Volume 3, Issue 8, August 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Bimodal Emotion Recognition: A Comparative Study of Rule Based System Vs Classification Algorithms

Preeti Khanna[*]
SBM, SVKM's NMIMS
India

Sasikumar, M.
Director R & D, CDAC Mumbai
India

Abstract— ***Emotions can be understood in a face to face interaction immediately while in a human computer interaction (HCI) this response is limited. Research studies have been undertaken to investigate and develop various approaches and technology to incorporate emotions in HCI. One major concern of HCI now is the need to improve the interactions between humans and computers through justifications and explanations. Now a day's HCI is experimenting with alternate input mechanisms through speech, gesture, posture and facial expression. Endowing computers with this kind of intelligence is a complex task. It becomes more complex with the fact that the interaction of humans with their environment (including other humans) is naturally multimodal. In reality, one uses a combination of modalities and they are not entirely independent. As compared to unimodal approaches, various problems arise in case of multimodal emotion recognition especially concerning fusion architecture of multimodal information. We introduce a rule based hybrid approach to combine the information from different sources like facial expression and speech for recognizing emotions. The results presented in this paper shows that it is feasible to recognize human affective states with a reasonable accuracy by combining the modalities together using rule based system. The performances of our hybrid model i.e. rule based approach is compared with the traditional model i.e. feature based fusion model.***

Keywords— ***Human computer interaction (HCI); emotion recognition; rule based system; emotional state; modalities.***

I. INTRODUCTION

People dependencies on machines have been increasing significantly. For quicker and more effective interaction, machines are being equipped with more and more user friendly interfaces. For example, more and more people enter an online community of the global networks (Internet) and communicate with each other through computers. This kind of wide adoption leads to increasingly importance to HCI. One of the current focuses of HCI is to improve the interaction between human and computers through justification and explanation. In this context, it may be noted that emotions play an important role in our everyday lives, in whatever we do. Emotions, sentiments and moods affect each aspect of response and interactions in our lives, be it with humans, technology or human via technology. Research studies have been undertaken to investigate and develop various approaches to incorporate emotions in HCI. Some of the recent research focuses on how a computer can automatically detect the emotional state of a user and then adapt its behaviour accordingly. Affective computing is a relatively new field of study researching the use of emotional capabilities in computing machines. Emotion enhanced human computer interfaces is expected to acquire a higher rate of acceptance and range of applicability as a result of the increased trust the users develop in such technology. There is increasing research interest along these lines. Some of the prominent areas include e-commerce, help desks, customer support, e-learning, game and entertainment industry etc. Surveillance is another application domain in which the reading of emotions may lead to better performance in predicting the future actions of subjects. Yet, human computer interactive systems capable of sensing stress, inattention and confusion and hence capable of adapting and responding to the affective states of end users are likely to be perceived as more natural and trustworthy [1], [2] and [3].

The issue of enhancing HCI with emotion raises a number of questions. What are the sources of information that a machine can use to decode the emotional state of the user? What kind of information (emotional cues) are available from these sources? How does one use these sources to estimate the emotional state? What are the emotional states of interest for us from the perspective of enhancing HCI? How to combine multiple modalities? Does the performance of the multimodal emotion recognition for a specific set of target emotions depend on type of fusion model? In realistic scenarios, can the modalities be treated separately? There is a plethora of existing work that bears on one or more of these questions. The paper begins by defining problem domain regarding emotion recognition and its applications. In this paper, we concentrate on the problem of integrating the inputs from sources like facial expression and speech. Also we look at the problem of fusing the inputs through different options other than traditional method of fusing inputs known as feature based fusion approach. Section II discusses the complete framework of our rule based system i.e. hybrid approach of integrating inputs from different modality. This approach is based on certainty factor i.e. the MYCIN approach. Section III talks about the overall framework of emotion recognition independent of any modalities. Then we explain our approach of rule based system for emotion recognition explaining with the running case scenarios of 'speech' in section

IV. We did few experiments on bimodal data and tested on the rule based system and feature based fusion method (using classification algorithms).The comparison of these has been explicitly mentioned in section V. We conclude the paper by summarizing the results and consider some challenges facing the researchers in this area.

A. The Problem Domain for Bimodal Emotion Recognition: from Fusion Perspective

Most of the existing work in emotion recognition addresses the unimodal analysis with various emotion sources considered independent of each other. It is well understood that humans recognize emotion, fusing information from multiple sources: speech signals, facial expressions, gesture, bio-signals and others. To have the human computer interaction be as natural as possible, it is desirable that computers should be able to interpret the human actions effectively. Hence, inadequacies of unimodal recognition systems drive the need to go for multimodal recognition. In literature, some attempts like [4], [5], [6] and [7] have considered the integration of information from facial expressions and speech. This paper explores how to combine the information from various sources [23], [24] and [25] to achieve better recognition of emotional state using rule based approach. And then the performance of rule based approach (which is hybrid approach, explained later) has been compared with the feature based fusion approach, which is traditional approach.

There are two broad approaches to design of a bimodal or multimodal recognition engine: feature fusion and decision fusion. Feature level fusion involves simply merging the features of each modality into a single feature vector. In this method of fusing, all the features are mixed together irrespective of their nature and type. Feature can be position of some feature points on the face or the prosodic features of a speech signal. Feature sets can be quite large as we will see later. This high cardinality can result in soaring computational cost for this fusion approach [8]. Decision level fusion is based on the fusion of decisions from each modality where the input coming from each modality is processed independently and these unimodal recognition results are combined at the end [9]. This fusion has advantage of avoiding synchronization issues over the feature level fusion. Decision level fusion ignores possible relationships between features coming from different modalities. Several works [10], [11] and [12] have discussed multimodal fusion; in particular [13] discusses many issues and techniques of multimodal fusion. Finding an optimal fusion type for a particular combination of modalities is not straightforward. A good initial guess can be based on the knowledge of the interaction and synchronization of those modes in a natural environment. Hybrid fusion attempts to combine the benefits of both feature level and decision level fusion methods. This may be a good choice for some multimodal fusion problems. However, based on existing knowledge and methods, how to combine the information coming from different modalities for the target set of emotions is still an open problem. We propose a rule based structure, a hybrid approach by fusing the data from modalities like facial expression and speech which combine the benefits of both methods of fusion i.e. feature level and decision level fusion. Also we plan to use feature based approach to see the influence of closely coupled features from different modalities like lip movement with the speech. Feature level fusion involves simply concatenating each modality feature or each stream of sensory data feature into a single feature vector. Decision level fusion ignores possible relationships between features coming from different modalities. We, therefore, present a comparative analysis to compare the performances for recognizing the target emotions across the two approaches - rule based hybrid approach and feature based fusion approach. The research design for the first approach is discussed below.

II. Rule Based System: Base For Our Hybrid Model

A rule based system consists of if-then rules, a bunch of facts, and an interpreter controlling the application of the rules. A simple if-then rule has the form ‘if x is A, then y is B’. One of the major strength of rule based representation is its ability to represent various uncertainties. Uncertainty is inherently part of most human decision making. This uncertainty could arise from various sources like incomplete data or domain knowledge used being unreliable.

A. Approaches for Handling Uncertainty

To handle these uncertainties, there are two broad approaches - those representing uncertainty using numerical quantities and those using symbolic methods. In numerical approaches, one models the uncertainty by numbers and provides some algebraic formulae to propagate these uncertainty values to the conclusions. These approaches are useful for handling the issues related to “unreliable or inaccurate knowledge”. For example, Bayesian reasoning [15], Evidence theory [16] and Fuzzy set approaches [17] are numerical models. On the other hand, symbolic characterization of uncertainty is mostly aimed at handling incomplete information, e.g., Assumption Based Reasoning [18], Default Reasoning [19] and Non-monotonic Logic [20]. For example, if there is not enough information available, the system makes assumptions that can be corrected later, when more information is received. In our domain, the basic problem is that there are hardly any features or feature combinations which can infer any emotion to complete certainty. Therefore, we concentrate on numerical approaches for handling the uncertainty. We have adopted the ‘Confirmation Theory’ as used in MYCIN approach [15] to deal with uncertainty in our domain. This approach works well with rule based representation of domain knowledge.

B. Reasoning with Certainty Factors: The MYCIN Approach

Shortliffe and Buchanan [15] developed the Certainty Factor (CF) model in the mid 1970s for MYCIN, an expert system for the diagnosis and treatment of infections of the blood. Since then, the CF model has been widely adopted for uncertainty management in many rule based systems. Each rule is assigned CF by domain experts. This is meant to represent the uncertainty of the rule. Higher CF indicates that the conclusion can be asserted with higher confidence when the conditions are true. Similarly every fact in the domain is also given CF indications how confident one is in that. Shortliffe and Buchanan [15] intended a CF to represent the change in belief in a hypothesis given some evidence. In particular, a CF between 0 and 1 means that the person's belief in h given e increases, whereas a CF between -1 and 0 means that the belief decreases. A value of +1.0 indicates absolute belief and -1.0 indicates absolute disbelief. The method generally used to propagate the measure of uncertainty in the antecedents and the uncertainty attached to the rule to the conclusions being derived is briefly explained below. This propagation is done in two steps [14].

(i) The different antecedents in the rule, in general, have different values of uncertainty attached to them. Some formula is required to combine these measures and provide a consolidated uncertainty number. This option considers the strength of the weakest link in a chain as the strength of the chain. This is defined as:

$$CF\textit{antecedents} = \{\text{minimum of CFs of all antecedents}\} \qquad (1)$$

(ii) Then this measure (uncertainty for the set of antecedents) is combined with the measure of uncertainty attached to the rule to give a measure of uncertainty for the conclusion of the rule.

$$\text{CF of the conclusion from rule} = \{\text{CF associated with rule R1}\}*\{CF\textit{antecedents}\}, \text{ provided } CF\textit{antecedents} >= \text{ threshold} \qquad (2)$$

It can be seen that the CF obtained for a conclusion from a particular rule will always be less than or equal to the CF of the rule. This is consistent with the interpretation of the CF used by MYCIN, that is, the CF of a rule is the CF to be associated with the conclusion if all the antecedents are known to be true with full certainty. In a typical rule based system, there may be more than one rule in the rule base that is applicable for deriving a specific conclusion. Some of them will not contribute any belief to the conclusion, because CF of antecedents has a CF less than the threshold. The contributions from all the other rules for the same conclusion have to be combined. For MYCIN model, initially CF of a conclusion is taken to be 0.0 (that is, there is no evidence in favour or against) and then as different rules for the conclusion fires, the CF gets updated. MYCIN uses a method that incrementally updates the CF of the conclusion as more evidence for and against is obtained. Let CFold be the CF of the conclusion so far, say, after rules R1, R2,...Rm have been fired. Let CFin be the CF obtained from firing of another rule Rn. The new CF of the conclusion (from rules R1, R2………..Rm and Rn), CFnew, is obtained using the formulae given below.

$$CFnew = CFold + CFin * (1 - CFold) \quad \text{when } (CFold, CFin > 0) \qquad (3)$$

$$CFnew = CFold + CFin * (1 + CFold) \quad \text{when } (CFold, CFin < 0) \qquad (4)$$

$$CFnew = (CFold + CFin) / (1 – \min (|CFold|, |CFin|)) \text{ otherwise} \qquad (5)$$

We adopt this calculus in our model and explain this later with a running example with speech in section IV. Before that we first discuss the overall framework of emotion recognition system.

III. GENERAL FRAMEWORK FOR EMOTION RECOGNITION

The conceptual framework for emotion recognition includes pre processing, feature extraction, feature analysis, selection of the features, formulation of rules and measuring performance to classify the target emotional states. We will explain each of these in brief below. We will use speech input as the running example to illustrate these stages later. The framework remains same across all modalities.

A. Pre Processing and Feature Extraction

The objective of this step is to make the input data in a standard format and suitable for extracting the desired features. Usual pre-processing steps include size normalization of the frontal image, noise removal from speech signal etc. The next step is feature extraction. The work in this step involves identifying relevant features and formulating algorithms to extract these features from their respective input data.

B. Feature Analysis and Selection

Once the basic feature set is ready, the next step is analysis of these features. The question, 'how does each of the features vary with the emotion' needs to be answered here. Usually every feature doesn't contribute to the same extent to recognize different emotional states. Thus feature analysis and selection is an important step.

C. Formulation of Rules

To design the rules for classifying emotions, all the relevant features needs to be studied in more detail to see its ability to distinguish between different target emotional states. Influential and useful features can be used to define rules. This approach remains broadly same across different modalities and is as follows:

(i) Feature analysis has been done for each feature to see its ability to distinguish among the target emotional states, and accordingly useful features were shortlisted.

(ii) Rules are formed using each of these features for different target emotional states. A feature may yield one or more rules. Generally these rules have the form: if feature F1 has value less than or greater than T1 and feature F1 has value less than or greater than T2 then conclude emotion = e1. For each rule, the cut- off points T1 and

T2 for a given emotion class is taken to be the approximate average of the value of that class with its immediate neighbor emotional class.

(iii) Corresponding to each rule, we associate CF values for each emotional class. These values of CFs are decided on the conditions mentioned in Table- 1.

This heuristic has been arrived at based on empirical studies of the various feature graphs and behavior of the CF calculus. There may be multiple rules associated with each feature. Multiple rules, when they fire simultaneously (based on values of different features) may saturate the values of CF associated with them. To minimize this possibility, we have chosen relatively lower range of CF values. Given our observation that most features do not provide a high degree of discrimination for any of the emotions, a high value did not appear justified for any individual feature. The chosen range also allows the CF value to climb steadily to a high range, when there are many features supporting an emotion. The rules may point to a specific emotional state or a set of emotional states. If the distance of an emotion with its neighboring emotion is found to be less than 5– 6% of the entire spread (overall range) for that features value, then these emotions are grouped as a subset.

TABLE 1: DEFINING CERTAINTY FACTOR (CF) FOR RULES

Range of the CF	**CF Values**	**Belief and Disbelief**	**Indicated by**
Greater than 0.2 and up to 0.4	0.3	High evidence	High Inter class distance
Greater than 0.1 and up to 0.2	0.2	Moderate evidence	Medium Inter class distance
Equal to 0.1	0.1	Low evidence	Low Inter class distance

Allocation of the values of CF to these classes is done based on the following rules, derived based on analysis of the emotion profile.

- High Interclass Distance: If the interclass distance of an emotional class (either singleton or non-singleton) with its neighbors (left side and right side) is more than 15% of the entire spread for that feature, then the chances of a confusion with the neighboring class is low and hence the CF value associated with this class for that feature is 0.3.

- Medium Interclass Distance: If the interclass distance of a emotional class (either singleton or non-singleton) with its neighbors (left side and right side) is in between 6-15% of the entire spread for that feature, then the CF value associated with this class is 0.2.

- Low Interclass Distance: If the interclass distance of a emotional class (either singleton or non-singleton) with its neighbors (left side and right side) is less than 6% of the entire spread for that feature, then the CF value associated with this class is 0.1.

D. Classification into Emotional Categories

As known from the literature, a number of algorithms are available for classification- each with its own strength and weaknesses. We plan to investigate different classification algorithms like Multilayer Perceptron, Simple Logistics, Logistics, BayesNet, Multiclass classifier and Classification via regression for emotion recognition. The software used to implement the models for the classification is Weka software (open source software). The exercise is done for the modalities like facial expression and speech. The next section discusses one of the case scenarios for speech.

IV. CASE STUDY FOR SPEECH

We illustrate the process with a concrete example of emotion recognition from speech. We have used Danish Emotional Speech (DES) [21] databases that cover six emotional states (anger, happy, fear, sad, neutral and bore). We utilize 477 utterances from 10 subjects. We have 5 female and 5 male subjects having 260 and 217 emotional utterances respectively for the emotional states of anger, happy, fear, sad, neutral and bore.

A. Pre-processing and Feature Extraction

To deal with discrete-time signal x(n), framing is used to divide the signal into frames. The speech signal is first segmented into several data frames of length 20ms. Each frame overlaps with the adjacent frames by 10ms. The next step is to apply the Hamming window to each individual frame to minimize the signal discontinuities at the beginning and end of each frame. The further process consists of looking for specific indicators (features) in the speech signal which carries relevant information about our select set of emotions. We are considering basic group of prosodic features and Mel

Frequency Cepstral Coefficients, MFCC (13 coefficients). 13 MFCC coefficients as the feature set for each input speech signal for all the six emotions have been extracted. Then the vector quantization based on Linde-Buzo-Gray (LBG) k-means (4, 8, 16 and 32, i.e. by varying the size of the codebook) clustering algorithm [22] has been performed. Finally three different distance method to classify emotional states have been used. These are chebyshev distance, euclidean distance and manhattan distance. Each of these is calculated with respect to neutral state of the sample. This whole experiment (LBG-VQ on MFCC) is repeated by changing the size of the codebook from 4, 8, 16 and 32 to see the variation across it. Table -2 describes the feature set which is extracted using various algorithms and methods for our experiments. The pitch or the fundamental frequency of the speech signal is calculated using cepstrum analysis of the signal as well as using autocorrelation method. Also the first four formant frequencies of the speech signal using Linear Predictive Coding (LPC) coefficients have been determined. These frequencies F1, F2, F3 and F4 represent the resonance frequencies of the vocal tract.

TABLE2: LIST OF FEATURES CHOSEN FOR STUDY OF SPEECH MODALITY

Set of Features	
Statistics related to pitch using autocorrelation method	F0auto (mean, min, max, range, standard deviation, mean average slope)
Statistics related to pitch using cepstrum method	F0cep (mean, min, max, range)
Statistics over the individual voiced and unvoiced parts	number of voiced regions, number of unvoiced region, average voiced length, average unvoiced length
Statistics related to intensity	average energy (Short term energy)
Statistics related to rhythm	speaking rate
Formants	F1, F2, F3 and F4
Total no. Prosodic Features	*20 Prosodic Features*
Mel-scale Frequency Cepstral Coefficient	*13 MFCC Coefficients*

B. *Feature Analysis*

We first studied each of these features individually to see their variation across the various emotional states, and their discrimination ability against the set of emotions for each of the individual feature. We used the average percentage deviations w.r.t neutral state of selected features for each of the six emotional states. The average percentage deviations were calculated as follows:

Percentage Deviation = ((Measured Value – Neutral State Value) / Neutral State Value)*100

For example, Fig. 1 shows the variations for the features 'speaking rate' across emotions. It is observed that 'anger' has the highest value of speaking rate as compared to other emotional states and then 'happy' followed by 'fear' and then 'bore'. But 'sad' value is coming closer to 'neutral' hence hard to recognize separately using speaking rate only as a single input. This kind of analysis is important to know which kind of feature is good for a set of emotions, or a specific emotion and hence worth considering for emotion recognition. We found that set of features (includes {F0(mean), F0(min), F0(max), range} using cepstral method, {F0(mean), F0(min), F0(max), range, SD, mean avg slope} using autocorrelation method, formant frequencies (F1, F2, F3 and F4) and derived features {((F1*1000)/(F2*F3), (F4*F3)/ (F1*F2), (F0CEP)/F1}, speaking rate) shows significance variation across the interested emotional states. This analysis tells us that each of these features doesn't contribute to the same extent to recognize a set of emotional state or a specific emotional state.

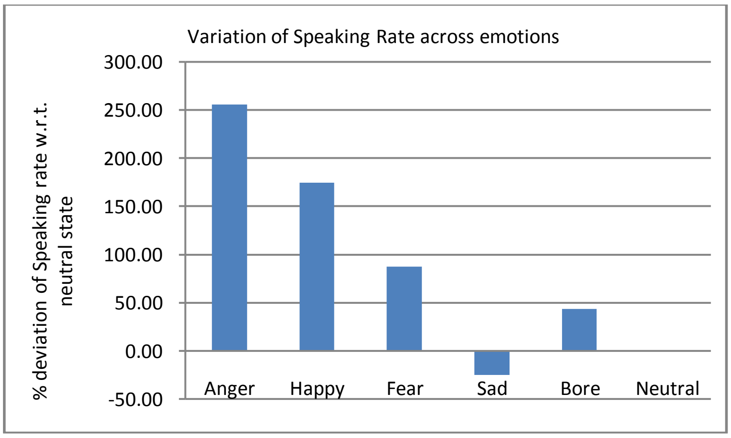
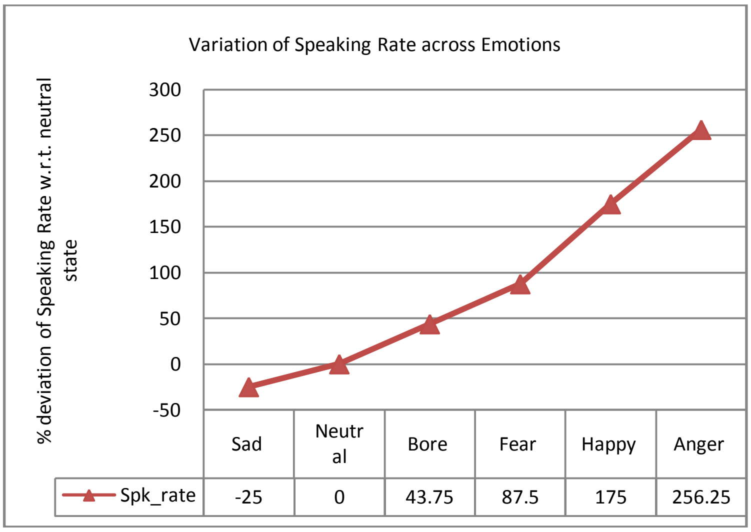


Fig 1: Variation of Speaking Rate across Emotions

C. Formulation of Rules

As discussed earlier all features might not be useful in forming the rules. Individually each of these has to be analysed. Fig. 2 shows the variation of speaking rate across emotions for gender independent scenario which forms six singleton classes belonging to each individual emotion. Rules have been designed with proper CFs to these singleton classes as per interclass distances.



	Sad	Neutr al	Bore	Fear	Happy	Anger
Spk_rate	-25	0	43.75	87.5	175	256.25

Fig 2: Variation of Speaking Rate across Emotions

Depending on the interclass distances of these classes CFs has been allocated (Table -1) and rules have been formed. For each rule (of the type if – then), the cut off point (i.e., upper limit and lower limit) belonging to the emotion class is taken to be the average of the value of that class with its immediate emotional class. For example, for 'bore' emotion the cut off points to be considered are 22 and 66, forming the singleton class and due to high inter class distances the CF values is to be considered as 0.3. Similarly, all other emotional states are formed to be singleton class having CF as 0.3. We found a total of six conditions each for the feature 'speaking rate' to classify emotions. Examples of rules are shown below.

Example Rule 1: Using the feature speaking rate for emotion identification

```
(i)   if (Speaking_rate <= -12)
                 CFSad=0.3;
(ii)  if ((Speaking_rate > -12) && (Speaking_rate < 22))
                 CFNeu=0.3;
(iii) if ((Speaking_rate >= 22) && (Speaking_rate < 66))
                 CFBore=0.3;
(iv)  if ((Speaking_rate >= 66) && (Speaking_rate < 131))
                 CFFear=0.3;
(v)   if ((Speaking_rate >= 131) && (Speaking_rate < 216))
                 CFHap=0.3;
(vi)  if(Speaking_rate >= 216)
                 CFAng=0.3;
```

Total of 12 features are found to be relevant for designing rules for emotion identification using speech. These features include – speaking rate, formant frequency (F1, F2, F3 and F4), fundamental frequency (F0auto and F0cep), measure of euclidean distance, measure of chebyshev distance, mean_average_slope (using autocorrelation method), range of fundamental frequency (using autocorrelation method) and standard deviation of fundamental frequency (using autocorrelation method). Total of 13 rules have been formulated using these features and tested on the DES database and final value of CF has been updated corresponding to each of 6 emotional states (sad, neutral, bore, happy, fear, anger).

D. Recognizing Emotions from Speech Using Rules

These rules have been tested on the DES database and final value of CF has been computed corresponding to each of the 6 emotional states. The emotion with the highest value of final CF is considered and counted against the expected emotion class for each image for all the subjects. For example, Table - 3 shows the computed values of CF labelled as CF_Sad, CF_Neu, CF_Ang, CF_Happy, CF_Fear and CF_Disgust corresponding to all the six emotions - sad (S), neutral (N), anger (A), happy (H), fear (F) and disgust (D). Each row of this table-3 indicates an input signal tested for a particular emotion belonging to an individual subject labelled as 1, 2 etc. Each instance has been tested across emotions.

Final outcome for the same is indicated with CF values under the six columns labelled from CF_Sad to CF_Fear. For example, instance -3 shows the maximum value of CF under the emotion class of 'happy' (0.87). In this case the actual emotion of the subject is 'anger'; but it is not able to recognize correctly. Instance 1 is detecting the correct emotion 'sad' with CF (0.91) associated with it - the highest. Similarly each computed value of CF in each instance, has been analyzed for each of the emotions.

TABLE 3: EXAMPLES OF COMPUTED VALUES OF CF USING RULES FROM FACE FOR FEMALE SUBJECT

Subjects	**Actual Emotions**	**Updated Value of CF computed using rules for respective emotion**					
		CF _ Sad	**CF_ Neu**	**CF_ Ang**	**CF_ Happ y**	**CF_ Bore**	**CF_ Fear**
s1	S	0.91	0.76	0.20	0.30	0.51	0.44
s1	N	0.30	0.96	0.00	0.00	0.78	0.30
s1	A	0.73	0.56	0.66	0.87	0.28	0.30
s1	H	0.51	0.51	0.83	0.76	0.00	0.76
s1	B	0.44	0.89	0.20	0.00	0.92	0.30
s1	F	0.73	0.88	0.00	0.73	0.51	0.66

The overall correctness of recognizing emotions using rule based approach in a unimodal system from speech is found to be 71.66%. The recognition rates are found to be 70% and 73.33% for female and male subjects respectively.

V. PERFORMANCE MEASURED OF RULE BASED SYSTEM VS CLASSIFICAITON ALGORITHMS

A. Bimodal Emotion Recognition using Rule Based Method: Hybrid Approach

We extend emotion recognition based on our rule based model by combing facial expression with speech. This model is based on preparing a set of rules derived from the individual modalities. The rules are mixed together independent of the modality into a single group. In order to test, we propose to include the data source as facial expressions with speech. The database used in the experiments consists of audio samples and static frontal images of different people (graduate students in the age group of 21 to 28 years). Total of 11 subjects participated in our experiment (5 female and 6 male). Each of these subjects was told to read a single sentence under four emotional states (anger (A), happy (H), sad (S) and neutral (N)). For the process of inducing the desired emotional state, individual subjects were shown a small video clipping of 2-3 minute corresponding to each of the four emotional categories. During this, facial expression was captured by the digital camera. The subjects chosen in our experiment don't wear 'glasses' and males don't have 'beard' on their face – this made the analysis easier. We have total of 20 images with utterances (5 each of 'anger', 'happy', 'sad', and 'neutral') of female and 23 images (6 each of 'anger', 'sad' and 'neutral' but 5 is of 'happy') with utterance of male subjects. The compiled set of rules for speech and facial expression was run against this dataset. Table - 4 shows the results obtained using facial expression and speech modalities in unimodal as well in bimodal scenarios using rule based approach.

The average emotion recognition rate of the system using our own database is found to be 65% (for female subjects), 65.21% (for male subjects) and 67.44% overall using facial expressions. The emotion 'sad' is the best recognized and has 82% recognition rate overall. But this is not true with male subjects. 'Anger' is hard to distinguish from others and hence having the least accuracy. The average emotion recognition rate of the system was found to be 55% (for female subjects), 62.5% (for male subjects) and 56.6% overall using speech. It has been observed that the emotion 'happy' is hard to recognize both in female as well as in male subjects. The emotion 'sad' shows reasonably good recognition rate for male as well as female subjects.

TABLE 4: RECOGNITION RATE (%) FOR UNIMODAL VS BIMODAL USING RULE BASED APPROACH

Recognition rate in %			
Rule Based Approach	Gender Dependent		Gender Independent
	Female	Male	
Facial Expression	65	65.21	67.44
Speech	55	62.5	56.6
Bimodal (facial expression with speech)	75	65.21	67.44

The average emotion recognition rate of the bimodal emotion recognition system (adding the two sets of rules together) using rules is found to be 75% (for female subjects), 65.21% (for male subjects) and overall 67.44%. It has been observed that overall performance has increased by combining the inputs from speech signal and facial image in case of gender independent as well as gender dependent scenario.

B. Bimodal Emotion Recognition using Classification Algorithms: Feature Based Fusion Approach

Experiments have been done using Multilayer Perceptron, Simple Logistics, NaiveBayes, Multiclass Classifier and classification via regression classification algorithms with 10-fold validation method using weka software by selecting the features set (as selected for rules formulation). Table - 5 shows the results obtained using facial expression and speech modalities in unimodal as well in bimodal scenarios using Multilayer Perceptron classifier.

TABLE 5: RECOGNITION RATE (%) FOR UNIMODAL VS BIMODAL USING FEATURE BASED FUSION APPROACH

Recognition rate in % using Multilayer Perceptron			
Classification Algorithm	Gender Dependent		Gender Independent
	Female	Male	
Facial Expression	65	57	58
Speech	55	50	64
Bimodal (facial expression with speech)	70	73.91	76.74

For the Multilayer Perceptron classifier, the overall performance is 65% (for female), 57% (for male) and 58% for gender independent case using facial expression. As observed from our experiment, the two emotions are found to be confused more. ‘Anger’ is highly confused with ‘sad’ (9% with happy, 27% with neutral and 36% with sad). The emotion ‘sad’ is highly confused with ‘neutral’ (20% with anger, 20% with happy and 27% with neutral). Classifiers like Logistics, NaiveBayes and Multiclass classifier are doing well and recognition rate is found to be in between 65% to 68%. Using speech classifier like Multilayer Perceptron don’t show reasonable recognition rate to classify the target emotions for male as well as for female subjects. In case of gender independent case, some of the classifiers like Simple Logistics, BayesNet, and Multilayer Perceptron are showing a little better classification rate (i.e. from 63% to 68%). Table- 6 shows the comparison of recognition rate (%) for different types of classifiers.

TABLE 6: RECOGNITION RATE (%) FOR VARIOUS CLASSIFICATION ALGORITHMS FOR UNIMODAL VS BIMODAL

Recognition rate in %			
Classification Algorithms	Gender Dependent Scenario		Gender Independent
	Female	Male	
Multilayer Perceptron	70	73.91	76.74
Logistics	75	65.21	67.44
Simple Logistics	75	60.86	74.41
BayesNet	65	NA	60.46
Multiclass classifier	65	69.56	62.79
Classification via regression	70	52.17	74.41

In this table-6, the field value corresponding to ‘NA’ indicates that particular classifier was not suitable for such database as recognition rate is coming out to be less than 50%. It is observed that overall performance (rate of recognition in %) has increased by combining speech signal with facial expression, with almost all the classifiers tried. Now the highest recognition rate of 75% using Logistics and Simple Logistics classifier for female subjects and 73.91% using Multilayer Perceptron for male subjects. For gender independent case, the highest recognition rate of 76.74% was observed using Multilayer Perceptron.

C. Comparative Analysis of Rule Based Hybrid Approach and Feature Based Fusion Approach using Classification Algorithms

The overall performance of rule based hybrid model is found to be better as compared to feature level fusion. This is true for female and male subjects tested for unimodal systems but for bimodal the feature level fusion perform better. The rate of recognition of emotions for female is better than male subjects for both feature level fusion as well as rule based system. This is true for unimodal as well as bimodal system. We analysed and compared the performance of unimodal and bimodal system using facial expression and acoustic information as an input. We have seen that some pairs of emotions are usually misclassified. But the overall performance of the bimodal emotion recognition classifier was found to be higher than each of the unimodal systems. The results presented in this research show that it is feasible to recognize human affective states with a reasonable accuracy by combining the modalities together. Therefore, the next generation of human computer interfaces might be able to perceive humans feedback, and respond appropriately and opportunely to changes of user affective states, improving the performance and engagement of the current interfaces.

VI. CONCLUSION

The focus of the paper is to compare the performance of the traditional approach which is feature based fusion approach and hybrid approach i.e. rule based approach to recognize the user’s emotional state from facial expressions and speech. The bimodal emotion recognition framework (by considering two modalities - facial expression and speech) has been formulated which built around if-then rules using certainty factors to capture uncertainty and unreliability of individual features. To the best of our knowledge, this approach for emotion recognition has not been tried in the literature. This technique appears to be simple and effective for this problem. This rule based approach for emotion recognition could be extended to multimodality. The process for the formulation of rules (derived from facial expression and speech) for classifying into different emotions have been defined. A general framework for defining the certainty factors associated to such rules have been defined which is independent of the modality and feature used. Performances of unimodal and bimodal system using facial expression and acoustic information as an input were compared. The overall performance of 'rule based hybrid model' is found to be better as compared to feature level fusion. This is true for female and male subjects tested for unimodal systems but for bimodal the feature level fusion perform better. The overall performance of 'rule based hybrid model' is found to be better compared to feature level fusion. The rate of recognition of emotions for females is better than male subjects (except for speech modality) for both feature level fusion as well as rule based system.

There are a number of avenues for extending this work. A more realistic evaluation with large data and more modalities is, perhaps, the most important. At present, we have used the Confirmation theory as used in MYCIN approach [12]. One of the major concerns against the use of certainty factor is that they have no sound theoretical basis; though, they often work well in practice. We allocated the values of CF to the emotional classes based on heuristic rules as defined in section III. These have been derived based on the analysis of the individual features across different emotions. In this work, we have ignored the possibility of having more than one emotional state at a time. We would also like to investigate alternative uncertainty models like the Dempster-Shafer Theory. Demspter Shafer theory provides more flexibility in assigning belief to various subsets of emotions. The multimodal data fusion for emotion recognition remains an open challenge as several problems still persist, related to finding optimal features, integration and recognition. Completely automated multimodal emotion recognition system is still at the preliminary phase, shows very limited performance and is mostly restricted to the lab environment.

REFERENCES

[1] R.W. Picard, *Affective computing*, The MIT Press, Cambridge, MA, (1997).

[2] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A.W. Senior, "Recent Advances in the Automatic Recognition of Audio-Visual Speech", Invited, Proceedings of the IEEE, 91(9), 1306-1326, (2003).

[3] M. Pantic, N. Sebe, J.F. Cohn, and T.S. Huang "Affective multimodal human–computer interaction", in: *Proceedings of the 13th annual ACM international conference on Multimedia*, 669–676, (2005).

[4] L.S. Chen, T.S. Huang, T. Miyasato and R. Nakatsu, "Multimodal Emotion/Expression Recognition", in *Proceedings of the 3rd International Conference on Face and Gesture Recognition*, 366-371, (1998).

[5] De Silva and Ng, "Bimodal Emotion Recognition", Automatic Face and Gesture Recognition, *IEEE International Conference*, 332 – 335, (2000).

[6] N. Sebe, I. Cohen, T. Gevers and T.S. Huang, "Emotion Recognition Based On Joint Visual and Audio Cues", *Pattern Recognition, International Conference*, 1, 1136–1139, (2006).

[7] Z. Zeng, Tu Jilin, Liu, Huang, Pianfetti, Roth and Levinson, "Audio-Visual Affect Recognition", *IEEE Transactions on multimedia*, 9(2), 424-428, (2007).

[8] B.V. Dasarathy, "Sensor Fusion Potential Exploitation Innovative Architectures and Illustrative Approaches", in *Proceeding of IEEE*, 85, 24–38, (1997).

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, L.M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of Emotion Recognition using Facial Expressions, Speech and Multimodal Information", in *Proceedings of ACM 6th International Conference on Multimodal Interfaces*, 205-211, (2004).

[10] A. Corradini, M. Mehta, N. Bernsen, and J. C. Martin, "Multimodal input fusion in human-computer interaction" on the example of the on-going nice project. In *Proceedings of the NATO-ASI conference on Data Fusion for Situation Monitoring, Incident Detection, Alert and Response Management*, Yerevan (Armenia), August 2003.

[11] H. Liao, Multimodal Fusion, Master's thesis, University of Cambridge, July (2002).

[12] S. Kettebekov, and R. Sharma, "Understanding Gestures in Multimodal Human Computer Interaction", *International Journal on Artificial Intelligence Tools*, 9(2), 205-223, (2000).

[13] R. Sharma, V. Pavlovic and T. Huang, "Toward Multimodal Human Computer Interface", *In Proceedings of the IEEE*, 86(5), 853-860, (1998).

[14] Sasikumar, M., S. Ramani, S.M. Raman, K.S.R. Anjaneyulu, K.S.R. and R. Chandrasekar, *Rule Based Expert Systems – A Practical Introduction*, Narosa Publishers, (2007).

[15] E.H. Shortliffe and B.G. Buchanan, "A Model of Inexact Reasoning in Medicine", *Mathematical Biosciences*, 23, 351-379, (1975).

[16] J. Gordon and E.H. Shortliffe, The Dempster-Shafer Theory of Evidence, in *[Buchanan and Shortliffe, 1984]*, 272-292, (1984).

[17] C.V. Negoita, Expert Systems and Fuzzy Systems, Benjamin/Cummings, (1985).

[18] J.A. Doyle, Truth Maintenance System, *Artificial Intelligence*, 12, 231-272, (1979).

[19] R.A. Reiter, Logic for Default Reasoning, *Artificial Intelligence*, 13, 81-132, (1980).

[20] D. McDermott and D. Doyle, Non-monotonic Logic I, Artificial *Intelligence*, 13, 41-72, (1980).

[21] [Online]. Available: DES database: http://www.emotionreseach.net/biblio/; accessed on June (2008).

[22] Y. Linde, A. Buzo and R. Gray, “An Algorithm for Vector Quantizer Design”, *IEEE Transactions on Communications*, 28, 84-95, (1980).

[23] P. Khanna and M. Sasikumar, “Recognizing Emotions from Keyboard Stroke Pattern”, International Journal of Computer Applications, 11(9): December, (2010).

[24] P. Khanna and M. Sasikumar,”Recognizing Emotions from Human Speech”, Think Quest 2010, International Conference on “Contours of Computing Technology in association with Springer Publications, March (2010).

[25] P. Khanna and M. Sasikumar, “Rule based System for recognizing Emotions using Multimodal Approach”, International Journal of Advanced Computer Science and Applications” 4(7), (2013).