Evaluating Evaluators: A Case Study in Understanding the Benefits and Pitfalls of Multi-Evaluator Modeling

Emily Mower[† ‡], Maja J Matarić[‡], Shrikanth Narayanan[†]

University of Southern California
[†] Signal Analysis and Interpretation Laboratory, [‡]Interaction Laboratory
University Park, Los Angeles, California, USA 90089
mower@usc.edu, mataric@usc.edu, shri@sipi.usc.edu

Abstract

Emotion perception is a complex process, often measured using stimuli presentation experiments that query evaluators for their perceptual ratings of emotional cues. These evaluations contain large amounts of variability both related and unrelated to the evaluated utterances. One approach to handling this variability is to model emotion perception at the individual level. However, the perceptions of specific users may not adequately capture the emotional acoustic properties of an utterance. This problem can be mitigated by the common technique of averaging evaluations from multiple users. We demonstrate that this averaging procedure improves classification performance when compared to classification results from models created using individual-specific evaluations. We also demonstrate that the performance increases are related to the consistency with which evaluators label data. These results suggest that the acoustic properties of emotional speech are better captured using models formed from averaged evaluations rather than from individual-specific evaluations.

Index Terms: Emotion, Hidden Markov Model, Perception

1. Introduction

Quantitative models of user perception have the potential to facilitate the design of synthetic emotional expressions. These models could lead to computer agents and robots that more naturally and functionally blend into human society [1, 2]. User specific emotional modeling and synthesis requires an understanding of human emotional perception, often measured using stimuli presentation experiments. Unfortunately, the evaluation process is non-stationary. Subjective emotional appraisals of evaluators change as they tire and as they are exposed to increasing numbers of emotional utterances. It is common practice to estimate emotional ground truth by averaging evaluations from multiple evaluators. The question remains as to whether this averaging between evaluators with different internal emotional representations sacrifices important individual information.

Emotion recognition has been studied extensively [3–6]. However, these studies do not provide analyses of inter-evaluator differences. In [7–9], the authors present an analysis of the differences between self-evaluations and the evaluations of others. In [10], the authors present a new emotion classification accuracy metric that considers common inter-evaluator emotion classification errors. The question of inter-evaluator averaging remains unexplored.

This paper presents an analysis of human emotion evaluation. The foci of this paper are: the measurement of the consistency between evaluators and the evaluation of evaluators based on automatic emotion recognition. Firstly, we study the consistency between the categorical emotion labels (e.g., angry, happy, sad, neutral) of the utterances and the evaluators' internal representation of valence (positive vs. negative) and activation (calm vs. excited), using Naïve Bayes classification. This classification framework is used to estimate the categorical emotion of an utterance given individual evaluators' ratings of valence and activation. It is hypothesized that higher performance will be observed for evaluators with internally consistent representations of the dimensional emotional space, since emotional categories are located in specific regions of this space [11]. Secondly, we model the relationship between the temporal acoustic properties of a clip and the subjective valence or activation rating using Hidden Markov Models. The conventional approach in emotion recognition is to use subjective evaluations to measure the performance of the system. We propose the opposite approach: to use the results of an emotion recognition system to measure the accuracy of subjective evaluations. We hypothesize that the performance of the automatic system will increase, if the emotional labels are accurate. These models are compared across three conditions: a) training and testing on ***individual*** data; b) training and testing on ***averaged*** data; and c) training on ***averaged*** data and testing on ***individual*** data.

Human evaluators are as unique as snowflakes. Consequently, one would expect that HMMs trained on individual-specific data would better capture the variability inherent in the individual's evaluation style. However, the presented experiments demonstrate that models trained on averaged data either outperform or perform comparably to those trained solely on the individual-specific data. The results also suggest that evaluations from individuals with a higher level of internal emotional consistency are more representative of the emotional acoustic properties of the clips than those of less consistent evaluators.

The remainder of the paper will be presented as follows: The data are described in Section 2. The approach is outlined in Section 3. The results are presented in Section 4. The discussion is provided in Section 5. Finally, the conclusion is presented in Section 6.

2. Data

2.1. IEMOCAP Data

Our data were drawn from the audio-visual Interactive Emotional Dyadic Motion Capture Database (IEMOCAP) [12]. In this dataset, five mixed-gender pairs of actors (ten actors total) performed selections from emotionally salient plays and improvisations based on emotionally salient scenarios. We use the audio files from the five female actresses.

The data were evaluated using two methods: categorical and dimensional evaluation. In the categorical evaluation task, evaluators tag a clip with a semantic emotional label (e.g., angry or happy). In the dimensional evaluation task, evaluators assign a clip to a location within an emotional space. We use an emotional space consisting of the dimensions of valence and activation, both discretized to a five-point scale. Valence describes the positive vs. negative aspect of the emotion [1 = most negative, 5 = most positive]. Activation describes the calm vs. excited aspect of the emotion [1 = most calm, 5 = most excited].

In this paper we present dimensional classification analyses. We consider only clips labeled (using majority voting over the categorical evaluators) as angry, happy, sad, or neutral. We present results using dimensional evaluations from the two evaluators who evaluated the largest quantity of data (evaluators one and two). Evaluator one analyzed 1,773 clips and evaluator two analyzed 1,682 clips from the angry, happy, sad, neutral set. The clips evaluated by evaluator two are a subset of those evaluated by evaluator one. Please see [12] for more database details.

2.2. Audio Features

We extracted 13 filterbanks of Mel Filterbanks (MFB), their delta, and acceleration from the audio files. MFBs model the human auditory system by creating filterbanks of increasing width as the frequency increases. This structure approximates the increasing de-sensitivity in human hearing to deviations in frequency as the frequency content of the signal increases. MFBs are used in the calculation of Mel Frequency Cepstral Coefficients (MFCC). They have been shown [3] to contain more emotional information than MFCCs.

2.3. Treatment of Evaluations

This paper presents two types of evaluator studies: ***individual*** and ***averaged***. The experiments based upon *individual* evaluations study the dimensional evaluator behaviors of the two evaluators, evaluator one and evaluator two, separately. These evaluations are neither averaged nor normalized. The *averaged* evaluations are the averages of the valence and activation ratings of evaluators one and two. The averaged rating is always rounded up to the nearest integer.

3. Approach

There are two points of interest that arise when considering evaluator performance: evaluator consistency (how similarly evaluators rate clips of the same semantic label) and evaluator reliability (how representative the labels are of the acoustic properties of the utterance). To answer these questions we consider two probabilistic modeling techniques: Naïve Bayes and Hidden Markov Models, respectively. Previous work has demonstrated the efficacy of utilizing Naïve Bayes to recognize the emotional content of speech [13] and Hidden Markov Models to capture its underlying temporal properties [3].

3.1. Naïve Bayes

Research has shown that categorical emotions can be depicted as occupying specific portions of a dimensional space defined by valence and activation [5,11,14,15]. For example, archetypal angry emotions lie within an area defined by negative valence and high activation, while archetypal happy emotions lie within an area defined by positive valence and mid to high activation. This suggests that given only an evaluator's subjective evaluation of the valence and activation of an utterance, it should be possible to estimate the categorical emotion label of the utterance [5]. One measure of evaluator consistency is to determine how well a simple classification algorithm predicts the categorical emotion label of a clip given the subject evaluation of valence and activation. We use Naïve Bayes classification for this analysis. In this classification task, evaluator one's and two's subjective valence and activation ratings are used to predict the *majority voted* categorical label.

3.2. Hidden Markov Models

In the emotional evaluation process, there exists a dependency between assigned evaluation and the temporal acoustic properties of an utterance. We use Hidden Markov Models (HMM) to model the relationship between the temporal fluctuations of the acoustic properties and the resulting reported emotion perception. The HMM classification accuracies provide insight regarding how representative the subjective dimensional tags of the evaluators are of the underlying emotional acoustic properties of the utterances. The accuracies are also used to analyze the effectiveness of averaging subjective, unnormalized evaluations obtained from multiple individuals.

We describe two separate classification tasks: valence and activation. In these tasks, the original five point scale was collapsed into a three point scale to combat a data sparsity issue. Classes one and five were not tagged with sufficient frequency to form models across both evaluators. Class three, representing neutral (for both valence and activation), remained unchanged. Classes one and two (either negative valence or lowly activated) were collapsed into a single class and classes four and five (either positive valence or highly activated) were collapsed into a single class. This resulted in three model groups, one for each, of the valence and activation classification tasks.

In each model group the data were modeled at the phoneme level. The phonemes were clustered into seven classes to retain a large quantity of data per model. The seven classes included: front vowels, back/mid vowels, dipthong, liquid, nasal, stop consenants, and fricatives (see [3] for a detailed mapping). The utterances each had accompanying transcription files at the word and phoneme level, generated using forced alignment [12]. The phoneme-level transcription files were modified for each utterance, replacing the original phonemes with phoneme classes (see Table 1 for an example). In each classification task there were a total of three model groups, each with seven phoneme class models, plus emotion-independent models for silence and laughter for a total of 23 models.

The HMMs were trained using HTK [16]. Each model (valence, activation) had three-states and eight mixture components. The emotional HMMs were trained in two ways: a) using individual-specific evaluations, and b) using averaged evaluations. The individual-specific HMMs were tested using individual-specific evaluations. The averaged HMMs were tested using both individual-specific evaluations and averaged evaluations. The testing procedure utilized word-level forced alignment using *-I in HVite*. This focused the classification task on the identification of the correct emotional phoneme class, rather than the correct phoneme and the correct emotion class. The output of the HMM classification consisted of a transcript file containing the estimated emotional phoneme states over specified time windows. The final emotion of the utterance was assigned using majority voting over the estimated emotional phonemes, weighted by the time duration of each assigned emotional phoneme class. The emotion class represented most frequently in the output transcription was assigned as the final class label.

Original Data		Transformed Data		
Time	Content	Time	Ph.	Ph. Class
0 - 31	silence	0 - 47	sil	sil
48 - 57	what	48 - 51	W	one_liquid
		52 - 54	AH	one_back/mid
		55 - 57	T	one_stop
58 - 70	was	58 - 60	W	one_liquid
		61 - 63	AX	one_back/mid
		64 - 70	Z	one_fricative
71 - 87	that	71 -75	DH	one_fricative
		76 - 83	AE	one_front
		84 - 87	TD	one_stop
88 - 145	silence	88 - 145	sil	sil

Table 1: Data format used for HMM categorical emotion training, original sentence: "What was that," expressed with a valence rating of one (ph. is an abbreviation for phoneme)

4. Results

4.1. Naïve Bayes classification of evaluator consistency

The subjective appraisals of valence and activation are linked to the categorical emotion label [11]. This link can be simply modeled using Naïve Bayes (NB). We used an NB classifier, implemented in PRTools [17] to predict the categorical emotion label of the clip given only the subjective valence and activation evaluations of: a) evaluator one, and b) evaluator two. In all cases, the clips are chosen from the set evaluated by both evaluators one and two and with a categorical label of angry, happy, sad, or neutral. The analysis was performed using five-fold cross-validation. The results show that evaluator one's valence and activation ratings predicted the correct categorical label 59.51% of the time while evaluator two's dimensional evaluations predicted the correct categorical label 66.80% of the time (see Table 2 for the evaluator-specific confusion matrices).

4.2. HMM classification for correspondence between content and evaluation

In this sections, models are referred to as "A – B", where "A" represents the training set and "B" represents the testing set. The accuracies of the models trained with averaged data (models "Ave - (Ave or Ind)" in Tables 3 and 4) are either better or comparable to the accuracies of the models trained with individual data (models "Ind - Ind" in Tables 3 and 4). The models trained and tested on averaged data (models "Ave - Ave" in Tables 3 and 4) had a higher accuracy than either of the individual models (models "Ind - Ind" and "Ave - Ind" in Tables 3 and 4) for both valence and activation. The classification performance of the "Ave - Ave" model improves significantly only with respect to evaluator one activation and evaluator two valence ($\alpha = 0.01$, difference of proportions). In all other conditions, the change in classification performance between the averaged and individual models is not significant ($\alpha = 0.01, 0.05$, difference of proportions).

The models trained and tested on individual data performed unequally for the valence and activation classification tasks. The evaluator one model outperformed the evaluator two model for the valence task ($\alpha = 0.01$, difference of proportions). The evaluator two model outperformed the evaluator one model for the activation task ($\alpha = 0.01$, difference of proportions). The models trained and tested on individual data did not perform significantly differently than the models trained on averaged data and tested on individual data ($\alpha = 0.01, 0.05$, difference of proportions).

(a) Confusion matrix for evaluator 1, Accuracy = 59.51%

	A	H	S	N
A	61	4	13	22
H	6	74	0	20
S	35	8	23	34
N	10	12	6	73

(b) Confusion matrix for evaluator 2, Accuracy = 66.80%

	A	H	S	N
A	81	2	1	15
H	0	56	0	44
S	23	4	39	35
N	6	3	9	82

Table 2: Confusion matrices for the categorical emotion classification task (A = angry, H = happy, S = sad, N = neutral)

Type	Evaluator	1 (%)	2 (%)	3 (%)	Total
Ind - Ind	Evaluator 1	50.00	65.90	23.71	52.18
Ind - Ind	Evaluator 2	37.47	60.28	46.65	44.33
Ave - Ave	Average	47.86	64.70	40.00	52.68
Ave - Ind	Evaluator 1	44.28	61.28	38.44	50.91
Ave - Ind	Evaluator 2	36.01	69.72	41.34	44.39

Table 3: Classification: **valence** across the three levels

5. Discussion

The NB and HMM classification indicated that the evaluation styles and strengths of the two evaluators differed across tasks. However, when the evaluations from both evaluators were combined, the HMM classification accuracies across the valence and activation classification problem either improved or did not change significantly. This suggests that models constructed from averaged evaluator data may capture the emotional acoustic properties of the utterance more closely even given different evaluation styles and internal representations of the relationship between the dimensional and categorical emotion labels.

The NB results suggest reasons for the discrepancies between the classification performance for the HMMs modeled on evaluator-specific data. The NB classification for evaluator one indicates that evaluator one's internal representation of valence is more strictly defined than that of evaluator two (Table 2). Evaluator one's confusion matrix demonstrates that based on the subjective valence and activation ratings, there exists a smaller confusion between happiness and other emotions than is observed for evaluator two. Happiness is the only emotion with positive valence and should be differentiable based on the valence rating. It should be noted, that evaluator one's confusion matrix suggests that there is an increased confusion between happiness and sadness. This may be due to a misrepresentation of activation, discussed in the following paragraph. The differences between the inter-evaluator dimensional consistency may explain why the individual-specific HMM valence model for evaluator one outperformed that of evaluator two.

The NB results show an opposite trend for the dimensional ratings of activation. These results suggest that evaluator two's dimensional rating of activation is more internally consistent when compared to that of evaluator one. For example, evaluator one's results indicate that there exists a higher level of confusion between anger and sadness than is observed in evaluator two's results. Anger and sadness are emotion classes that should be differentiable based on their activation (high vs. low, respectively). The difference in evaluator activation consistency is supported by the HMM classification accuracies. The HMM activation classification performance is higher for evaluator two than for evaluator one.

The comparisons between the NB and HMM results suggest that evaluators one and two have different evaluation styles and internal dimensional representation of emotions. However, when the ratings of these two evaluators are combined, the per-

Type	Evaluator	1 (%)	2 (%)	3 (%)	Total
Ind - Ind	Evaluator 1	64.55	23.16	66.76	47.79
Ind - Ind	Evaluator 2	68.81	39.37	62.83	55.79
Ave - Ave	Average	64.50	47.00	65.93	56.86
Ave - Ind	Evaluator 1	41.18	42.20	71.60	47.55
Ave - Ind	Evaluator 2	60.57	49.00	63.70	57.70

Table 4: Classification: **activation** across the three levels

formance of the HMM classification on valence and activation improved (significantly with respect to evaluator one activation and evaluator two valence). This suggests that even given large quantities of data, it may be more beneficial to create averaged models of dimensional evaluation, rather than evaluator-specific models (given evaluation styles that are not divergent).

The user evaluations utilized in this study were not normalized per evaluator. While the results of normalized evaluations may improve overall classification accuracies, such techniques are not necessarily representative of real-world user interactions. It is not good practice to discount the feedback of a user regarding emotion expression. It is important, from a user initiative standpoint, to work with the evaluations as provided. Furthermore, given new users in a human-computer or human-robot interaction scenario, it may not be possible to develop normalization constants in real time, necessitating the use of raw user input.

It is also important to note that the data utilized in this experiment come from unconstrained dyadic acted speech (both scripted and improvised). The utterances were not recorded on a turn-by-turn basis with rigid emotional targets. As a result, the emotional utterances in this database are not archetypal emotion expressions. Consequently, one cannot expect the classification accuracies of these more natural and subtle human emotional expressions to match those of classifications performed on read speech databases.

A weakness of this study is the small number of dimensional evaluators considered. Future work includes incorporating the evaluations of additional evaluators. The IEMOCAP database contains both audio and facial motion capture information. Future work also includes utilizing the video information to improve the accuracies and to understand the temporal interaction between the audio information, video information, and user perception.

6. Conclusion

This paper presents evidence suggesting that given different evaluation styles and different levels of evaluator consistency, averaged models of emotion perception can outperform individual models. As we move towards a society with ever increasing computing power, we will begin to see emotionally personalized technology. These systems must be able to meet both the interaction needs and expectations of the users with whom they work. This necessitates an understanding and an ability to anticipate these preferences. Initially it may seem wise to model these expectations at a per user level. However, this work suggests that the variability of individuals with respect to their dimensional appraisal may lead to inaccuracies due to user self-misrepresentation. To mitigate this problem, it may be beneficial to adapt averaged models of user perception to accommodate individual users.

Additional work is needed to determine how to integrate and interpret raw user evaluations. Researchers [10] have suggested that new evaluation metrics should be created. Emotion evaluation experimental techniques should also be updated. This may lead to the creation of new emotional ground truthing techniques that are more evaluator-intuitive and evaluator-independent.

7. Acknowledgements

This work was supported by the National Science Foundation, the US Army, and the Intel Foundation PhD Fellowship.

8. References

[1] K. Dautenhahn, A. Bond, L. Cañamero, and B. Edmonds, *Socially intelligent agents*, vol. 3, chapter 1, pp. 1–20, Springer, 2002.

[2] M. Pantic, N. Sebe, J.F. Cohn, and T. Huang, "Affective multimodal human-computer interaction," in *Proceedings of ACM Int. Conf. on Multimedia*. New York, NY, USA, 2005, pp. 669–676.

[3] C. Busso, S. Lee, and S. Narayanan, "Using neutral speech models for emotional speech analysis," in *InterSpeech ICSLP*, Antwerp, Belgium, Aug. 2007, pp. 2225–2228.

[4] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J.G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, pp. 32–80, Jan. 2001.

[5] M. Grimm, K. Kroschel, E. Mower, and S. Narayanan, "Primitives-based evaluation and estimation of emotions in speech," *Speech Communication*, vol. 49, pp. 787–800, 2007.

[6] Z. Zeng, J. Tu, BM Pianfetti, and TS Huang, "Audio–Visual Affective Expression Recognition Through Multistream Fused HMM," *IEEE Transactions on Multimedia*, vol. 10, no. 4, pp. 570–577, 2008.

[7] S. Biersack and V. Kempe, "Tracing vocal emotion expression through the speech chain: do listeners perceive what speakers feel," in *ISCA Workshop on Plasticity in Speech Perception*, London, UK, June 2005, pp. 211–214.

[8] C. Busso and S.S. Narayanan, "The expression and perception of emotions: Comparing assessments of self versus others," in *Interspeech 2008 - Eurospeech*, Brisbane, Australia, Sept. 2008, pp. 257–260.

[9] K.P. Truong, M.A. Neerincx, and D.A. van Leeuwen, "Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data," in *Interspeech 2008*, Brisbane, Australia, Sept. 2008, pp. 318–321.

[10] S. Steidl, M. Levit, A. Batliner, E. Noth, and H. Niemann, ""Of all things the measure is man": Automatic classification of emotions and inter-labeler consistency," in *ICASSP*, 2005, vol. 1, pp. 317–320.

[11] J.A. Russell and L.F. Barrett, "Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant," *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, 1999.

[12] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, pp. 335–359, Nov. 5 2008.

[13] T. Vogt and E. Andre, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in *IEEE Int. Conf. on Multimedia and Expo (ICME)*, July 2005, pp. 474–477.

[14] R. Cowie and R.R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, no. 1-2, pp. 5–32, 2003.

[15] E. Mower, M. Matarić, and S. Narayanan, "Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information," *IEEE Transactions on Multimedia*, Accepted for Publication.

[16] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK book*, Entropic Cambridge Research Laboratory Cambridge, England, 1997.

[17] R.P.W. Duin, P. Juszczak, P. Paclik, E. Pekalska, D. De Ridder, and DMJ Tax, "PRTools, a matlab toolbox for pattern recognition," *Delft University of Technology*, 2004.