions

Introduction

Can specific acting methods be used to mitigate the limitations of recording emotional data from actors?

- Acting provide opportunities to tackle the problem in a systematic and controlled fashion [Enos, 2006]
- How?
  - Using better elicitation techniques
  - Make use of acting techniques
  - Make connection with real-life scenarios
  - Create suitable social settings in the recording

- In this talk we present
  - Guidelines for designing new emotional corpora
  - Our new IEMOCAP database

Outline

- ✓ Introduction
- ➢ Guidelines to record databases from actors
- ➢ The USC IEMOCAP corpus
- ➢ Conclusions

Guidelines to record databases from actors

Contextualization and social setting

- Discourse context is important for expressing emotion [Cauldwell, 2000]
- Isolated sentences or short dialogs are not appropriate for eliciting emotions
- Read speech is different from spontaneous speech

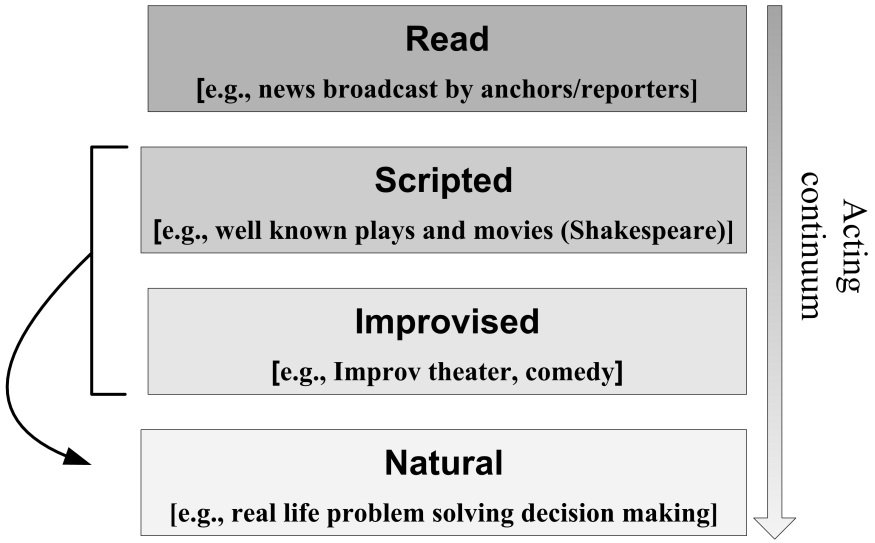
Suggested guidelines

- Database should contain natural dialogs
- Dialogs should be long enough to contextualize the flow of emotions [Douglas-Cowie, 2003]
- Semantic content should be congruent with the intended emotions
- Record interaction between two or more actors rather than monologues

Guidelines to record databases from actors

Acting styles

- Adopt acting techniques that are well-known to the actors
  - Reading sentences portraying different emotions may not be adequate
- Acting continuum
  - From fully predetermined to fully undetermined
- The use of specific acting methods and styles can be used to control aspects of the recording
  - Balance the tradeoff between controllability and naturalness
- Two appealing acting genres are
  - Scripted plays
  - Improvisation



Guidelines to record databases from actors

Trained actors

- Exaggeration or caricature of emotions can be avoided by recording skilled actors
- As the subjects display facial expression that are closer to genuine emotions they may start feeling the emotions [Ekman, 1993]

Additional guidelines

- Use audition sessions to select the actors
- Rehearsing the material in advance
  - Get familiar with the material
  - Get familiar with their colleagues
- Use an experienced professional to supervise the audition, rehearsing and recording sessions

Guidelines to record databases from actors

Emotional descriptors

- Defining emotional descriptions is a key aspect
- Categorical descriptions (happiness, anger)
  - Which emotions labels to target (material)
  - Extensive list, poor inter-evaluator agreement
  - Limited list, poor emotional description of the corpus
- Continuous descriptions (valence, activation)
  - More general
  - Useful for emotion expression variability
- Emotional descriptors should be assigned based on perceptual evaluations
  - As many subjects as possible
  - Pilot tests are highly suggested
  - Warning: differences between expression and perception
  - Evaluators should judge the emotional content based on sequential development of the dialogs
  - All available modalities should be presented

Outline

- ✓ Introduction
- ✓ Guidelines to record databases from actors
- ➢ The USC IEMOCAP corpus
- ➢ Conclusions

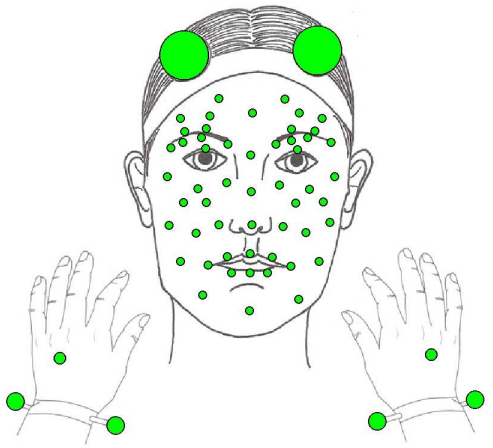
The USC IEMOCAP corpus

Requirements

- The database must contain genuine realizations of emotions
- It should contain natural dialogues, in which the emotions are naturally elicited
- Many experienced actors should be recorded
- The emotional and linguistic content should be as controlled as possible
- The database should have detailed acoustic and visual information
- The emotional labels should be assigned based on human subjective evaluations

Design of the corpus

- Direct and detailed motion capture information
  - Face (53), head motion (2), hand gestures (6)
- Ten actors were recorded in dyadic sessions
- Scripts and improvisation of hypothetical scenarios
- Happiness, anger, sadness, frustration and neutral
- Approximately 12 hours of audiovisual data



The USC IEMOCAP corpus

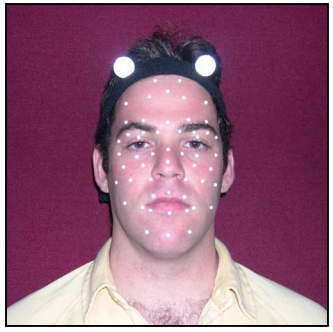
Material selection

- Scripted sessions: use of plays
  - Three 10-minute scripts were selected
  - The subjects were asked to memorize and rehearse
  - The emotions are expressed within a suitable context
- Spontaneous sessions: Improvisation based on hypothetical scenarios
  - The topics were selected following the guidelines provided by Scherer et al. [Scherer,1986]
  - Common situations such us loss of a friend and separation
  - For example, one subject is telling his/her friend that he/she is getting married

The USC IEMOCAP corpus

Actor selection

- 7 actors and 3 senior students (USC)
- Selected based on audition sessions
- Rehearsal practices were supervised by an experienced professional



Recording setting

- Facial markers were placed according to the MPEG-4 standard
- VICON motion capture system with 8 cameras
- Audio was recorded with 2 high quality shotgun microphones (Schoeps CMIT5U)
- Two high-resolution digital cameras recorded semi frontal views of the actors
- The recordings were synchronized by using a clapboard

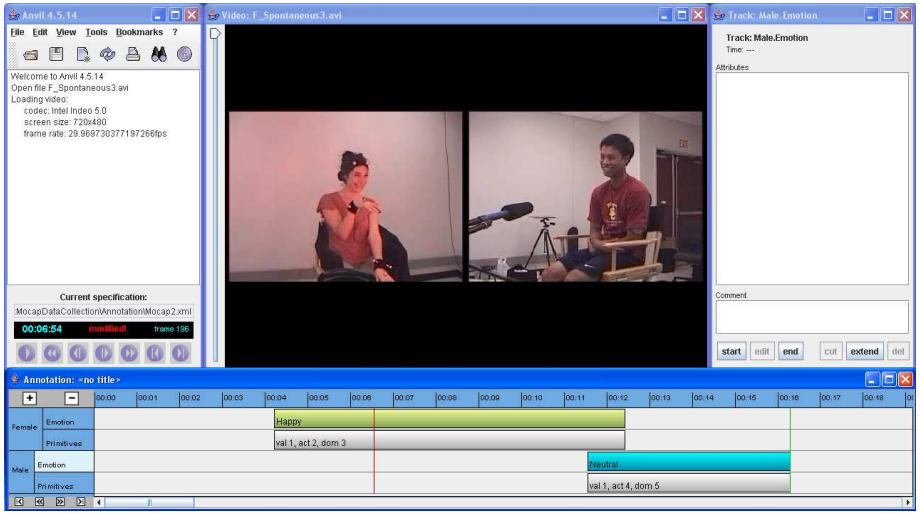
The USC IEMOCAP corpus

Segmentation and transcription of the data

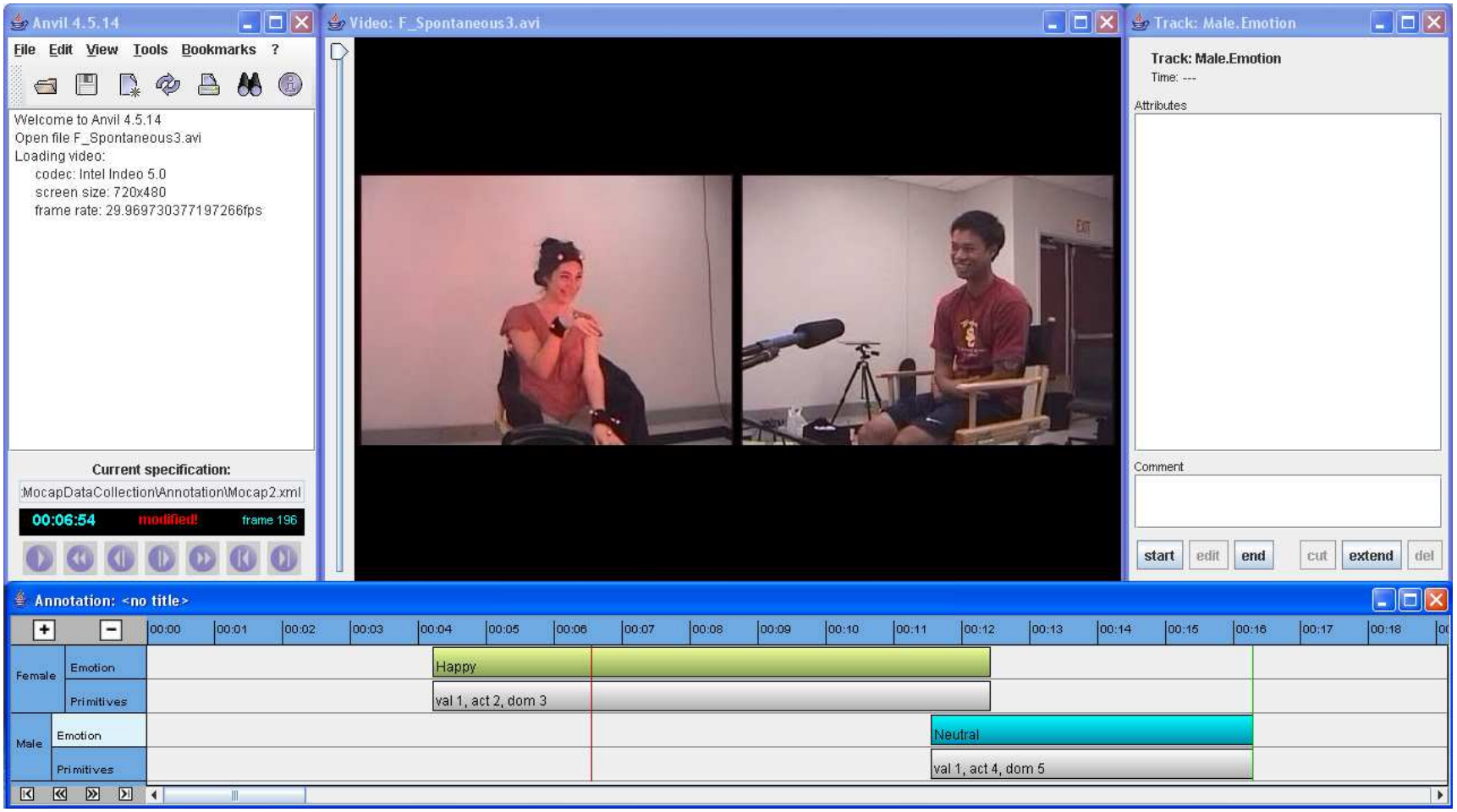
- The dialogs were manually segmented at the dialog turn level
- Multi-sentence utterances were segmented as single turns
- The corpus contained 10039 turns
- Transcription of the audio dialogs were obtained from Ubiqus

Emotional annotation of the data

- Emotional labels were assigned based on subjective emotional evaluations
- The ANnotation of VIdeo and spoken Language tool (ANVIL) was used
- The evaluators sequentially assessed the turns, after watching the videos



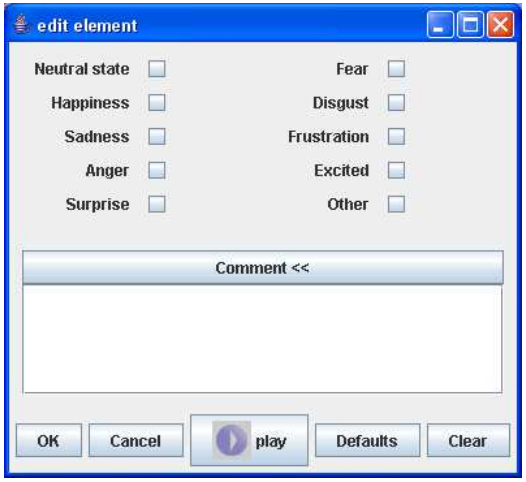
The USC IEMOCAP corpus



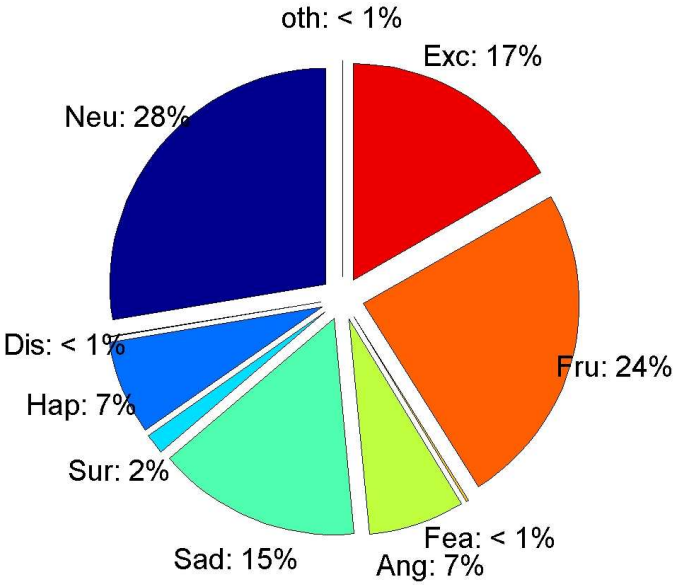
The USC IEMOCAP corpus

Categorical descriptors

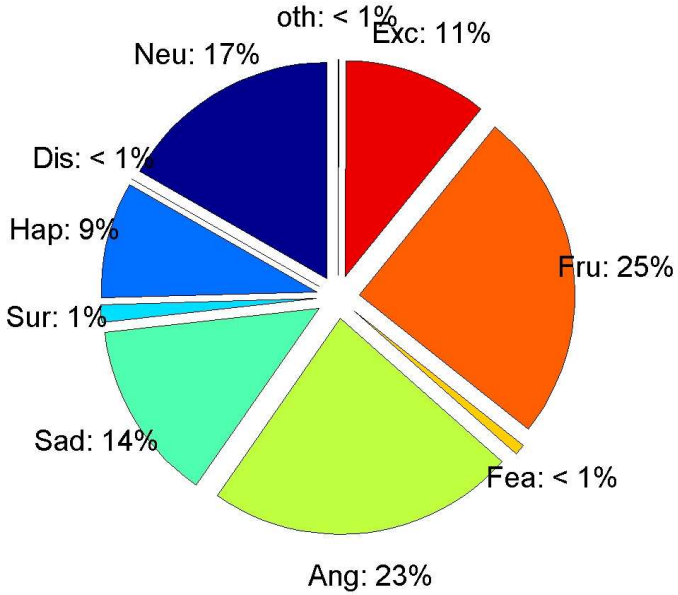
- Three different evaluators assessed each utterance
- The evaluators could select more than one label
- Majority voting was used
- 74.6% of the turns were assigned a label



Spontaneous sessions



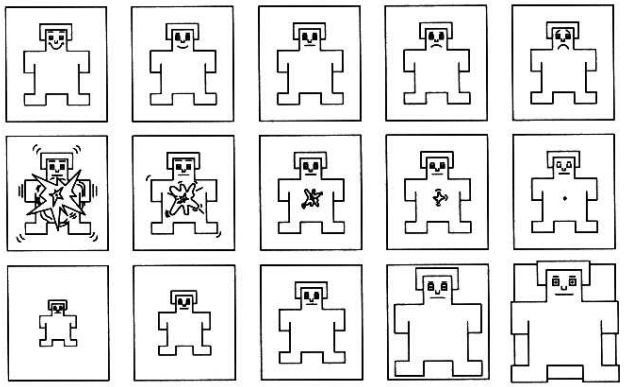
Scripted sessions

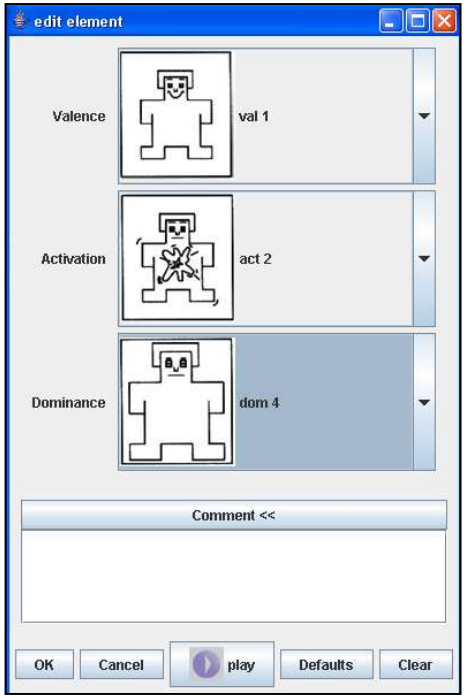


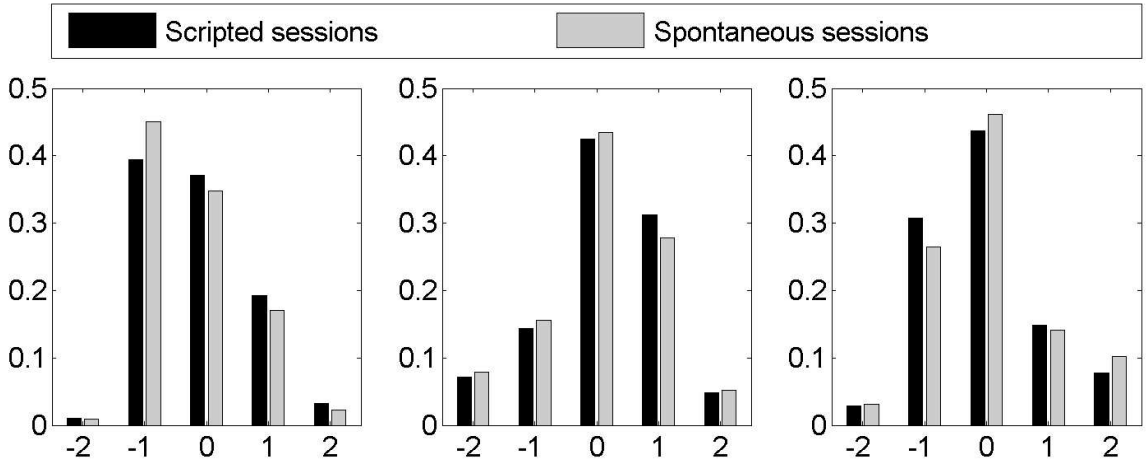
The USC IEMOCAP corpus

Continuous descriptors

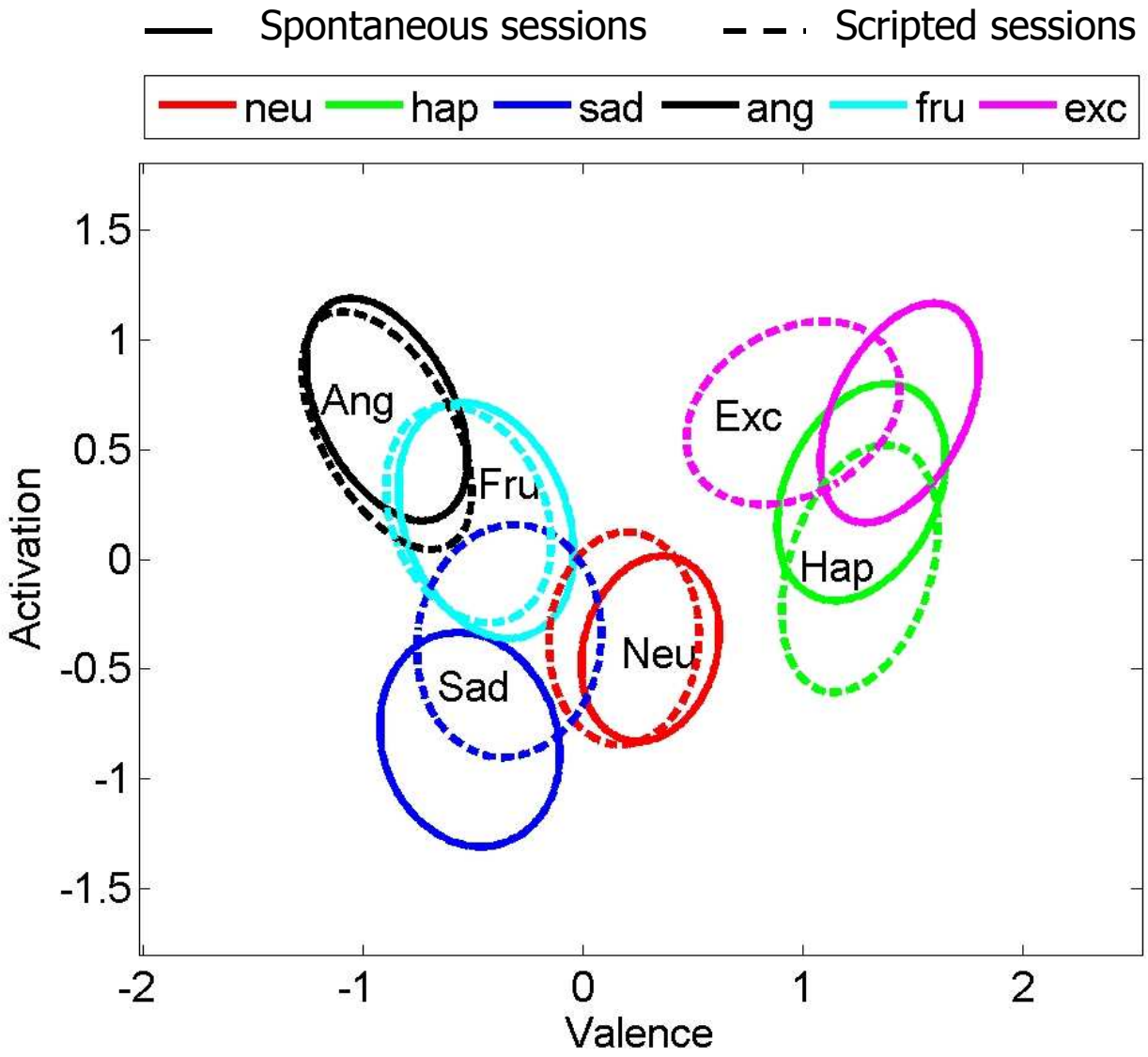
- Two evaluators per turn (80% done)
- Valence
  - 1-positive, 5-negative
- Activation
  - 1-excited, 5-calm
- Dominance
  - 1-weak, 5-strong







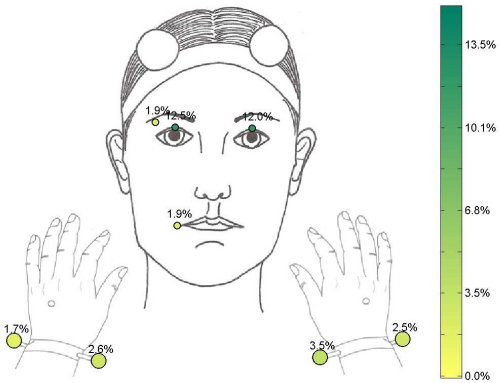
The USC IEMOCAP corpus



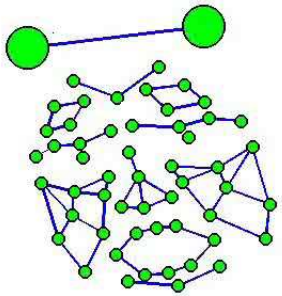
The USC IEMOCAP corpus

Reconstruction of marker data

- The markers' trajectories were reconstructed using the VICON iQ 2.5
- Cubic interpolation was used to fill small gaps
- Few markers were lost during the recording
  - Eyelids and the hands
- Nose marker is assumed to be local coordinate center (translation effect)
- A rotational matrix is estimated for each frame (rotational effects)
  - The technique is based on Singular Value Decomposition (SVD)
- Headband markers were used to ensure good head motion estimation



The USC IEMOCAP corpus



Outline

- ✓ Introduction
- ✓ Guidelines to record databases from actors
- ✓ The USC IEMOCAP corpus
- ➢ Conclusions

Conclusions

- We offered guidelines for designing controlled emotional databases from actors that are closer to the emotions observed in real-life scenarios
  - Contextualization
  - The use of skilled actors
  - The use of different acting styles
  - Suitable emotional descriptors

IEMOCAP

- The IEMOCAP was designed to satisfy many of the key requirements
- This database addresses some of the core limitations of the existing databases
- It can be useful for studies on expressive human communication

Conclusions

Limitations and challenges

- Overlapped speech
- Disfluencies

Future direction

- Study new acting methodologies
- Identify the recording methodologies that will aid emotional recording from actors that resemble real emotions observed in daily human interaction