# Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information

Carlos Busso, Zhigang Deng*, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann*, Shrikanth Narayanan

Emotion Research Group, Speech Analysis and Interpretation Lab
Integrated Media Systems Center, Department of Electrical Engineering, *Department of Computer Science
Viterbi School of Engineering, University of Southern California, Los Angeles
http://sail.usc.edu

## ABSTRACT

The interaction between human beings and computers will be more natural if computers are able to perceive and respond to human non-verbal communication such as emotions. Although several approaches have been proposed to recognize human emotions based on facial expressions or speech, relatively limited work has been done to fuse these two, and other, modalities to improve the accuracy and robustness of the emotion recognition system. This paper analyzes the strengths and the limitations of systems based only on facial expressions or acoustic information. It also discusses two approaches used to fuse these two modalities: decision level and feature level integration. Using a database recorded from an actress, four emotions were classified: sadness, anger, happiness, and neutral state. By the use of markers on her face, detailed facial motions were captured with motion capture, in conjunction with simultaneous speech recordings. The results reveal that the system based on facial expression gave better performance than the system based on just acoustic information for the emotions considered. Results also show the complementarily of the two modalities and that when these two modalities are fused, the performance and the robustness of the emotion recognition system improve measurably.

## Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *interaction styles, Auditory (non-speech) feedback.*

## General Terms

Performance, Experimentation, Design, Human Factors

## Keywords

Emotion recognition, speech, vision, PCA, SVC, decision level fusion, feature level fusion, affective states, human-computer interaction (HCI).

## 1. INTRODUCTION

Inter-personal human communication includes not only spoken language but also non-verbal cues such as hand gestures, facial expressions and tone of the voice, which are used to express feeling and give feedback. However, the new trends in human-computer interfaces, which have evolved from conventional mouse and keyboard to automatic speech recognition systems and special interfaces designed for handicapped people, do not take complete advantage of these valuable communicative abilities, resulting often in a less than natural interaction. If computers could recognize these emotional inputs, they could give specific and appropriate help to users in ways that are more in tune with the user's needs and preferences.

It is widely accepted from psychological theory that human emotions can be classified into six archetypal emotions: surprise, fear, disgust, anger, happiness, and sadness. Facial motion and the tone of the speech play a major role in expressing these emotions. The muscles of the face can be changed and the tone and the energy in the production of the speech can be intentionally modified to communicate different feelings. Human beings can recognize these signals even if they are subtly displayed, by simultaneously processing information acquired by ears and eyes. Based on psychological studies, which show that visual information modifies the perception of speech [17], it is possible to assume that human emotion perception follows a similar trend. Motivated by these clues, De Silva et al. conducted experiments, in which 18 people were required to recognize emotion using visual and acoustic information separately from an audio-visual database recorded from two subjects [7]. They concluded that some emotions are better identified with audio such as sadness and fear, and others with video, such as anger and happiness. Moreover, Chen et al. showed that these two modalities give complementary information, by arguing that the performance of the system increased when both modalities were considered together [4]. Although several automatic emotion recognition systems have explored the use of either facial expressions [1],[11],[16],[21],[22] or speech [9],[18],[14] to detect human affective states, relatively few efforts have focused on emotion recognition using both modalities [4],[8]. It is hoped that the multimodal approach may give not only better performance, but also more robustness when one of these modalities is acquired in a noisy environment [19]. These previous studies fused facial expressions and acoustic information either at a decision-level, in which the outputs of the unimodal systems are integrated by the use of suitable criteria, or at a feature-level, in which the data from both modalities are combined before classification. However, none of these papers attempted to compare which fusion approach is more suitable for emotion recognition. This paper evaluates these two fusion approaches, in terms of the performance of the overall system.


*ICMI'04*, October 13–15, 2004, State College, Pennsylvania, USA.

This paper analyzes the use of audio-visual information to recognize four different human emotions: sadness, happiness, anger and neutral state, using a database recorded from a actress with markers attached to her face to capture visual information (the more challenging task of capturing salient visual information directly from conventional videos is a topic for future work but is hoped to be informed by studies such as in this report). The primary purpose of this research is to identify the advantages and limitations of unimodal systems, and to show which fusion approaches are more suitable for emotion recognition.

# 2. EMOTION RECOGNITION SYSTEMS

## 2.1 Emotion recognition by speech

Several approaches to recognize emotions from speech have been reported. A comprehensive review of these approaches can be found in [6] and [19]. Most researchers have used global suprasegmental/prosodic features as their acoustic cues for emotion recognition, in which utterance-level statistics are calculated. For example, mean, standard deviation, maximum, and minimum of pitch contour and energy in the utterances are widely used features in this regard. Dellaert et al. attempted to classify 4 human emotions by the use of pitch-related features [9]. They implemented three different classifiers: Maximum Likelihood Bayes classifier (MLB), Kernel Regression (KR), and K-nearest Neighbors (KNN). Roy and Pentland classified emotions using a Fisher linear classifier [20]. Using short-spoken sentences, they recognized two kinds of emotions: approval or disapproval. They conducted several experiments with features extracted from measures of pitch and energy, obtaining an accuracy ranging from 65% to 88%.

The main limitation of those global-level acoustic features is that they cannot describe the dynamic variation along an utterance. To address this, for example, dynamic variation in emotion in speech can be traced in spectral changes at a local segmental level, using short-term spectral features. In [14], 13 Mel-frequency cepstral coefficients (MFCC) were used to train a Hidden Markov Model (HMM) to recognize four emotions. Nwe et al. used 12 Mel-based speech signal power coefficients to train a Discrete Hidden Markov Model to classify the six archetypal emotions [18]. The average accuracy in both approaches was between 70 and 75%. Finally, other approaches have used language and discourse information, exploring the fact that some words are highly correlated with specific emotions [15].

In this study, prosodic information is used as acoustic features as well as the duration of voiced and unvoiced segments.

## 2.2 Emotion recognition by facial expressions

Facial expressions give important clues about emotions. Therefore, several approaches have been proposed to classify human affective states. The features used are typically based on local spatial position or displacement of specific points and regions of the face, unlike the approaches based on audio, which use global statistics of the acoustic features. For a complete review of recent emotion recognition systems based on facial expression the readers are referred to [19].

Mase proposed an emotion recognition system that uses the major directions of specific facial muscles [16]. With 11 windows manually located in the face, the muscle movements were extracted by the use of optical flow. For classification, K-nearest neighbor rule was used, with an accuracy of 80% with four emotions: happiness, anger, disgust and surprise. Yacoob et al. proposed a similar method [22]. Instead of using facial muscle actions, they built a dictionary to convert motions associated with edge of the mouth, eyes and eyebrows, into a linguistic, per-frame, mid-level representation. They classified the six basic emotions by the used of a rule-based system with 88% of accuracy.

Black et al. used parametric models to extract the shape and movements of the mouse, eye and eyebrows [1]. They also built a mid- and high-level representation of facial actions by using a similar approach employed in [22], with 89% of accuracy. Tian et al. attempted to recognize Actions Units (AU), developed by Ekman and Friesen in 1978 [10], using permanent and transient facial features such as lip, nasolabial furrow and wrinkles [21]. Geometrical models were used to locate the shapes and appearances of these features. They achieved a 96% of accuracy. Essa et al. developed a system that quantified facial movements based on parametric models of independent facial muscle groups [11]. They modeled the face by the use of an optical flow method coupled with geometric, physical and motion-based dynamic models. They generated spatial-temporal templates that were used for emotion recognition. Without considering sadness that was not included in their work, a recognition accuracy rate of 98% was achieved.

In this study, the extraction of facial features is done by the use of markers. Therefore, face detection and tracking algorithms are not needed.

## 2.3 Emotion recognition by bimodal data

Relatively few efforts have focused on implementing emotion recognition systems using both facial expressions and acoustic information. De Silva et al. proposed a rule-based audio-visual emotion recognition system, in which the outputs of the uni-modal classifiers are fused at the decision-level [8]. From audio, they used prosodic features, and from video, they used the maximum distances and velocities between six specific facial points. A similar approach was also presented by Chen et al. [4], in which the dominant modality, according to the subjective experiments conducted in [7], was used to resolve discrepancies between the outputs of mono-modal systems. In both studies, they concluded that the performance of the system increased when both modalities were used together.

Yoshitomi et al. proposed a multimodal system that not only considers speech and visual information, but also the thermal distribution acquired by infrared camera [24]. They argue that infrared images are not sensitive to lighting conditions, which is one of the main problems when the facial expressions are acquired with conventional cameras. They used a database recorded from a female speaker that read a single word acted in five emotional states. They integrated these three modalities at decision-level using empirically determined weights. The performance of the system was better when three modalities were used together.

In [12] and [5], a bimodal emotion recognition system was proposed to recognize six emotions, in which the audio-visual data was fused at feature-level. They used prosodic features from audio, and the position and movement of facial organs from

video. The best features from both unimodal systems were used as input in the bimodal classifier. They showed that the performance significantly increased from 69.4% (video system) and 75% (audio system) to 97.2% (bimodal system). However they use a small database with only six clips per emotion, so the generalizability and robustness of the results should be tested with a larger data set.

All these studies have shown that the performance of emotion recognition systems can be improved by the use of multimodal information. However, it is not clear which is the most suitable technique to fuse these modalities. This paper addresses this open question, by comparing decision and features level integration techniques in term of the performance of the system.

# 3. METHODOLOGY

Four emotions -- sadness, happiness, anger and neutral state --are recognized by the use of three different systems based on audio, facial expression and bimodal information, respectively. The main purpose is to quantify the performance of unimodal systems, recognize the strengths and weaknesses of these approaches and compare different approaches to fuse these dissimilar modalities to increase the overall recognition rate of the system.

The database used in the experiments was recorded from an actress who read 258 sentences expressing the emotions. A VICON motion capture system with three cameras (left of Figure 1) was used to capture the expressive facial motion data with 120Hz sampling frequency. With 102 markers on her face (right of Figure 1), an actress was asked to speak a custom phoneme-balanced corpus four times, with different emotions. The recording was made in a quiet room using a close talking SHURE microphone at the sampling rate of 48 kHz. The markers' motion and aligned audio were captured by the system simultaneously. Notice that the facial features are extracted with high precision, so this multimodal database is suitable to extract important clues about both facial expressions and speech.

**Figure 1: Data recording system**

In order to compare the unimodal systems with the multimodal system, three different approaches were implemented all using support vector machine classifier (SVC) with 2nd order polynomial kernel functions [3]. SVC was used for emotion recognition in our previous study, showing better performance than other statistical classifiers [13][14]. Notice that the difference between the three approaches is in the features used as inputs, so it is possible to conclude the strengths and limitations of acoustic and facial expressions features to recognize human emotions. In all the three systems, the database was trained and tested using the leave-one-out cross validation method.

## 3.1 System based on speech

The most widely used speech cues for audio emotion recognition are global-level prosodic features such as the statistics of the pitch and the intensity. Therefore, the means, the standard deviations, the ranges, the maximum values, the minimum values and the medians of the pitch and the energy were computed using Praat speech processing software [2]. In addition, the voiced/speech and unvoiced/speech ratio were also estimated. By the use of sequential backward features selection technique, a 11-dimensional feature vector for each utterance was used as input in the audio emotion recognition system.

## 3.2 System based on facial expressions

In the system based on visual information, which is described in figure 4, the spatial data collected from markers in each frame of the video is reduced into a 4-dimensional feature vector per sentence, which is then used as input to the classifier. The facial expression system, which is shown in figure 4, is described below.

After the motion data are captured, the data are normalized: (1) all markers are translated in order to make a nose marker be the local coordinate center of each frame, (2) one frame with neutral and close-mouth head pose is picked as the reference frame, (3) three approximately rigid markers (manually chosen and illustrated as blue points in Figure 1) define a local coordinate origin for each frame, and (4) each frame is rotated to align it with the reference frame. Each data frame is divided into five blocks: forehead, eyebrow, low eye, right cheek and left cheek area (see Figure 2). For each block, the 3D coordinate of markers in this block is concatenated together to form a data vector. Then, Principal Component Analysis (PCA) method is used to reduce the number of features per frame into a 10-dimensional vector for each area, covering more than 99% of the variation. Notice that the markers near the lips are not considered, because the articulation of the speech might be recognized as a smile, confusing the emotion recognition system [19].

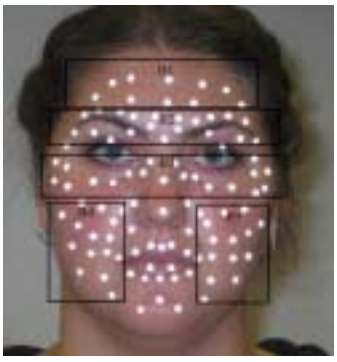

**Figure 2: five areas of the face considered in this study**

In order to visualize how well these feature vectors represent the emotion classes, the first two components of the low eye area vector were plotted in figure 3. As can be seen, different emotions appear in separate clusters, so important clues can be extracted from the spatial position of these 10-dimensional features space.

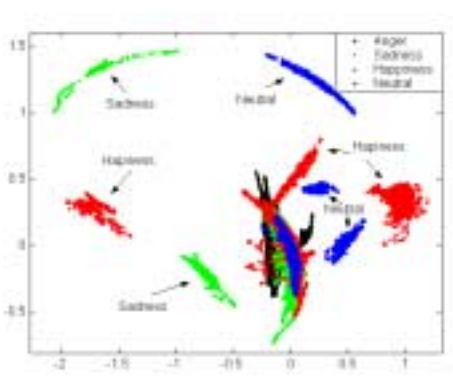

**Figure 3: First two components of low eye area vector**

Notice that for each frame, a 10-dimensional feature vector is obtained in each block. This local information might be used to train dynamic models such as HMM. However, in this paper we decided to use global features at utterance level for both unimodal systems, so these feature vectors were preprocessed to obtain a low dimensional feature vector per utterance. In each of the 5 blocks, the 10-dimensional features at frame level were classified using a K-nearest neighbor classifier (k=3), exploiting the fact that different emotions appear in separate clusters (Figure 3). Then, the number of frames that were classified for each emotion was counted, obtaining a 4-dimentional vector at utterance level, for each block. These feature vectors at utterance level take advantage not only of the spatial position of facial points, but also of global patterns shown when emotions are displayed. For example, when happiness is displayed in more than 90 percent of the frames, they are classified as happy, but when sadness is displayed even more than 50 percent of the frames, they are classified as sad. The SVC classifiers use this kind of information, improving significantly the performance of the system. Also, with this approach the facial expression features and the global acoustic features do not need to be synchronized, so they can be easily combined in a feature-level fusion.

As described in figure 4, a separate SVC classifier was implemented for each block, so it is possible to infer which facial area gives better emotion discrimination. In addition, the 4-dimensional features vectors of the 5 blocks were added before classification, as shown in figure 4. This system is referred as the combined facial expressions classifier.

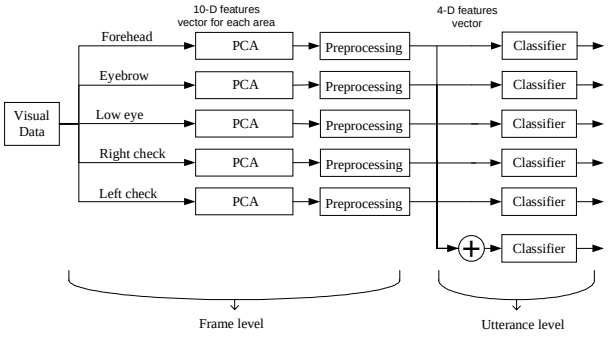


**Figure 4: System based on facial expression**

## 3.3 Bimodal system

To fuse the facial expression and acoustic information, two different approaches were implemented: feature-level fusion, in which a single classifier with features of both modalities are used (left of Figure 5); and, decision level fusion, in which a separate classifier is used for each modality, and the outputs are combined using some criteria (right of Figure 5). In the first approach, a sequential backward feature selection technique was used to find the features from both modalities that maximize the performance of the classifier. The number of features selected was 10. In the second approach, several criteria were used to combine the posterior probabilities of the mono-modal systems at the decision-level: maximum, in which the emotion with greatest posterior probability in both modalities is selected; average, in which the posterior probabilities of each modalities are equally weighted and the maximum is selected; product, in which the posterior probabilities are multiplied and the maximum is selected; and, weight, in which different weights are applied to the different unimodal systems.

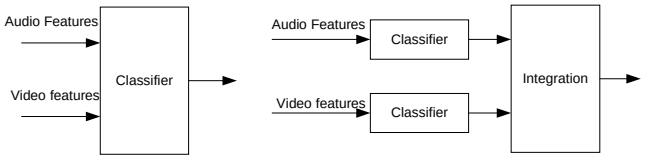


**Figure 5: Features-level and decision-level fusion**

# 4. RESULTS

## 4.1 Acoustic emotion classifier

Table 1 shows the confusion matrix of the emotion recognition system based on acoustic information, which gives details of the strengths and weaknesses of this system. The overall performance of this classifier was 70.9 percent. The diagonal components of table 1 reveal that all the emotions can be recognized with more than 64 percent of accuracy, by using only the features of the speech. However, table 1 shows that some pairs of emotions are usually confused more. Sadness is misclassified as neutral state (22%) and vise versa (14%). The same trend appears between happiness and anger, which are mutually confused (19% and 21%, respectively). These results agree with the human evaluations done by De Silva et al. [7], and can be explained by similarity patterns observed in acoustic parameters of these emotions [23]. For example, speech associated with anger and happiness is characterized by longer utterance duration, shorter inter-word silence, higher pitch and energy values with wider ranges. On the other hand, in neutral and sad sentences, the energy and the pitch are usually maintained at the same level. Therefore, these emotions are difficult to be classified.

**Table 1: Confusion matrix of the emotion recognition system based on audio**

| | Anger | Sadness | Happiness | Neutral |
|---|---|---|---|---|
| Anger | 0.68 | 0.05 | 0.21 | 0.05 |
| Sadness | 0.07 | 0.64 | 0.06 | 0.22 |
| Happiness | 0.19 | 0.04 | 0.70 | 0.08 |
| Neutral | 0.04 | 0.14 | 0.01 | 0.81 |

## 4.2 System based on facial expressions

Table 3 shows the performance of the emotion recognition systems based on facial expressions, for each of the five facial blocks and the combined facial expression classifier. This table reveals that the cheek areas give valuable information for emotion classification. It also shows that the eyebrows, which have been widely used in facial expression recognition, give the poorest performance. The fact that happiness is classified without any mistake can be explained by the figure 3, which shows that happiness is separately clustered in the 10-dimentional PCA spaces, so it is easily to recognize. Table 2 also reveals that the combined facial expression classifier has an accuracy of 85%, which is higher than most of the 5 facial blocks classifiers. Notice that this database was recorded from a *single* actress, so clearly more experiments should be conducted to evaluate these results with other subjects.

**Table 2: Performance of the facial expression classifiers**

| Area | Overall | Anger | Sadness | Hapiness | Neutral |
|---|---|---|---|---|---|
| Forehead | 0.73 | 0.82 | 0.66 | 1.00 | 0.46 |
| Eyebrow | 0.68 | 0.55 | 0.67 | 1.00 | 0.49 |
| Low eye | 0.81 | 0.82 | 0.78 | 1.00 | 0.65 |
| Right cheek | 0.85 | 0.87 | 0.76 | 1.00 | 0.79 |
| Left cheek | 0.80 | 0.84 | 0.67 | 1.00 | 0.67 |
| Combined classifier | 0.85 | 0.79 | 0.81 | 1.00 | 0.81 |

The combined facial expression classifier can be seen as a feature-level integration approach in which the features of the 5 blocks are fused before classification. These classifiers can be also integrated at decision-level. Table 3 shows the performance of the system when the facial block classifiers are fused by the use of different criteria. In general, the results are very similar. All these decision-level rules give slightly worse performance than the combined facial expression classifier.

**Table 3: Decision-level integration of the 5 facial blocks emotion classifiers**

| | Overall | Anger | Sadness | Hapiness | Neutral |
|---|---|---|---|---|---|
| Mayority voting | 0.82 | 0.92 | 0.72 | 1.00 | 0.65 |
| Maximum | 0.84 | 0.87 | 0.73 | 1.00 | 0.75 |
| Averaging combining | 0.83 | 0.89 | 0.72 | 1.00 | 0.70 |
| Product combining | 0.84 | 0.87 | 0.72 | 1.00 | 0.77 |

Table 4 shows the confusion matrix of the combined facial expression classifier to analyze in detail the limitation of this emotion recognition system. The overall performance of this classifier was 85.1 percent. This table reveals that happiness is recognized with very high accuracy. The other three emotions are classified with 80 percent of accuracy, approximately. Table 4 also shows that in the facial expressions domain, anger is confused with sadness (18%) and neutral state is confused with happiness (15%). Notice that in the acoustic domain, sadness/anger and neutral /happiness can be separated with high accuracy, so it is expected that the bimodal classifier will give good performance for anger and neutral state. This table also shows that sadness is confused with neutral state (13%). Unfortunately, these two emotions are also confused in the acoustic domain (22%), so it is expected that the recognition rate of sadness in the bimodal classifiers will be poor. Other discriminating information such as contextual cues are needed.

**Table 4: Confusion matrix of the combined facial expression classifier**

| | Anger | Sadness | Happiness | Neutral |
|---|---|---|---|---|
| Anger | 0.79 | 0.18 | 0.00 | 0.03 |
| Sadness | 0.06 | 0.81 | 0.00 | 0.13 |
| Happiness | 0.00 | 0.00 | 1.00 | 0.00 |
| Neutral | 0.00 | 0.04 | 0.15 | 0.81 |

## 4.3 Bimodal system

Table 5 displays the confusion matrix of the bimodal system when the facial expressions and acoustic information were fused at feature-level. The overall performance of this classifier was 89.1 percent. As can be observed, anger, happiness and neutral state are recognized with more than 90 percent of accuracy. As it was expected, the recognition rate of anger and neutral state was higher than unimodal systems. Sadness is the emotion with lower performance, which agrees with our previous analysis. This emotion is confused with neutral state (18%), because none of the modalities we considered can accurately separate these classes. Notice that the performance of happiness significantly decreased to 91 percent.

**Table 5: Confusion matrix of the feature-level integration bimodal classifier**

| | Anger | Sadness | Happiness | Neutral |
|---|---|---|---|---|
| Anger | 0.95 | 0.00 | 0.03 | 0.03 |
| Sadness | 0.00 | 0.79 | 0.03 | 0.18 |
| Happiness | 0.02 | 0.00 | 0.91 | 0.08 |
| Neutral | 0.01 | 0.05 | 0.02 | 0.92 |

Table 6 shows the performance of the bimodal system when the acoustic emotion classifier (Table 1) and the combined facial expressions classifier (Table 4) were integrated at decision-level, using different fusing criteria. In the weight-combining rule, the modalities are weighted according to rules extracted from the confusion matrices of each classifier. This table reveals that the maximum combining rule gives similar results compared to the facial expression classifier. This result suggests that the posterior probabilities of the acoustic classifier are smaller than the posterior probabilities of the facial expression classifier. Therefore, this rule is not suitable for fusing these modalities, because one modality will be effectively ignored. Table 6 also shows that the product-combining rule gives the best performance.

**Table 6: Decision-level integration bimodal classifier with different fusing criteria**

| | Overall | Anger | Sadness | Hapiness | Neutral |
|---|---|---|---|---|---|
| Maximum combining | 0.84 | 0.82 | 0.81 | 0.92 | 0.81 |
| Averaging combining | 0.88 | 0.84 | 0.84 | 1.00 | 0.84 |
| Product combining | 0.89 | 0.84 | 0.90 | 0.98 | 0.84 |
| Weight combining | 0.86 | 0.89 | 0.75 | 1.00 | 0.81 |

Table 7 shows the confusion matrix of the decision-level bimodal classifier when the product-combining criterion was used. The overall performance of this classifier was 89.0 percent, which is very close to the overall performance achieved by the feature-level bimodal classifier (Table 5). However, the confusion matrices of both classifiers show important differences. Table 7 shows that in this classifier, the recognition rate of anger (84%) and neutral

states (84%) are slightly better than in the facial expression classifier (79% and 81%, Table 4), and significantly worse than in the feature-level bimodal classifier (95%, 92%, Table 5). However, happiness (98%) and sadness (90%) are recognized with high accuracy compared to the feature-level bimodal classifier (91% and 79%, Table 5). These results suggest that in the decision-level fusion approach, the recognition rate of each emotion is increased, improving the performance of the bimodal system.

**Table 7: Confusion matrix of the decision-level bimodal classifier with product-combining rule**

| | Anger | Sadness | Happiness | Neutral |
|---|---|---|---|---|
| Anger | 0.84 | 0.08 | 0.00 | 0.08 |
| Sadness | 0.00 | 0.90 | 0.00 | 0.10 |
| Happiness | 0.00 | 0.00 | 0.98 | 0.02 |
| Neutral | 0.00 | 0.02 | 0.14 | 0.84 |

# 5. DISCUSSION

Humans use more than one modality to recognize emotions, so it is expected that the performance of automatic multimodal systems will be higher than automatic unimodal systems. The results reported in this work confirm this hypothesis, since the bimodal approach gave an improvement of almost 5 percent (absolute) compared to the performance of the facial expression recognition system. The results show that pairs of emotions that were confused in one modality were easily classified in the other. For example, anger and happiness that were usually misclassified in the acoustic domain were separated with greater accuracy in the facial expression emotion classifier. Therefore, when these two modalities were fused at feature-level, these emotions were classified with high precision. Unfortunately, sadness is confused with neutral state in both domains, so its performance was poor.

Although the overall performance of the feature-level and decision-level bimodal classifiers was similar, an analysis of the confusion matrices of both classifiers reveals that the recognition rate for each emotion type was totally different. In the decision-level bimodal classifier, the recognition rate of each emotion increased compared to the facial expression classifier, which was the best unimodal recognition system (except happiness, which decreased in 2%). In the feature-level bimodal classifier, the recognition rate of anger and neutral state significantly increased. However, the recognition rate of happiness decreased 9 percent. Therefore, the best approach to fuse the modalities will depend on the application.

The results presented in this research reveal that, even though the system based on audio information had poorer performance than the facial expression emotion classifier, its features have valuable information about emotions that cannot be extracted from the visual information. These results agree with the finding reported by Chen et al. [4], which showed that audio and facial expressions data present complementary information. On the other hand, it is reasonable to expect that some characteristic patterns of the emotions can be obtained by the use of either audio or visual features. This redundant information is very valuable to improve the performance of the emotion recognition system when the features of one of the modal are inaccurately acquired. For example, if a person has beard, mustache or eyeglasses, the facial expressions will be extracted with high level of error. In that case, audio features can be used to overcome the limitation of the visual information.

Although the use of facial markers are not suitable for real applications, the analysis presented in this paper give important clues about emotion discrimination contained in different blocks of the face. Although the shapes and the movements of the eyebrows have been widely used for facial expression classification, the results presented in this paper show that this facial area gives worse emotion discrimination than other facial areas such as the cheeks. Notice that in this work only four affective states were studied, so it is possible that eyebrows play an important role in other emotions such as surprise.

The experiments were conducted by using a database based on one female speaker, so the three systems were trained to recognize her expressions. If the system is applied to detect the emotions of other people it is expected that the performance will vary. Therefore, more data collected from other people are needed to ensure that the variability that human beings display emotions are well represented by the database, a subject of ongoing work. Another limitation of the approach reported in this research is that the visual information was acquired by the use of markers. In real applications, it is not feasible to attach these markers to users. Therefore, automatic algorithm to extract facial motions from video without markers should be implemented. An option is to use optical flow, which has been successfully implemented in previous research [11][16].

The next steps in this research will be to find better methods to fuse audio-visual information that model the dynamics of facial expressions and speech. Segmental level acoustic information can be used to trace the emotions at a frame level. Also, it might be useful to find other kind of features that describe the relationship between both modalities with respect to temporal progression. For example, the correlation between the facial motions and the contour of the pitch and the energy might be useful to discriminate emotions.

# 6. CONCLUSION

This research analyzed the strengths and weaknesses of facial expression classifiers and acoustic emotion classifiers. In these unimodal systems, some pairs of emotions are usually misclassified. However, the results presented in this paper show that most of these confusions could be resolved by the use of another modality. Therefore, the performance of the bimodal emotion classifier was higher than each of the unimodal systems.

Two fusion approaches were compared: feature-level and decision-level fusion. The overall performance of both approaches was similar. However, the recognition rate for specific emotions presented significant discrepancies. In the feature-level bimodal classifier, anger and neutral state were accurately recognized compared to the facial expression classifier, which was the best unimodal system. In the decision-level bimodal classifier, happiness and sadness were classified with high accuracy. Therefore, the best fusion technique will depend on the application.

The results presented in this research show that it is feasible to recognize human affective states with high accuracy by the use of audio and visual modalities. Therefore, the next generation of human-computer interfaces might be able to perceive humans feedback, and respond appropriately and opportunely to changes

of users affective states, improving the performance and engagement of the current interfaces.

**Acknowledgements**: Work supported in part by NSF (through the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152 and a CAREER award) and the Department of the Army under contract number DAAD 19-99-D-0046. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

## 7. REFERENCES

[1] Black, M. J. and Yacoob, Y. *Tracking and recognizing rigid and non-rigid facial motions using local parametric model of image motion*. In Proceedings of the International Conference on Computer Vision, pages 374–381. IEEE Computer Society, Cambridge, MA, 1995.

[2] Boersma, P., Weenink, D., *Praat Speech Processing Software*, Institute of Phonetics Sciences of the University of Amsterdam. http://www.praat.org

[3] Burges, C. *A tutorial on support vector machines for pattern recognition.* Dat Mining and Know. Disc., vol. 2(2), pp. 1–47, 1998.

[4] Chen, L.S., Huang, T. S., Miyasato T., and Nakatsu R. *Multimodal human emotion / expression recognition*, in Proc. of Int. Conf. on Automatic Face and Gesture Recognition, (Nara, Japan), IEEE Computer Soc., April 1998

[5] Chen, L.S., Huang, T.S. *Emotional expressions in audiovisual human computer interaction*. Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on, Volume: 1, 30 July-2 Aug. 2000. Pages: 423 - 426 vol.1

[6] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J.G. *Emotion recognition in human-computer interaction*. Signal Processing Magazine, IEEE, Volume: 18, Issue: 1, Jan 2001. Pages: 32 – 80

[7] De Silva, L. C., Miyasato, T., and Nakatsu, R. *Facial Emotion Recognition Using Multimodal Information*. In Proc. IEEE Int. Conf. on Information, Communications and Signal Processing (ICICS'97), Singapore, pp. 397-401, Sept. 1997.

[8] De Silva, L.C., Ng, P. C. *Bimodal emotion recognition*. Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, 28-30 March 2000. Pages: 332 – 335.

[9] Dellaert, F., Polzin, T., Waibel, A. *Recognizing emotion in speech*. Spoken Language, 1996. ICSLP 96. Proceedings. Fourth International Conference on, Volume: 3, 3-6 Oct. 1996. Pages: 1970 - 1973 vol.3.

[10] Ekman, P., Friesen, W. V. *Facial Action Coding System: A Technique for Measurement of Facial Movement*. Consulting Psychologists Press Palo Alto, California, 1978.

[11] Essa, Pentland, A. P. *Coding, analysis, interpretation, and recognition of facial expressions*. IEEE Transc. On Pattern Analysis and Machine Intelligence, 19(7):757–763, JULY 1997.

[12] Huang, T. S., Chen, L. S., Tao, H., Miyasato, T., Nakatsu, R. *Bimodal Emotion Recognition by Man and Machine*. Proceeding of ATR Workshop on Virtual Communication Environments, (Kyoto, Japan), April 1998.

[13] Lee C. M., Narayanan, S.S., Pieraccini, R. *Classifying emotions in human-machine spoken dialogs*. Multimedia and Expo, 2002. ICME '02. Proceedings. 2002 IEEE International. Conference on , Volume: 1 , 26-29 Aug. 2002. Pages:737 - 740 vol.1

[14] Lee, C. M., Yildirim, S., Bulut, M., Kazemzadeh A., Busso,C., Deng, Z., Lee, S., Narayanan, S.S. *Emotion Recognition based on Phoneme Classes*. to appear in Proc. ICSLP'04, 2004.

[15] Lee C. M., Narayanan S.S. *Towards detecting emotions in spoken dialogs*. IEEE Trans. on Speech & Audio Processing, in press, 2004.

[16] Mase K. *Recognition of facial expression from optical flow*. IEICE Transc., E. 74(10):3474–3483, 0ctober 1991.

[17] Massaro, D. W. *Illusions and Issues in Bimodal Speech Perception*. Proceedings of Auditory Visual Speech Perception '98. (pp. 21-26). Terrigal-Sydney Australia, December, 1998.

[18] Nwe, T. L., Wei, F. S., De Silva, L.C. *Speech based emotion classification*. Electrical and Electronic Technology, 2001. TENCON. Proceedings of IEEE Region 10 International Conference on, Volume: 1 , 19-22 Aug. 2001. Pages: 297 - 301 vol.1

[19] Pantic, M., Rothkrantz, L.J.M. *Toward an affect-sensitive multimodal human-computer interaction*. Proceedings of the IEEE , Volume: 91 Issue: 9 , Sept. 2003. Page(s): 1370 – 1390.

[20] Roy, D., Pentland, A. *Automatic spoken affect classification and analysis*. Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on , 14-16 Oct. 1996. Pages: 363 – 367

[21] Tian, Ying-li, Kanade, T. and Cohn, J. *Recognizing Lower Face Action Units for Facial Expression Analysis*. Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00), March, 2000, pp. 484 – 490.

[22] Yacoob, Y., Davis, L. *Computing spatio-temporal representations of human faces*. Computer Vision and Pattern Recognition, 1994. Proceedings CVPR '94., 1994 IEEE Computer Society Conference on , 21-23 June 1994 Page(s): 70 –75.

[23] Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Deng, Z., Busso, C., Lee, S., Narayanan, S.S., *Analysis of acoustic correlates in emotional speech*. to appear in ICSLP'04, 2004.

[24] Yoshitomi, Y., Sung-Ill Kim, Kawano, T., Kilazoe, T. *Effect of sensor fusion for recognition of emotional states using voice, face image and thermal image of face*. Robot and Human Interactive Communication, 2000. RO-MAN 2000. Proceedings. 9th IEEE International Workshop on, 27-29 Sept. 2000. Pages: 178 – 18.