The expression and perception of emotions: Comparing Assessments of Self versus Others

Carlos Busso and Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory (SAIL)
Electrical Engineering Department
University of Southern California, Los Angeles, CA 90089
busso@usc.edu, shri@sipi.usc.edu

Abstract

In the study of expressive speech communication, it is commonly accepted that the emotion perceived by the listener is a good approximation of the intended emotion conveyed by the speaker. This paper analyzes the validity of this assumption by comparing the mismatches between the assessments made by naïve listeners and by the speakers that generated the data. The analysis is based on the hypothesis that people are better decoders of their own emotions. Therefore, self-assessments will be closer to the intended emotions. Using the IEMOCAP database, discrete (categorical) and continuous (attribute) emotional assessments evaluated by the actors and naïve listeners are compared. The results indicate that there is a mismatch between the expression and perception of emotion. The speakers in the database assigned their own emotions to more specific emotional categories, which led to more extreme values in the activation-valence space.

Index Terms: Emotion, emotional perception, expression of emotion, inter-evaluator agreement

1. Introduction

An important paralinguistic aspect of human interaction is the emotional state of the speaker. The face [1], speech [2], and body posture [3] are all used as channels to convey the intended emotion. In many studies, perceptual experiments are usually conducted to define the emotional state conveyed by the subjects recorded in the corpus, and are used to baseline further research and development such as automatic emotion recognition [4, 5]. The implicit assumption in these studies is that the perceived emotion matches the intended emotions of the speaker. Although we are very good at recognizing even subtle emotional cues, it is not guaranteed that this assumption always holds.

An interesting model to study the mismatch between expression and perception of emotions is the modified version of the Brunswik's lens model, proposed by Scherer [2]. This model includes three main processes: the encoding, the transmission, and the decoding of emotions. In the encoding step, the speaker modifies his/her communicative channels (*distal cues*) to convey his/her internal affective state. The observer will perceive these cues (*proximal cues*) and will make an inference about the emotional state of the speaker. Although the proximal cues (percepts) are based on the distal cues, they are not necessarily equivalent since they may be corrupted in the transmission or in the interpretation of the emotions [2]. In this context, self-reports are useful to analyze mismatch between distal and proximal cues.

This paper addresses aspects of the mismatch between expression and perception of emotion, an important problem not only in theory, but also for practical research areas such as automatic emotion recognition. The proposed approach is based on comparing the emotional evaluations as perceived by naïve speakers (*others*), and the self-evaluations from the participants used for the recordings (*self*). The underlying assumption is that people can better decode, up to some extent, their own emotions than do naïve listeners. Therefore, their evaluations will be closer to their intended emotions (*distal indications*). For this purpose, the *interactive emotional dyadic motion capture* (IEMOCAP) database is used [6, 7]. In this corpus, ten actors were recorded in two-party interactions. Two elicitation techniques were used: scripted dialogs (scripted sessions) and improvisation of hypothetical scenarios (spontaneous sessions). Six evaluators assessed the emotional content of the corpus in terms of categorical emotional labels, and continuous primitive attributes. In addition, six of the actors were asked to evaluate their emotions in the spontaneous sessions using the same descriptors. The *self* and *others* categorical evaluations are compared in terms of the entropy-based approach proposed by Steidl *et al.* [8] and the *Kappa statistic*. For the attribute-based evaluations, the *self* and *others* assessments are compared in terms of the Euclidean distance and the correlation between the raters.

The results indicate that the categorical evaluations from the *self*-reports differ from the assessments evaluated by naïve listeners. In fact, the inter-labeler agreement significantly decreases when the *self*-reports are considered. When the dialog turns are plotted in the valence-activation space, the emotional category clusters for the *self*-reports are shifted toward more extreme values. This result suggests that the conceptualization of the emotional categories for the actors was more intense than for naïve listeners, in terms of activation and valence.

The paper is organized as follows. Section 2 describes the IEMOCAP corpus and the metrics used to compare the emotional evaluation from the actors and the naïve listeners. Section 3 presents the analysis of the differences between *self* and *others* evaluations in terms of categorical and continuous emotional descriptors. Finally, Section 4 gives the discussion, future directions and final remarks.

2. Methodology

2.1. IEMOCAP database

The *interactive emotional dyadic motion capture* (IEMOCAP) database was used for the analysis. In this corpus, seven professional actors and three senior students in the Drama Department at the University of Southern California were each recorded in dyadic interaction. The actors' motion was captured during the recording by attaching markers on the face, head and hands. Furthermore, two digital cameras and two directional shotgun microphones were used to produce audio-visual recordings of the sessions. The actors were asked to perform scripted dialogs

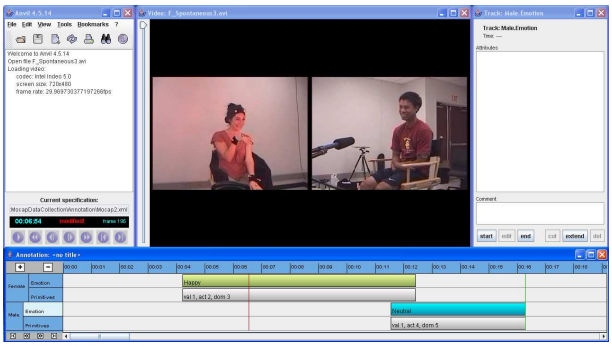


Figure 1: ANVIL annotation tool used to evaluate the emotional content of the IEMOCAP database, in terms of discrete (categories) and continuous (attribute) emotional descriptors.

(scripted sessions) and improvise hypothetical scenarios (spontaneous sessions).

The corpus was transcribed and segmented at the dialog turn level. The emotional content was evaluated with subjective experiments using the ANVIL annotation tool [9] (Fig. 1). Six naïve listeners assessed the corpus in terms of emotional categories. Although the corpus was designed to target anger, happiness, sadness, frustration and neutral state, the emotional categories were extended to provide a wider emotional description (surprise, excited, fear, disgust, and other). The evaluation was arranged such that three different raters assessed each turn. The subjects were allowed to assign multiple labels if necessary. If none of the labels was adequate to describe the emotional content of the turns, they were asked to choose "other" and type their own emotional label.

Likewise, two different evaluators assessed the emotional content in terms of continuous attributes: *valence* [1-negative, 5-positive], *activation* [1-calm, 5-excited], and *dominance* [1-weak, 5-strong] (VAD). At the time of this writing, approximately 85.5% of the turns have been evaluated. In both types of emotional evaluation, the raters where asked to sequentially assess the turn, after watching the videos. Therefore, the acoustic and visual channel, and the context of the dialogs were available to make the decision.

In addition, six of the actors that participated in the recording were asked to evaluate their own dialogs using both categorical and attribute descriptors. Since the actors were asked to evaluate only the spontaneous/unscripted sessions, this paper analyzes only that portion of the database. They assessed the emotional content of the corpus using the same settings and tools used by the naïve evaluators. Details of the corpus are given in [6, 7].

2.2. Proposed approach to study inter-evaluator agreement

In our previous work, we presented a simple comparison between *self* and *others* evaluations using categorical descriptors (i.e., sadness, happiness) [6]. The listener recognition accuracy between emotional classes was estimated for each actor, after assigning reference labels to the turns based on the naïve rater assessments (majority vote). Table 1 gives recognition accuracy for the actors (*self*) and naïve raters (*others*). This table shows that the recognition rate for the actors was consistently lower than the recognition rate of naïve listeners, suggesting that there is a mismatch between the expression and perception of the emotions. However, the methodology used to create this table has two main limitations. First, the ground reference for the emotional labels was estimated only with the results from the naïve labelers. Therefore, the recognition accuracy results are expected to be lower for the actors. Second, this analysis only considered the portion of the database in which the evaluators reached agreement.

Table 1: Comparison of the recognition rate in percentage between *self* and *others* evaluations [6].

	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

In this paper, we proposed an improved approach to address the mismatch between *self* and *others* evaluations. Instead of defining a ground emotional reference, the labels of one evaluator are compared with the labels of the rest of the evaluators (leave-one-evaluator-out approach) [8].

For category emotional annotation, we adopted the ideas of the entropy-based metric proposed by Steidl *et al.*[8]. This metric was initially proposed to measure the performance of automatic emotion recognition systems by weighting the classification results according to the performance (agreement) observed by human labelers. The entropy is defined as a measure of uncertainly of a random variable [10]. If the distribution of this random variable is uniformly distributed, the entropy will reach its maximum. In contrast, if we know the value of the random variable, the entropy is zero. This concept can be applied to measure inter-evaluator agreement (the higher the agreement, the lower the entropy). Let's assume that we have a probability distribution p with the emotional evaluations from n subjects (i.e., 0.5 happiness, 0.25 excited, 0.25 surprise). If the results from a new rater is available, we can estimate a new distribution, $\bar{p}$ (e.g., 0.6 happiness, 0.2 excited, 0.2 surprise). We propose to estimate the difference between the entropies of the two distributions to measure the impact of the new labeler in the overall agreement (Eq. 1). A negative value means entropy decreased, resulting in higher agreement probability (in the example, $S_{ent} = 1.37 - 1.5 = -0.13$).

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

We also measure the inter-evaluation agreement using Fleiss' *Kappa* statistic [11]. For the attribute-based annotation, we used two different metrics: Euclidean distance in the VAD space and correlation between evaluations. More details are given in Sections 3.1 and 3.2.

3. Analysis of self and others evaluation

3.1. Categorical emotional descriptors

As mentioned in Section 2.1, six evaluators and six of the actors that participated in the recording were asked to asses the emotional content of the corpus in terms of discrete categories for each dialog turns. Figure 2 shows the distribution of the emotional labels assigned by the naïve raters (*others*) and the actors (*self*). An interesting result is that the actors were more specific in their classification of the emotion, choosing the label "other" for 9.6% of the turns (1.4% for naïve evaluators). Some of the emotional labels suggested by the actors are irritation, curious, shocked, and emphatic.

The results of the inter-labeler agreement for the actors and naïve evaluators measured in terms of the entropy metric and the *Kappa* statistic are presented in Table 2. This table also shows the number of turns evaluated for each subject. Table 2 shows that the average entropy score for the naïve evaluators is lower than for the self evaluations. A one-way analysis of variance (ANOVA) test indicates that the observed difference is significant ($p << 0.005$). In fact, only the evaluator E5 presented higher average score than the any of the actors (although this rater evaluated only 56 turns). This result indicates that the

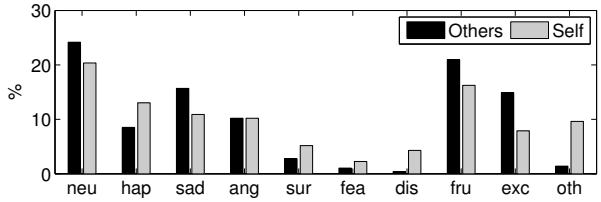


Figure 2: Distribution of the emotional content of the dialog turns in terms of the emotional categories.

Table 2: Inter-evaluator agreement for categorical descriptors, using leave-one-evaluator-out approach.

		Number turns	Entropy S_{ent}	Kappa statistic w/o eval.	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**

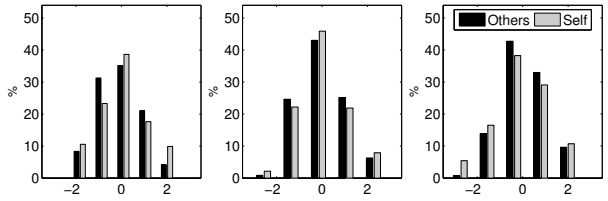
emotional perception of the actors does not match the emotional perception of other evaluators.

Table 2 also shows the inter-evaluator results using the *Kappa* statistic. The procedure used to measure the impact of each evaluator in the inter-labeler agreement was to compute the standard *Kappa* statistic in two conditions: with the assessments from all the evaluators except one of them (w/o eval.), and with the assessments from all the evaluators (with eval.). For example, if the labeler E1 is not considered, the *Kappa* statistic is equal to $\kappa = 0.358$. When this labeler is considered, the *Kappa* statistic decreased to $\kappa = 0.331$. The results also indicate that emotional self-evaluation differs from the assessments made by naïve listeners. Notice that the *Kappa* statistic decreases for all the actors when their assessments are included. In contrast, the inter-labeler agreement increases for 4 out of 6 of the evaluators (E2, E3, E4 and E6).

3.2. Continuous emotional descriptors

As mentioned in Section 2.1, two evaluators and six of the actors assessed the emotional content of the IEMOCAP corpus in terms of the attribute valence, activation and dominance, using a 5-point scale. Since there is no overlap between the self-evaluations, each dialog turn was evaluated by only three subjects (two naïve raters and one actor). To compensate for inter-evaluator variation, speaker dependent z-normalization was applied for each attribute (zero mean and standard deviation equal to one). Figure 3 shows the emotional content of the IEMOCAP corpus, in terms of the VAD attributes. The figure indicates that the actors labeled more turns with extreme values (1 or 5, before normalization) than naïve evaluators (valence: 20.4% versus 12.5%, activation: 10.0% versus 7.2%, dominance: 16.1% versus 10.4%).

As mentioned in Section 2.2, we use a leave-one-evaluator-out approach to contrast the emotional assessments of one rater with the others. Here, the average values of the VAD attributes between two evaluators are compared with the assessments of



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

Figure 3: Distribution of the emotional content of the turns in terms of the attributes *valence* [negative-possitive], *activation* [calm-excited] and *dominance* [weak-strong].

Table 3: Inter-evaluator agreement for attribute descriptors, using leave-one-evaluator-out approach.

		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**

the third evaluator. The results of the *self* and *other* evaluation in terms of Euclidean distance and the correlation in the VAD space are presented in Table 2. This table also gives the number of turns evaluated for each subject. Notice that the number of turns for the actors is different from the values reported in Table 2. The reason of this difference is that we are comparing only the turns that have been also evaluated by the naïve speakers (85.5% of the corpus). This table suggests that the self-assessments are similar to the assessments made by the evaluators. In fact, a one-way ANOVA test indicates that there is no significant difference in the Euclidean distance between *self* and *others* evaluations ($p = 0.115$).

3.3. Categorical labels in the valence-activation space

The results in Section 3.1 suggests that the perception of emotional categories by the actors differs from the emotional perception by the naïve evaluators. In this section, the categorical assessments are projected into the valence-activation space, to further analyze the differences between *self* and *others* evaluations.

In Figure 4, the emotional labels from the naïve evaluators are used as a reference for both the *self* and *others* assessments. For the turns that the naïve evaluators' majority vote reached agreement, the mean vector and covariance matrix of the valence and activation were separately estimated for each emotional category. These statistics were used to plot ellipsoids that define 20% confidence regions. Figure 4 shows the results for the naïve evaluators (solid line) and the actors (dashed line). This figure shows that for all the emotional categories the ellipsoids for the actors are shifted to the center (0,0). This indicates that the turns that are perceived as happy, angry, or excited by the naïve evaluators, are perceived by the actors with neutral values of valence and activation. Notice that the emotional categories disgust, fear and surprise were not included in the figure since only a few turns were labeled with these emotional categories (Fig. 2).

In Figure 5, the emotional labels of the dialog turns are sep-

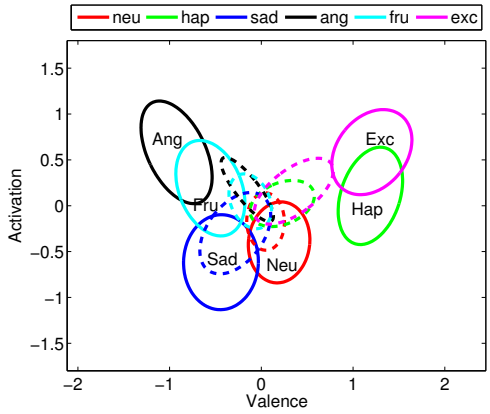


Figure 4: Clustering of the emotional categories in terms of the valence and activation. The categorical labels from the naïve evaluators are imposed to the actors' assessments. The results are presented for *others* (solid line), and *self* (dashed line) judgments.

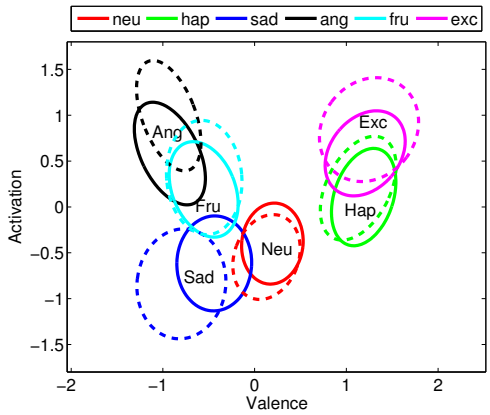


Figure 5: Clustering of the emotional categories in terms of the valence and activation. The categorical labels are separately assigned to the *self* and *other* reports. The results are presented for *others* (solid line), and *self* (dashed line) judgments.

arately assigned for the naïve evaluators and for the actors, according to their assessments. This figure shows that the ellipsoids for the actors are shifted away from the center. This result suggests that the concept of emotional categories for the self-reports is more intense than for the other evaluators. In other words, the levels of arousal and valence need to be higher to self-assign an emotional category to the turns.

4. Discussion and conclusions

This paper analyzed the emotional assessments of the IEMOCAP database made by naïve evaluators, and the actors that participated in the recording. Under the assumption that *self*-report evaluations are closer to the intended emotion of the speaker, this paper suggests that there is a mismatch between the expression and perception of emotions between speakers and listeners. Using an entropy-based metric and the *Kappa* statistic, we showed that the inter-labeler agreement significantly decreases when the self-reports are considered. The actors seem to be more selective in their assignment of emotional categories. This is exemplified in the activation-valence space, where the *self* values are more extreme than the *others* values.

The implication of these results is that subjective emotional evaluations made by listeners may not accurately describe the true emotions conveyed by the speaker. From an automatic recognition perspective, these results are relevant since the ultimate goal may not be just to recognize what others are perceiving, but what the user is expressing or feeling. However, with the current technology we do not have access the intended emotions of the speaker. Since self-reports are not always available, the emotional description obtained from subjective experiments remains the best approximation of the emotion conveyed by the speaker (but should be viewed as potentially inaccurate).

The claims made in this paper are limited by the assumption that *self*-reports are closer to the intended emotions of the speaker. Since humans may not even be aware of their internal affective state, the proposed approach to analyze the expression and perception of emotion provides a partial view to address the problem. Other methodologies could be designed to better approximate the *distal indicators* of the subjects. Likewise, we only analyzed the emotional assessments of 14 subjects. It would be interesting to replicate this study with more subjects. Also, we suggest analyzing a natural (non-acted) emotional database. Actors have been trained to induce emotional reactions in the audience, so they may look to different cues than naïve listeners. These are topics of future work.

5. Acknowledgements

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

6. References

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

[2] K. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech Communication*, vol. 40, no. 1-2, pp. 227–256, April 2003.

[3] M. Coulson, "Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence," *Journal of Nonverbal Behavior*, vol. 28, no. 2, pp. 117–139, June 2004.

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

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

[6] 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.

[7] 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.

[8] 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.

[9] M. Kipp, "ANVIL - a generic annotation tool for multimodal dialogue," in *European Conference on Speech Communication and Technology (Eurospeech)*, Aalborg, Denmark, September 2001, pp. 1367–1370.

[10] T. Cover and J. Thomas, *Elements of Information Theory*. New York, NY, USA: Wiley-Interscience, 2006.

[11] J. Fleiss, *Statistical methods for rates and proportions*. New York, NY, USA: John Wiley & Sons, 1981.