SCRIPTED DIALOGS VERSUS IMPROVISATION: LESSONS LEARNED ABOUT EMOTIONAL ELICITATION TECHNIQUES FROM THE IEMOCAP DATABASE

Carlos Busso and Shrikanth S. Narayanan
Viterbi School of Engineering, University of Southern California, Los Angeles, CA 90089
(busso@usc.edu, shri@sipi.usc.edu)

Motivation

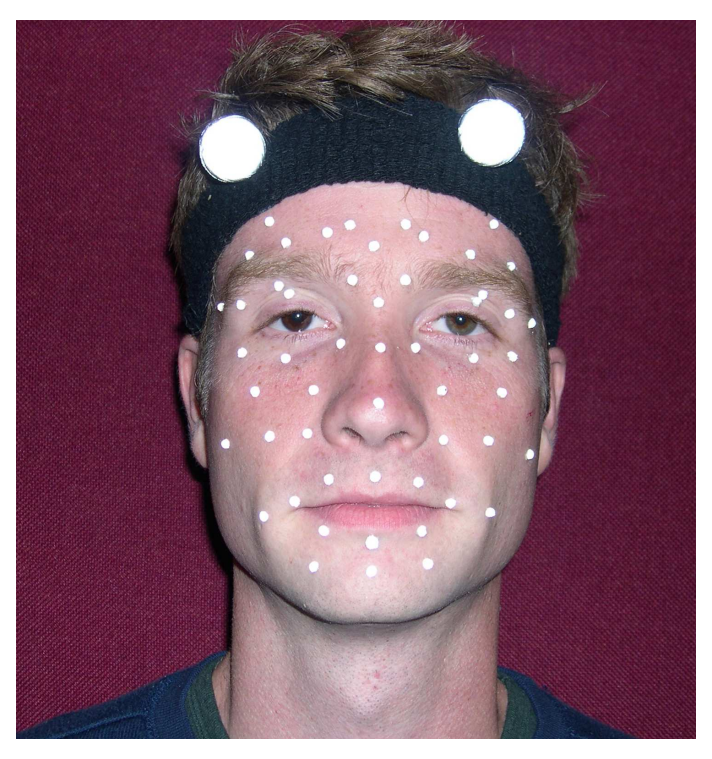
- Collecting natural (non-acted) emotional data present serious limitations
  - Ethical issues, restricted domain, or lack of control (e.g., type of sensors)
- The use of acting appears to be a viable research methodology to study emotions
- Recent efforts have focused on studying better elicitation techniques [1, 2]
- Two appealing elicitation approaches [2]:
  - The use of plays (**Scripted sessions**)
  - Improvisation based on hypothetical scenarios (**Spontaneous sessions**)
- These techniques are rooted in the core of acting training
- Our corpus: *Interactive Emotional Dyadic Motion Capture* database (IEMOCAP)

Goal

To analyze the advantages and limitations of scripted and spontaneous techniques to elicit expressive speech

IEMOCAP database

- Study patterns observed during expressive communication (ten actors) [3]
- Scripted sessions (55% of the corpus)
  - Three 10-minute plays with clear emotional content
  - The actors were asked to memorize and rehearse the scripts
- Spontaneous sessions (45% of the corpus)
  - Eight hypothetical scenarios (e.g., getting married [4])
- Target emotions: happiness, anger, sadness, frustration and neutral state

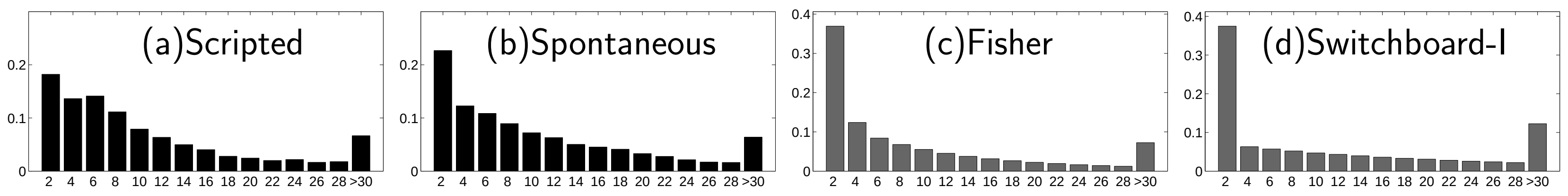


- Sixty-one markers were attached to one participant at a time (five dyadic sessions)
- VICON motion capture system with eight cameras
- The database was segmented and transcribed at the dialog turn level
- Categorical emotional evaluation (3 raters per turn)
  - Happiness,sadness,anger,surprise,fear,disgust,frustration,excited,neutral,and other
- Attribute based emotional evaluation (2 raters per turn, 85.5% completed)
  - *Valence* [1-neg,5-pos], *Activation* [1-calm,5-exc], *Dominance* [1-weak,5-strong]

Spontaneous versus scripted sessions

Lexical content

- Vocabulary size
  - Spontaneous sessions (2864) vs. scripted sessions (1489)
- Utterance duration
  - Scripted sessions tend to have longer utterances
  - 23% of the spontaneous sessions contain only one word (e.g., *yeah*, and *okay*)



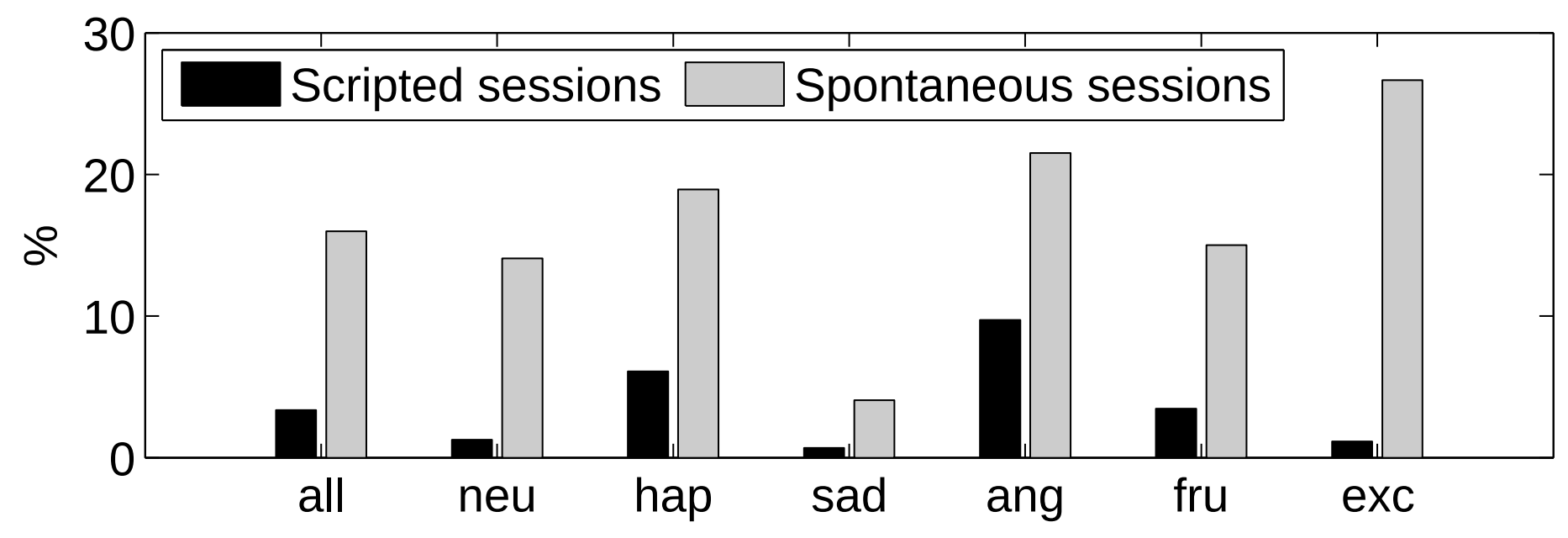
Disfluencies

- Rough approximation of disfluencies
  - Repetitions
  - Fillers (*uh, um, huh, ah, etc.*)
  - Discourse markers (*you know, well*)
  - Editing terms (*I mean,excuse me*)
- Improvisation has more disfluencies
  - Spontaneous sessions (44%)
  - Scripted sessions (30%)

	All disf.	Fillers	Discourse marker	Editing term	Repetition
			Scripted sessions		
All	30.1%	7.4%	14.3%	4.4%	8.6%
Neutral	30.2%	4.9%	23.0%	2.4%	3.9%
Anger	30.4%	8.0%	10.1%	2.8%	13.3%
Happiness	31.4%	11.8%	9.8%	5.9%	7.5%
Sadness	23.7%	1.7%	11.6%	8.6%	7.6%
Frustration	31.9%	5.8%	14.3%	4.9%	11.6%
Excited	44.7%	20.6%	16.1%	5.0%	15.1%
			Spontaneous sessions		
All	44.0%	13.4%	20.9%	10.4%	13.8%
Neutral	53.0%	19.8%	28.4%	13.7%	14.0%
Anger	32.3%	4.9%	12.5%	6.9%	13.5%
Happiness	49.3%	22.0%	24.1%	8.9%	14.2%
Sadness	39.2%	5.8%	21.9%	12.4%	12.7%
Frustration	42.1%	6.7%	17.2%	12.7%	17.5%
Excited	43.5%	18.2%	18.5%	6.8%	12.1%
			References		
Fisher	54.4%	30.5%	22.4%	4.1%	15.6%
Switchboard-I	42.8%	28.4%	16.2%	1.9%	12.9%

Overlapped speech

- Estimated from forced alignment
- Strong emotional dependency
  - Spontaneous (15%)
  - Scripted (5%)

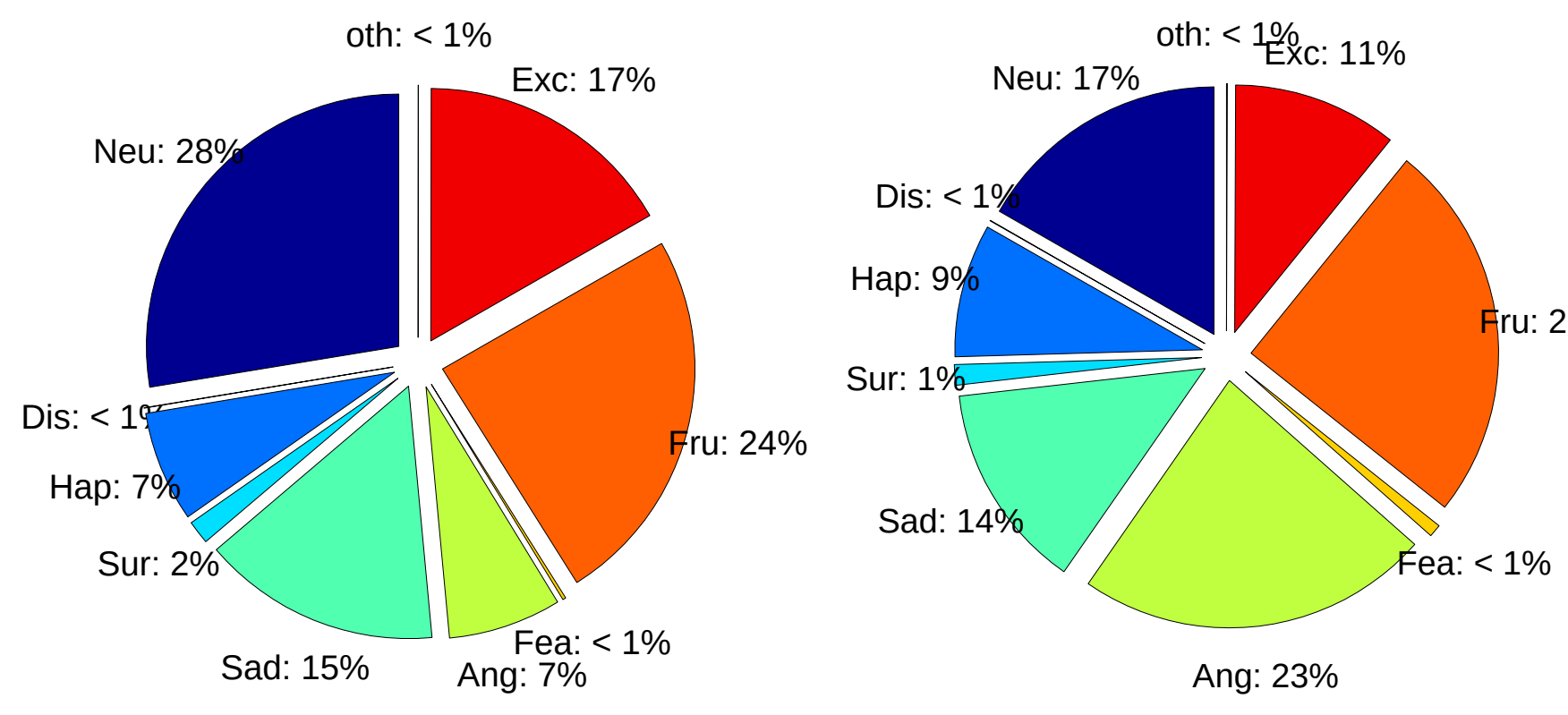


Emotional content

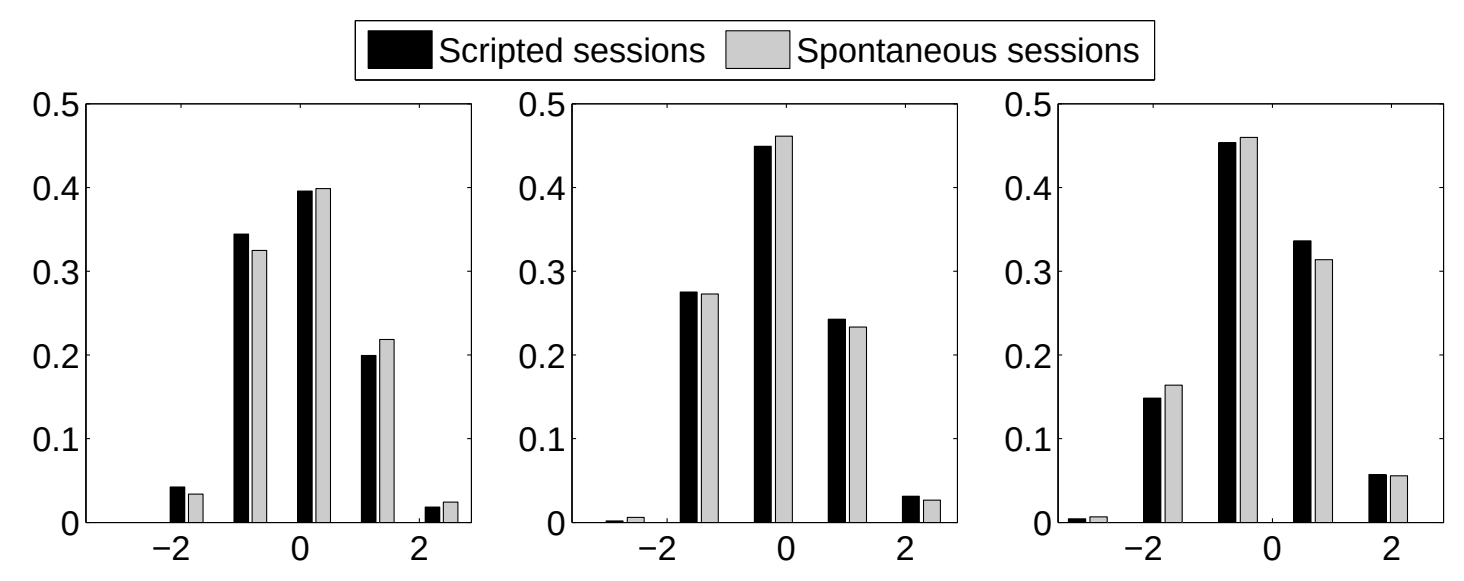
- Inter-evaluator agreement of the emotional categories

	Spontaneous sessions	Scripted sessions
Agreement (majority vote)	83.1%	66.9%
Kappa (Original labels)	$\kappa = 0.34$	$\kappa = 0.20$
Kappa (Combined labels)	$\kappa = 0.44$	$\kappa = 0.26$

- Scripted sessions include progressive changes from one emotional state to another
  - Elicits a wider spectrum of emotional content
  - Boundaries between emotional categories become closer

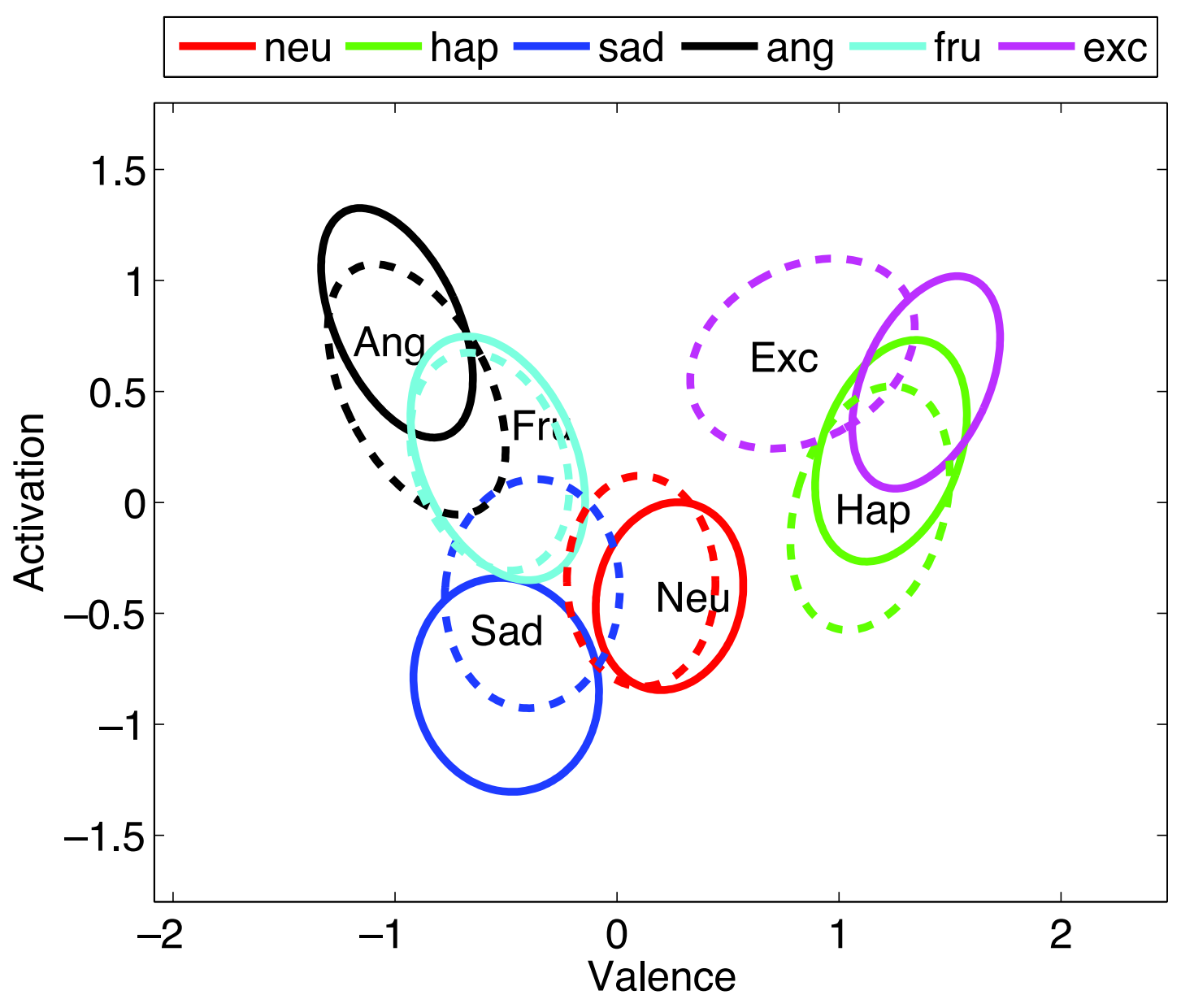


(a) Spontaneous (b) Scripted



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

- Ellipsoid defining confidence region (20%)
- Emotions for scripted sessions (dashed line) are shifted toward the center
  - Emotions in improvisation are more intense
  - They may be easier to recognize
- Actors concentrate on remembering scripts
  - Expression of emotions may be overlooked



- We cannot conclude which technique induces closer real-life emotions

Conclusions

Spontaneous sessions

√ Resulting corpus is similar to natural speech in many aspects
  - Disfluencies, overlapped speech, and turn-taking statistics

√ The scenarios can be easily designed to achieve emotionally balanced corpus

√ Higher vocabulary dimension

√ Spontaneous sessions are found to elicit more intense emotions

√ Higher inter-evaluator agreement on emotional content

× High levels of overlapped speech and disfluencies directly affect post analysis
  - Estimation of speech features (e.g., pitch measurements)

× It requires experienced actors willing to cooperate with each other

Scripted sessions

√ Lexical content is fixed beforehand

√ Low level of overlapped speech simplifies the post analysis steps

√ It may better represent the emotions observed in real-life scenarios

× Emotional boundaries in scripted sessions are more ambiguous

× Remembering dialogs may affect the emotional display
  - The use of experienced actors should mitigate this problem

Future work

Our ultimate goal is to identify better recording methodologies that resemble the emotions observed in real-life scenarios.

- Human perceptual experiments to assess the naturalness of the corpus
- We are planning to systematically analyze different acting styles
  - From fully predetermined (scripted) to fully undetermined (improvised)

References

[1] 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.

[2] F. Enos and J. Hirschberg, "A framework for eliciting emotional speech: Capitalizing on the actors process," in *First International Workshop on Emotion: Corpora for Research on Emotion and Affect (International conference on Language Resources and Evaluation (LREC 2006))*, Genoa,Italy, May 2006, pp. 6–10.

[3] 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. In press, 2008.

[4] K. Scherer, H. Wallbott, and A. Summerfield, *Experiencing emotion: A cross-cultural study*. Cambridge, U.K.: Cambridge University Press, 1986.

Acknowledgements

This research was supported in part by funds from the NSF, and Army