Classification of Emotional Content of Sighs in Dyadic Human Interactions

Rahul Gupta, Chi-Chun Lee, Shrikanth Narayanan

Signal Analysis and Interpretation Laboratory (SAIL)
University of Southern California, Los Angeles, CA 90089, USA
guptarah@usc.edu, chiclee@usc.edu, shri@sipi.usc.edu

ABSTRACT

Emotions are an important part of human communication and are expressed both verbally and non-verbally. Common non-verbal vocalizations such as laughter, cries and sighs carry important emotional content in conversations. Sighs often are associated with negative emotion. In this work, we show that emotional sighs exist along both ends of the valence axis (*positive-emotion* vs. *negative-emotion* sighs) in spontaneous affective dialogs and that they have certain distinct multimodal characteristics. Classification results show that it is possible to differentiate between the two types of emotionally valenced sighs, using a combination of acoustic and gestural features with an overall unweighted accuracy of 58.26%.

Index terms—Nonverbal vocalizations, multimodal fusion, support vector machine

1. INTRODUCTION

Non-verbal vocalizations, such as laughter, grins, giggles, cries, sighs, etc, often occur naturally as part spontaneous conversations. These cues bear no linguistic content, and studies have shown that they are often associated strongly with the speaker's affective state [1, 2]. In fact, these nonverbal vocal cues are considered key to understanding distressed and atypical human behaviors in a variety of interaction domains such as marital therapy, and autism [3, 4]. Effectively recognizing and analyzing these vocalizations is hence essential in better understanding their roles not only in expressing an individual speaker's emotion but also in sustaining natural spoken dialogs. In the present work, we focus on analyzing one specific type of non-verbal vocalizations, *sighs*, and their significance in affective spoken interactions. In this work, sighs are defined as the audible deep intake and release of breath. Research in psychology has shown that, in general, *sighs* can be associated with different affective states (positive vs. negative), e.g., frustrated sigh and sigh of relief [5]. The objective of this work is two folds. The first goal is to analyze the occurrences of sighs in relation to the speaker emotional state. The second goal is to analyze whether multimodal features describing sighs carry predictive power in differentiating between the affective states of sighs. To predict the emotional content of sighs, we perform a classification experiment to distinguish negative from positive valenced sighs.

Various previous research works have analyzed non-verbal vocalization cues, and shown a strong relationship between these cues and the emotional states of a speaker along with many other factors such as cultures and age [1, 2]. Whereas laughter and cries are among the most researched non-verbal vocalization cues [6-9], there is only limited engineering work in analyzing the affective nature of sighs. Research in psychology [5] has documented the existence of different types of emotional sighs. In this work, we analyze this phenomenon using multimodal cues extracted from a dyadic affective interactive database, the IEMOCAP database [10]. The IEMOCAP database was used because it contains spontaneous interactions exemplifying various human emotions. The database was annotated with two types of emotional descriptions, categorical (e.g., happy, sad, neutral, frustrated, etc.) and dimensional (e.g., valence, activation, and dominance) labels.

We manually annotated and segmented occurrences of sighs in the IEMOCAP database. The emotional content of each sigh is approximated by the emotion labels associated with the closest neighboring utterances from the same speaker. This led to the annotated sighs in our data being categorized into either *positive-emotion* sigh (indicated by high valence value) or a *negative-emotion* sigh (indicated by a low valence value), supporting the observation in [5]. Analysis into the characteristics of the sigh not only suggested their occurrence with negative and positive emotions, but a considerable portion was also associated with rather ambiguous emotions such as neutrality. A support vector machine classifier was used to perform automatic classification of *positive-emotion* sighs vs. *negative-emotion* sighs using multimodal information. We found that whereas the negative sighs are better characterized by the acoustic features, the gestural features have better discriminative power for the positive class. We achieve an unweighted accuracy of 58.26% in recognizing the emotional content of sighs.

In Section 2 we describe the research methodology, in section 3 we describe and discuss the experimental results, and finally report our conclusions in Section 4.

2. RESEARCH METHODOLOGY

2.1 Database and Annotation

2.1.1 Database Description

We used the IEMOCAP database [8] for the present study. This multimodal database offers an opportunity to investigate different modalities in expressive human-human spoken dialogue interaction. It is composed of five different sessions, each involving a different pair of male and female professional actors engaged in spoken dialog interactions. The actors performed from scripted plays as well as engaged in spontaneous improvisation conversations. In addition to audio-video recordings, for each spontaneous dialogue, 61 markers (two on head, 53 on face and three on each hand) were attached to one of the interlocutors to record (x, y, z) positions of each marker (MOCAP features). Figure 1 illustrates the placement of the markers. The markers were then placed onto the other actor and recorded again with the same set of scenarios to complete the session. For our study, we consider

only the spontaneous dialogue conversations, since those interactions more closely reflect natural human behaviors in dialogs.

The data were annotated with two types of emotional descriptions over the manually segmented speaking turns, referred as "utterances" in this work. One emotional attribute type was based on categorical labels (anger, sad, happy, neutral, excited, frustrated, etc.) using at least three naïve evaluators per utterance. Majority voting was used to assign the final categorical emotion label to each utterance. The second emotional attribute was based on dimensional evaluation (valence, activation, dominance) rated on a scale of 1 to 5 with at least two naïve evaluators. The average value of the dimension attribute is used to represent the final dimensional emotion attribute for each utterance.

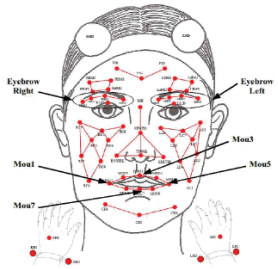

Fig 1. Motion Capture Markers Placement

2.1.2 Annotation of Sighs

The definition of sighs used in this work is a deep intake and release of breath that was audible in the audio channel. The annotation and segmentation of sighs were based on human judgment. Three different evaluators annotated sighs on non-overlapping portions of the IEMOCAP database; marking the start and end point of each sigh. The software Wavesurfer [11] was used to perform the annotation. This method of annotation requires that a sigh is *audible* using only the acoustic information. Also the sighs are discernible for individual speakers across each session in the audio stream hence annotated separately. There were a total of 502 sighs segmented; 346 of which were male sighs and 156 of which were female sighs. A table listing the number of sighs tagged per session is presented in Table 1.

Table 1. Distribution of Sighs across Sessions

Session Number	Number of Sighs Annotated	
	Male	Female
1	36	13
2	198	86
3	32	24
4	46	8
5	34	25
Total	346	156

In this work, we define the emotional content of each sigh by valence attribute of the closest utterance (or valence of the corresponding speech utterance if the sigh happened within a speech turn) of the same speaker. Figure 2 shows a histogram distribution of the valence values of all annotated sighs. The figure shows a clear bi-modal distribution of valence values for sighs. This support the claim that there are two different types of *emotional* sighs (positive-emotional sighs and negative-emotional sighs).

For the present work, we focus our analysis on these two types of emotional sighs: *positive-emotional* sighs and *negative emotional* sighs. *Positive-emotional* sighs are defined as sighs that have an associated valence dimension rating greater than 3, whereas *negative-emotional* sighs have a rating less than or equal to 3. As can be seen (Figure 2), there are more sighs associated with negative emotions. There are a total of 388 *negative-emotion* sighs and 114 *positive-emotion* sighs. Since we also use the MOCAP (visual gesture) features, and only one of the speakers is captured here, we have a subset of 252 negative-emotion sighs and 62 positive emotion sighs, totaling 314, with both acoustic and MOCAP information.

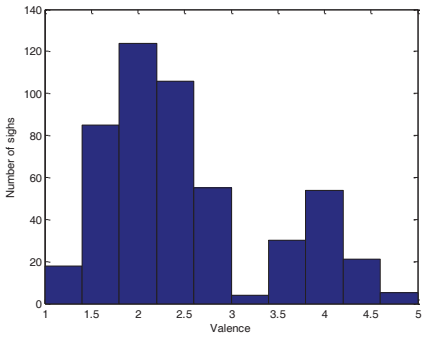

Fig 2. Valence distributions of sighs

2.2 Acoustic and MOCAP Feature Extraction

We extracted 13 Mel Frequency Cepstral coefficients. They have shown to be effective in emotion recognition tasks [12]. Using Praat [13], the features were extracted every 10ms over the all of the speakers' utterances. The mean and variance of each of the MFCCs were computed over the period of the sigh, resulting in a total of 26-dimensional features vector used for classification.

For the MOCAP features we calculated the mean and variance of the velocities of each of the markers over the sigh period. Speaker-wise normalization was performed on the velocity values using mean subtraction for each speaker in the database. The mean is the average velocity of each marker over the whole session. Since we had 63 markers in all, this gave as a 126 dimensional feature vector. Normalization was done to eliminate the effect of individual speaker characteristics. Since this feature dimension is too high given the number of data samples we reduce the dimensionality using principal component analysis.

3. EXPERIMENT RESULTS AND DISCUSSION

3.1 Emotional Content in sighs

In this section we analyze the categorical emotion labels associated with these two types of sighs. Table 2 and Table 3 show a breakup of the number of different categorical emotion labels for each type of sigh (Table 2 corresponds to female speakers and Table 3 corresponds to male speakers). *Unlabeled* corresponds to the utterances where there is no majority agreement in the emotion evaluation of the original IEMOCAP data.

The first observation is that for *negative-emotion* sighs, the dominant emotion class labels are *frustrated* and *sad*. The emotion label of *frustration* has been shown to be relatively

difficult to recognize compared with basic categorical emotion classes (happy, sad, and angry). This observation shows that there can be a potential application by utilizing the emotional type of sighs to improve the detection of *frustration*.

Table 2. Emotional Tags for Female Sighs

Emotion	*Negative Sighs*		*Positive Sighs*	
	Number	Percentage	Number	Percentage
Frustrated	139	27.68%	8	1.59%
Happy	0	0%	2.78%	10%
Sad	145	28.88%	6	1.20%
Excited	0	0%	39	7.77%
Neutral	44	8.76%	25	4.98%
Angry	14	2.78%	0	0%
Unlabeled	46	9.16%	22	4.38%
Total	388	77.29%	114	22.70%

Also there are a substantial number of emotional utterances across both types of sighs that have an emotion label of *neutral*. *Neutral* can often be viewed as an ambiguous emotion class (or a lack of true emotion) because of the fuzziness in the emotion definition. Therefore, *neutral* is often associated with having valence and activation close to a rating of 3 (on a typical scale from 1–5). It is interesting to observe that while a sigh is often perceived as a strong emotion marker (as shown in a clear bimodal distribution of valence in sighs), approximately 13.74% of sighs have associated categorical emotional labels of *neutral.* This suggests that any further analysis and characterization of the neutral "category" should also consider nonverbal cues in addition to the verbal ones.

The third observation is that the *unlabeled* sigh shows a similar trend to the *neutral* emotional class; it represents a substantial percentage of the sighs particularly in the case of *positive-emotion* sighs. While the majority of the *positive-emotion* sighs are labeled as excited, an almost equal percentage of them are associated with either *neutral* or *unlabeled*. One implication is that when a sigh is associated with positive emotion, it may be difficult to assign a canonical categorical label to describe it, e.g., sighs of relief. This observation shows that by simply stating the fact that there is an occurrence of sigh can be beneficial in providing a description of a subtle (non-prototypical) emotion class.

3.2 Classification on the Emotional Types of Sighs

3.2.1 Acoustic feature classifier

In this section, we describe an experiment to examine the predictive power of acoustic cues in recognizing the two different emotional types of sighs (*positive-emotion* sigh vs. *negative-emotion* sighs) using a support vector machine classifier (SVM) with acoustic cues. For this experiment we take the entire set of sighs (as we are not using the MOCAP data for this part of the experiment).

In the experiment, we use leave-one-speaker-out cross validation. Since the class label distribution is heavily biased toward *negative-emotion* sighs, the metric used in this classification task was the unweighted accuracy (average accuracy of each class label). A linear kernel was chosen for the SVM because it outperformed other kernels (quadratic and radial basis functions) through our empirical experiments.

When we consider the class-wise accuracy, there is a dramatic difference in performance between the positive and negatively valenced cases (Table 4). One of the possible reasons could be the difference in the sizes of the datasets for the two classes, with fewer instances for the positive class. To explore whether this can be improved with addition of visual features, we considered incorporating the available MOCAP features.

Table 3. Summary of SVM Classification

Model	Unweighted Accuracy (%)
Chance	50
SVM	**60.16**

Table 4. Class-wise Accuracy Summary

Experiment	*Positive*	*Negative*
Using only the MFCC derived features on the entire set of sighs	28.95	86.08

3.2.2 Multimodal classifier

For this experiment we used a subset of sighs where both the modalities were available, numbering 314. We performed a similar classification on the subset of sighs using only acoustic features, and then only the MOCAP features. Finally the two classifiers were combined at the decision level to improve the overall accuracy of the classification. These results are not directly comparable to the previous classification as we are using only a subset of the database now.

We performed principal component analysis on the MOCAP features as the dimensionality of the features vector is high, given the number of data samples. In order to find the optimal number of principal components, we divided the dataset into a training dataset composed of 8 speakers, 1 speaker for development set and 1 for testing. We noted the unweighted accuracy over the development set as the number of principal components was increased from 5 to 75 in steps of 5. We achieved maxima at 55 principal components. We use the same number of principal components for the test set. Our analysis on the reduced dimensionality reflects that we are able to preserve almost all the variance in the dataset, while reducing the dimensionality by more than half. In order to keep the experiments consistent we chose similar division of dataset on the acoustic features.

Finally we make the combined decisions based on the distances of the data-points from the decision hyper-plane in the two SVM classifiers. When there was a conflict in the decisions of the two classifiers, we compared the weighted distances of the data-points from the decision hyper-plane. A test on the development set suggested an optimum scaling parameter, α = .06 on the distances of the acoustic feature based classifier as compared to the MOCAP feature based classifier. This was achieved after a grid test on the weights by scaling the distances in the range of [.01 to .99] and then [1 to 100] in the steps of .01 and 1 respectively. This decision rule is represented in equation 1. If the scaled distance of one classifier was more than the other, we assigned the sigh to be of that class. In case there was same decision from both the classifiers that itself was taken as the final decision.

If $\text{class}_{\text{acoustic}} \neq \text{class}_{\text{mocap}}$ and

$$\alpha(w_{acoustic}.x_{acoustic} - b_{acoustic}) > (w_{mocap}.x_{mocap} - b_{mocap})$$

then class = $\text{class}_{\text{acoustic}}$

else class = $\text{class}_{\text{mocap}}$ (1)

Here $w_{acoustic}$ and w_{mocap} represent the weight vectors,

$x_{acoustic}$ and x_{mocap} the data points and b the bias terms for the individual classifiers. A point to note here is that the parameter α does not represent the actual weightage given to the individual features, as it scales the distance from two SVMs, which in themselves classify data-points having different dimensionality.

Table 6. Summary of SVM Classification

Model	Unweighted Accuracy (%)
Chance	50
SVM using acoustic features	**55.74**
SVM using MOCAP features	**57.66**
Fusion	**58.26**

Table 7. Class-wise Accuracy Summary

Experiment	*Positive*	*Negative*
SVM using acoustic features	24.19	87.30
SVM using MOCAP features	40.32	75.00
Fusion	40.32	76.19

3.2.3 Results of Classification Experiment

A summary of the classification accuracy is presented in Table 6 and 7. From the recognition rate, one can notice that the acoustic features are good at recognizing the negative class of sighs whereas the MOCAP features show a slightly better performance for the positive class. While further analysis is necessary to substantiate this observation further, we hypothesize that this may have been due to the fact that *positive-emotion* sighs have more reflection in the bodily expressions of the person and the acoustic cues carry more relevant information for the *negative-emotion* sighs.

We observe lower accuracies in this experiment as compared to the very first experiment as we are using a subset of database. But in order to make a more informed comparison between the role of MOCAP features and acoustic features it was important that we used the same dataset. The class accuracy for the positive class is lower in comparison to the negative class possibly because the positive sighs are fewer in number as compared to the negative sighs. Another drawback of the database is the uneven distribution in the number of sighs across each individual speaker. This sometimes leads to small number of training instances and a relatively larger testing set. This affects the accuracy of the classifier in that particular fold for cross-validation.

Despite these drawbacks we observe that there is an increase in the overall unweighted accuracy using the MOCAP features. The MOCAP features are substantially more efficient in capturing the positive sighs. We observed that during fusion, the algorithm chose the MOCAP classifier decision more often than the acoustic classifier. Only when the distance of a data sample in the acoustic feature space was very far away from the decision boundary, indicating more confidence, the fusion mechanism chose the acoustic classifier over the MOCAP classifier.

4. CONCLUSION AND FUTURE WORK

Nonverbal vocalization is a natural and integral part of human communication. It often carries strong emotional information. In this work, we analyzed a specific nonverbal vocalization, the *sigh*, using an emotionally-rich database of spontaneous dyadic interaction. The results underscore the importance of the emotional interpretation of sighs and suggest the feasibility of using low level acoustic cues as well as body movement to predict the different emotional content of each sigh.

One of the first critical steps that we are planning to do is to acquire more instances of sighs across multiple natural conversational databases. Also, since there exists such a strong interpretation of non-verbal vocalizations in understanding the specific affective state of the speaker, we need to incorporate this information to enhance the capabilities of automatic emotion recognizers. This may be especially useful for detecting hard-to-recognize or subtle ambiguous states, such as frustration and neutrality. Another interesting observation that we have found in this database is that there is a positive correlation between the numbers of occurrences of sighs between the dyad across sessions which suggests a phenomenon that interacting dyads often exhibit such nonverbal vocalizations together as they interact. With improved insights into the role of nonverbal vocalizations in human communication, we can contribute to the design of a better and natural dialog interface.

5. REFERENCES

[1] D. Sauter, F. Eisner, P. Ekman, and S. K. Scott, "Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations" in *Proceedings of the National Academy of Sciences*, 107(6), 2408-2412, 2010

[2] D. Sauter, , F.Eisner, , A. J.Calder, , and S. K. Scott, "Perceptual cues in nonverbal vocal expressions of emotion" *Quarterly Journal of Experimental Psychology*, 63(11), 2251-2272, 2010

[3] V. Reddy, E. Williams and A.Vaughan, "Sharing humour and laughter in autism and Down's syndrome" British Journal of Psychology, 93: 219–242

[4] J. L. Driver and J. M. Gottman, "Daily marital interactions and positive affect during marital conflict among newlywed couples" Family Process, 43: 301–314.

[5] K. Teigen, "Is a sigh "just a sigh"? Sighs as emotional signals and responses to a difficult task", *Scandinavian Journal of Psychology*, Volume 49, Number 1, February 2008 , pp. 49-57(9)

[6] M.Miranda, J.A.Alonzo, J.Campita, S. Lucila, M.Suarez, "Discovering emotions in Filipino laughter using audio features", *Human-Centric Computing (HumanCom),* 2010 3rd International Conference

[7] K. Kikuchi, K. Arakawa, "Estimation of babies' emotion by frequency analyses of their cries", IEEE-EURASIP *Nonlinear Signal and Image Processing,* NSIP 2005.

[8] K. Laskowski, "Finding emotionally involved speech using implicitly proximity-annotated laughter", *Acoustics Speech and Signal Processing (ICASSP)*, 2010 IEEE International Conference

[9] S. Sundaram and S. Narayanan, "Automatic acoustic synthesis of human-like laughter." *J. Acoust. Soc. Am*, vol. 121, no. 1. pp. 527–535, Jan 2007

[10] **C**. Busso, M. Bulut, C.-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*, 42:4, pp. 335-359, 2008.

[11] K. Sjölander and J. Beskow, "Wavesurfer – an open source speech tool", Centre for Speech Technology, KTH, Drottning Kristinas väg 31, SE-100 44, Sweden

[12] 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, 2007.

[13] P. Boersma, Praat, "A system for doing phonetics by computer", *Glot International* 5:9/10, 341-345, 2001

[14] C. Cortes and V. Vapnik, *Support-Vector Networks*, Machine Learning, 20(3):273-297, September 1995