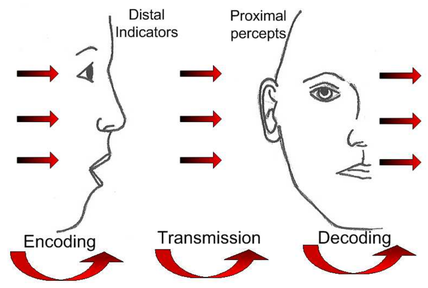
THE EXPRESSION AND PERCEPTION OF EMOTIONS: COMPARING ASSESSMENTS OF SELF VERSUS OTHERS

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

- Perceptual experiments are usually conducted to define emotional labels
- Provide baseline for further research and development (e.g. emotion recognition)
- Underlying assumption: perceived emotion matches intended emotion of the speaker
- It is not guaranteed that this assumption always holds

- Brunswik's lens model
  - Encoding
  - Transmission
  - Decoding



Goal

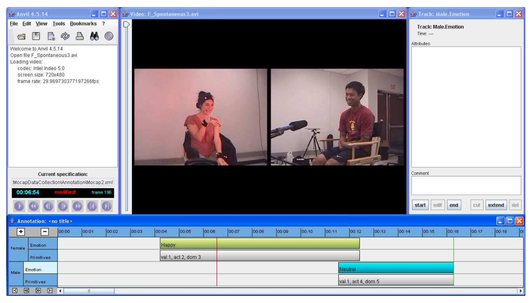
- **Study mismatch between expression and perception of emotion**

Hypothesis

- If self-reports are closer to the intended emotions, then the mismatch between subjective evaluation ("other") and intended emotions can be approximated

IEMOCAP database

- The *interactive emotional dyadic motion capture* (IEMOCAP) database [1]
- 10 actors, dyadic interaction (5 sessions)
- Markers were attached on the face (53), head (2) and hands (6)
- VICON system (8 cameras), 2 digital cameras, and 2 shotgun microphones
- Elicitation techniques: Scripted dialogs and Improvise hypothetical scenarios



- The database was segmented and transcribed at the dialog turn level
- The corpus was emotionally evaluated by
  - **OTHERS**: Naïve evaluators
  - **SELF**: Six of the actors (only spontaneous sessions)
- Categorical emotional evaluation (3 naïve raters, 6 actors per turn) (85.5%)
  - Happiness,sadness,anger,surprise,fear,disgust,frustration,excited,neutral,and other
  - The subjects were allowed to assign multiple labels
- Attribute based emotional evaluation (2 naïve raters, 6 actors per turn)
  - *Valence* [1-neg,5-pos], *Activation* [1-calm,5-exc], *Dominance* [1-weak,5-strong]

Preliminary results [1]

- Listener recognition accuracy between emotional classes
- Reference labels were assigned based on naïve rater assessments (majority vote)
- Recognition rate: *Others* (79%) vs. *Self* (60%)
- Two main limitations
  - The ground reference for the emotional labels is derived from naïve labelers
  - It considers only the turns in which the evaluators reached agreement

	F01	F02	F03	M01	M03	M05	Average
Self	0.79	0.58	0.44	0.74	0.57	0.54	0.60
Others	0.76	0.80	0.79	0.81	0.80	0.77	0.79

Proposed approach

- Leave-one-evaluator-out approach
- The labels of one rater are compared with the labels of the rest of the raters

Category based annotation

1- Fleiss' *Kappa* statistic

2- Entropy-based metric proposed by Steidl *et al*[2]

- Originality proposed to measure performance of emotion recognition systems
- Entropy is defined as a measure of uncertainly of a random variable
  - Variables uniformly distributed have maximum entropy
- The higher the agreement, the lower the entropy
  - 2 happiness, 1 excited, 1 surprise p=[0.50, 0.25, 0.25] (without evaluator)
  - 3 happiness, 1 excited, 1 surprise $\bar{p}$=[0.60, 0.20, 0.20] (with evaluator)
  - $S_{ent} = 1.37 - 1.5 = -0.13$

$$S_{ent} = H(\bar{p}) - H(p) = -\left(\sum \bar{p} \cdot log\bar{p} - \sum p \cdot logp\right) \quad (1)$$

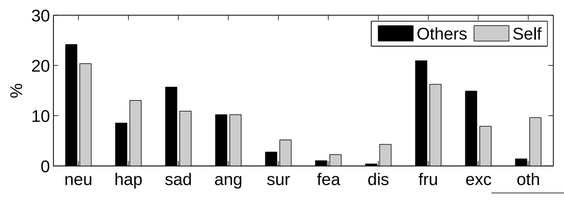
Attribute-based annotation

1- Euclidean distance in the VAD space

2- Correlation between evaluations

Analysis of self and others evaluation

Categorical emotional descriptors



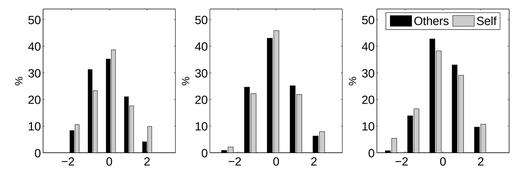
- Distribution of the emotional labels
- Self reports were more specific with label "other"
  - Self (9.6%) vs. Others (1.4%)
  - Irritation, curious, shocked, etc.

		Number turns	Entropy S_{ent}	Kappa statistic w/o eval.	Kappa statistic with eval.	$\Delta\kappa$
Others	Subject E1	2352	0.164	0.358	0.331	-0.028
	Subject E2	2246	0.064	0.284	0.333	0.049
	Subject E3	158	0.163	0.247	0.264	0.017
	Subject E4	2117	0.101	0.329	0.340	0.011
	Subject E5	56	0.301	0.197	0.165	-0.031
	Subject E6	290	0.115	0.205	0.241	0.036
	Average		**0.113**	**0.270**	**0.279**	**0.009**
Self	Actress F01	382	0.267	0.276	0.263	-0.013
	Actress F02	388	0.224	0.393	0.355	-0.038
	Actress F03	535	0.235	0.338	0.299	-0.039
	Actor M01	376	0.166	0.398	0.391	-0.007
	Actor M03	507	0.196	0.366	0.341	-0.024
	Actor M05	221	0.184	0.285	0.275	-0.010
	Average		**0.215**	**0.343**	**0.321**	**-0.022**

- Entropy metric
  - Entropy scores are higher in self-report
  - ANOVA ($p << 0.005$)
- *Kappa* statistic
  - Increases for naïve evaluators
  - Decreases for self evaluations
- Agreement decreases (self ratings)
- Mismatch in emotional perception

Continuous emotional descriptors

- Each dialog turn was evaluated by only three subjects (2 naïve raters, 1 actor)
- Speaker dependent z-normalization
- Self-reports have more extreme values (1,5)
  - 20.4% vs. 12.5% (Valence)
  - 10.0% vs. 7.2% (Activation)
  - 16.1% vs. 10.4% (Dominance)



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

		Number of turns	Euclidean distance	Correlation Val.	Act.	Dom.
Others	Subject E7	1811	1.363	0.746	0.527	0.497
	Subject E8	1811	1.324	0.828	0.643	0.397
	Average		**1.343**	**0.787**	**0.585**	**0.447**
Self	Actress F01	280	1.234	0.784	0.656	0.475
	Actress F02	297	1.394	0.839	0.593	0.239
	Actress F03	421	1.256	0.759	0.703	0.472
	Actor M01	274	1.338	0.785	0.647	0.564
	Actor M03	371	1.230	0.802	0.653	0.575
	Actor M05	168	1.456	0.801	0.488	0.084
	Average		**1.301**	**0.795**	**0.623**	**0.402**

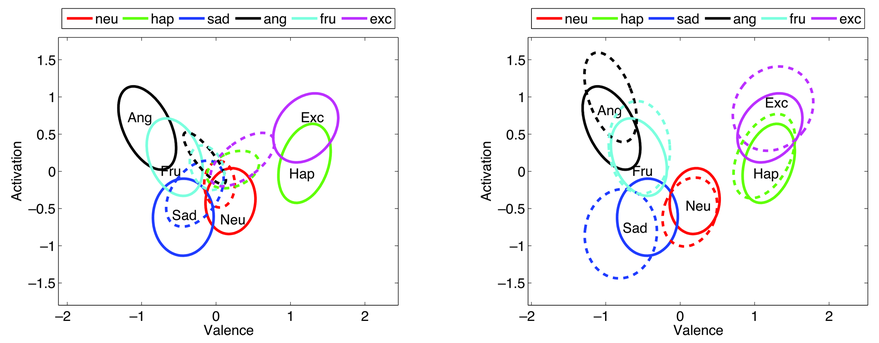
- Euclidean distance and correlation (VAD)
- Each rater compared with the mean between others evaluators
- Self and naïve assessments are similar
- ANOVA, Euclidean distance ($p = 0.115$)

Categorical labels in the valence-activation space

- Ellipsoids define confidence regions (20%)
- Naïve evaluators (solid line), self-evaluators (dashed line)

(a) Emotional labels from the naïve evaluators are used as a reference for both
  - Ellipsoids for the self-reports are shifted to the center (0,0)
  - Happy or angry turns are perceived more neutral by self-reports

(b) Emotional labels of the dialog turns are separately assigned
  - Ellipsoids for the self-reports are shifted away from the center
  - Concept of emotional categories for the self-reports is more extreme



(a) Labels from naïve raters are imposed (b) Labels are separately assigned

Discussion and conclusions

- There is a mismatch between the expression and perception of emotions
  - Especially with categorical assessment
- Inter-labeler agreement significantly decreases when the self-reports are considered
- Subjective evaluations may not accurately describe true emotions
  - Implication in automatic emotion recognition (potentially inaccurate)

Limitation and Future work

- Actors may look to different cues than naïve listeners
- Results rely in the assumption that *self*-reports are closer to the intended emotions
- Replicate this study with more subjects with natural (non-acted) emotional database

References

[1] 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.

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

Acknowledgements

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