Journal of Theoretical and Applied Information Technology
10th December 2015. Vol.82. No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

EMOTION DETECTION THROUGH SPEECH AND FACIAL EXPRESSIONS

[1]KRISHNA MOHAN KUDIRI, [2]ABAS MD SAID AND [3]M YUNUS NAYAN

[1] Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia

[2]Assoc. Prof., Computer and Information Sciences, Universiti Teknologi PETRONAS, Malaysia

[3]Senior Lecturer, Fundamental and Applied Sciences, Universiti Teknologi PETRONAS, Malaysia

E-mail: [1]kmk.iiit@gmail.com, [2]abass@petronas.com.my, [3]yunusna@petronas.com.my

ABSTRACT

Human-human communication in social environment is only possible through speech, facial expressions and bodily changes. In this research work, speech and facial expressions is used in order to estimate basic emotions (angry, sad, happy, boredom, disgusting and surprise**)** using a computer. A new asynchronous hybrid technique through Relative Bin Frequency Coefficients (RBFC) and Relative Sub-Image Base (RSB) is used in order to deal with the above modalities which are different in time, scale and dimension. Support Vector Machine (SVM) is used for classification. From experimental results, the proposed system performs better than the conventional systems.

Keywords: *Relative Bin Frequency Coefficient (RBFC), Relative Sub-Image Based Coefficients (RSB), Human Computer Interaction (HCI), Support Vector Machine (SVM), Principal Component Analysis (PCA).*

1. INTRODUCTION

Human-human or human-computer interaction in social environment is possible through speech, facial expressions and body poses. According to Darwin theory [1], identification of human emotions in social life is mainly possible thorough facial expressions. It is easy to predict emotions through facial expressions instead other modalities. However, human communication is a mixture of deliberate facial expressions and facial expressions during speech. Identification of human emotions through deliberate facial expressions is easy to detect using a computer than the facial expressions during speech. On the other hand, identification of deliberate facial expressions is different than identification of facial expressions during speech. This problem degrades the performance of the emotions detection through facial expressions. In order to counter the above issue, multimodal system is needed. Human body poses in social communication is culture dependent. Thus, in this research work, speech and facial expressions are considered in order to detect emotions.

Human emotions are infinite. Emotion can classify into two types namely, positive and negative emotions. Positive emotions are good for society. Identification of emotions is possible in many ways as follows: electrocardiogram (ECG), electroencephalogram (EEG), epidemic changes, speech or facial expression etc. However, natural communication is always possible through speech and facial expressions. But, identification of every human emotion is difficult through above two modalities. Ekman et al. [11] stated that, identification of basic emotions (angry, sad, disgust, fear, happy and boredom) are possible through speech and facial expressions. In addition, mixture of basic emotions helps to estimate other emotions.

Human speech contains two types of data namely, verbal and non-verbal. Verbal data gives meaning of the speech. Picard et al. [2] stated that, identification of emotions through meaning of the speech from verbal data is error prone. It creates misunderstanding in order to estimate emotions. To counter the above issue, this research work focused on non-verbal data to estimate emotions effectively. Non-verbal data tells us how the way the words were spoken.

Human emotions are asynchronous in nature which is shown in Figure 1. Arousal of emotion is not always possible to predict through both speech and facial expressions at a time. For instance, a person is angry. It is not always possible to understand through his facial expressions and speech at same time. Due to that, the performance of multimodal emotion detections through speech and facial expressions using feature level fusion decreases. On the other hand, if data is synchronous in nature, this will decrease the performance of multimodal emotion detection through speech and facial expressions using decision level fusion. In order to counter the issue, the research work introduced proposed hybrid system using Relative Bin Frequency Coefficients (RBFC) and Relative Sub-Image Based (RSB) features through speech and facial expressions respectively.

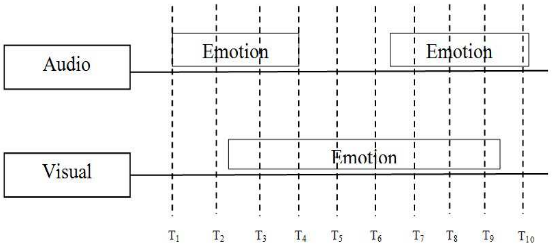


Figure 1. Time frame of emotion detection in different modalities

The rest of the paper is as follows: Section 2 shows the review of previous work in feature extraction and information fusion for the above modalities, Section 3 shows the proposed hybrid emotion detection system. Section 4 and 5 reports the results and discussion. Furthermore, Section 6 presents the conclusion and future work.

2. RELATED WORK

2.1 Feature Extraction

Visual and audio signals are different with respect to scale, dynamics and temporal patterns. Thus, feature extraction of visual data is different from audio data. In order to deal with the above modalities, a suitable fusion technique is needed. The above issues are discussed in subsections as follows.

2.1.1 Image feature extraction

Visual Feature extraction for 2D based methods is three type namely geometric based, appearance based and hybrid. Geometric based approach is mainly focused on complete face to estimate emotions, while appearance based approach is focuses on sub-parts like the corner of an eye, eyelids etc. Finally, the hybrid approach is a combination of the both geometric and appearance based approaches.

Geometric based approach is an optical flow which helps to estimate motion of the data. This approach is mainly dealing with complete facial area including textured areas such as cheeks and forehead. Principal component analysis [PCA] is a dimensionality reduction technique, which also useful as a geometric based feature on gray scale images for dimensionality reduction. Through PCA, the emotion detection system is able to estimate better due to its nature. PCA contains loss of data problem [3]. Gynander Verma et al. [4] stated that, wavelet and curve let transform is used as a geometric based features and showed better classification accuracy for emotion detection system through speech and facial expressions. However it is not suitable to detect adjacent emotion detection problem. Linear discriminate analysis (LDA) useful to detect emotions, but it decreases the overall efficiency of the system due to loss of data. Punitha and Geetha [5] used as a geometric based approach through gray level co-occurrence matrix (GLCM) and achieved better accuracy. In this case, the accuracy of the system depends on the resolution of the input image due to adjacent emotion detection problem. It's due to GLCM features.

From the above discussion it is concluded that geometric based approaches are not suitable for real time applications due to partial face occlusion problem. In order to counter the problem appearance based method is needed. Vukadinovic and Pantic [6] estimated human emotions through appearance based method. Here, the human face detected by Viola and Jones detector. The next is the detected face is divided into 20 regions of interest as follows: eyes, nose etc. The facial feature point of each region is computed with gray level summation and Gabor wavelet features. This system contains loss of data problem which decreases the performance of the system. Padgett and Cottrell [7] used Eigen vectors for the facial data. Murthy and Jadon [8] used appearance based Eigen faces in order to predict emotions. Eigen faces produces less accuracy due to loss of data. It also has adjacent emotion detection problem. Pantic and Patras [9] stated that the geometric based approach performs better than appearance based approach while dealing with emotions. However he also stated that, both approaches have merits and demerits. Thus, hybrid method came into the picture.

Pandzic and Forchheimer [10] showed a hybrid facial animation parameter [FAP] in order to identify facial muscle moment to detect emotions through facial expressions. The parameter location of the FAP is fixed and helps to get clear data of face muscle moments. In this case, a total of 84 locations are identified on the human face in order to cover full face to predict facial expressions. According to the moment of FAPs, the system detects the basic emotions. The major disadvantage of this method is the locations of FAP's, which renders this system not useful for real-time emotion detection systems. Ekman and Friensen [11] detected human emotions using action units (AUs). Facial expression of a person depends on his tradition and culture. Due to that, the method mainly focuses on the specific part which shows basic emotions. However the presence of skin fat of human body prohibits the identification of the actual facial expression. On the other hand, this method also requires manual interaction. Thus, this method is not suitable for real time systems.

In this research work, we are using the Relative Sub-image Based (RSB) for real time conditions in order to address the problem of loss of data. This is a appearance based feature extraction which helps to deal with the fast variations in shape, texture and color of facial expression as a linear combination of sub-images. It also useful to create a compact model for large training set in shape, texture and color. See Section 3 for more details.

2.1.2 Speech feature extraction

Acoustic feature extraction is two types namely direct feature extraction and indirect feature extraction. Direct feature extraction belongs to non-prosodic features namely fundamental frequency, energy and pitch etc. Indirect feature extraction belongs to prosodic features namely mean, median and range etc. Non-prosodic and prosodic features are frequency domain dependent features. As such, the efficiency of the system fully depends on the length of the input speech signal. This creates feature extraction delay which causes synchronization problem. Thus, this is not suitable for multimodal systems.

Binbin and Fengqin [12] introduced bimodal emotion detection system through speech and facial data, and prosodic features are used for acoustic data. The system was achieved better accuracy. However, due to the dependency problem with frequency domain, the acoustic feature extraction became more delayed. As the result, the entire efficiency of the bimodal system decreased. Busso and Lee [13] showed a comparative study with pitch and energy features. The features where belong to non-prosodic feature class. In this case, the aggressive emotion, which is named as the angry class, is identified by pitch and energy features but due to the dependency problem, the overall performance of the system decreases. Banse and Scherer [14] stated that prosodic or acoustic feature shows better accuracy while predicting emotions but less efficiency.

In order to counter the above issues, Relative Bin Frequency Coefficients (RBFC) is introduced for emotion detection through speech. This is a frequency domain independent technique and also helps to increase the overall system accuracy and response time in multimodal systems. For more details, see Section 3.

2.1.3 Information fusion

To date, a considerable amount of research work has been done with multimodal fusion methods to deal with multi domain data. Fusion methods are three types viz. rule based, classification based and estimation based. Researchers mostly use the rule based methods in order to deal with multimodal emotion detection systems to obtain better accuracy. So far, in the emotion detection domain, many researchers have used three fusion techniques namely feature level fusion, decision level fusion and hybrid.

Snoek and Worring [15] stated that feature level fusion which gives the better results for class dependent correlation between the features of different domains. The feature level fusion technique is primarily focused on the type of features used and the correlation between them. However, increasing the number of features always decreases the efficiency of the emotion detection system. Another disadvantage of the feature level fusion is it is significantly affected by environmental noise.

Atrey and Kankanhalli [16] stated that decision level fusion technique provides robustness to emotion detection system. The efficiency of the system does not depend on the type of features used or also the correlation among them. Decision level fusion gives scalability to the bimodal system through speech and facial expressions by allowing the use of a suitable method for a modality to increase the system accuracy. However, the disadvantage of this decision level fusion technique is it is less efficient due to increased time complexity of local decisions from the above modalities which affects the final decision time.

Xu and Chua [17] introduced a hybrid system using feature level and decision level techniques to overcome the above problems. This hybrid system is able to provide better accuracy than both unimodal and bimodal emotion detection systems through speech and facial expressions. Compared to the unimodal emotion detection systems, this hybrid system is complex in nature.

Even though the hybrid system accuracy is increased, the response time is decreased, which reduces the overall performance of the hybrid system. Thus, this hybrid system through speech and facial expressions is not suitable for real time applications. According to the above studies, multimodal emotion detection systems are able to give a better accuracy than unimodal emotion detection systems, but the latter are less efficient and have a slower response time due to their complex nature.

3. PROPOSED WORK

The new method being proposed is a technique aimed at solving the problem of multimodal emotion detection through speech and facial expressions in real time. This new method implemented as follows:

3.1 SYSTEM ARCHITECTURE

Figure 2 shows the input video file of 8 seconds duration. According to the given file, the facial expressions and speech modalities are correlated. Therefore, the two modalities need to process this correlated data within the same amount of time so that the fusion between the both modalities can be possible without any delay. This file shows facial expressions during speech. In this case, the emotion of the actor changes every second as shown in Figure 3. In order to estimate the emotions accurately, the system observation time should not be more than 1 to 3 seconds.

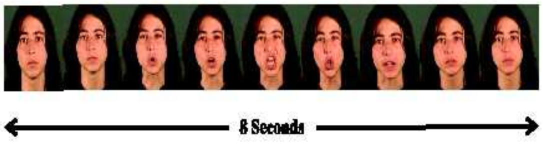


Figure 2. Video file of 8 seconds duration

The proposed emotion detection system is created to estimate emotions from asynchronous data such as shown in Figure 3. Here, this video file in Figure 4, contains 30 frames per second

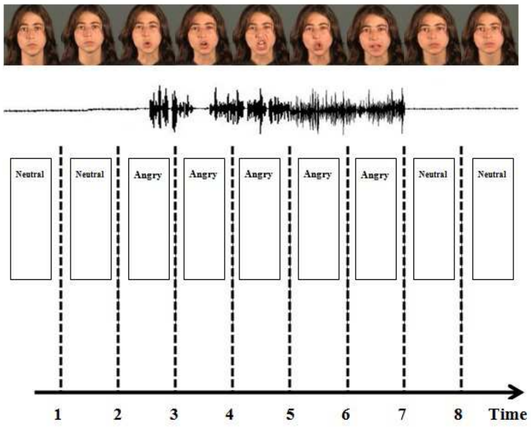


Figure 3. Data availability between two modalities

. The proposed bimodal emotion detection system through speech and facial expressions has three stages namely, feature extraction, classification and fusion. Each stage of the system contains delay, and the delay of one stage of the system can propagate to the subsequent processing stages, which eventually will cause the entire system to be delayed. For instance, when the feature extraction stage of speech and facial expression is being delayed, automatically the classification stage of the modalities needs to wait for some time as shown in Figure 4. In addition to that, each classifier (speech classifier, image classifier and speech – image classifier) also contains its own delay. The speech and image classifiers contain different delays because of different number of input vectors through different modalities. Due to that, this causes problem at the speech-image classifier. In this case, the speech-image classifier has a combination of two different feature vectors which are of different sizes, dynamics and temporal patterns. Due to this, the classification is expected to be difficult, and also will consume more processing time than the other classifiers. In the proposed emotion detection system, this problem is addressed by performing classification based fusion using SVM at the semantic level [20]. Furthermore, all outputs from the classifiers with respect to time, namely speech classifier, image classifier and speech-image classifier become an input to the decision level stage which is illustrated in Figure 6.

At the decision level fusion technique, all outputs of the above three classifiers are combined together to generate the final output of the system. This will be discussed in section 5. The number of classifiers which is needed is chosen by using select lines as shown in Figure 4. In general, increasing the number of classifiers always increases the overall system complexity. Thus, effective utilization of the classifiers using select lines supports to reduce the system complexity as shown in Table 2. Here, availability of speech and image data are represented in the form of 1's and 0's respectively. 1 represents the presence data and 0 represents absence of data. The classifiers which are shown in Table 2 are speech classifier, image classifier, and speech-image classifier. If there is no data for speech and image modality, then no classifier will be active. On the other hand, if there is no data for speech modality, but the data is available for image modality then both classifiers (image classifier and speech-image classifier) will be active. If data for speech modality is available but there is no data for image modality, then speech and speech-image classifiers will be active. Let us consider when data is available for both modalities at the same time then all the above classifiers will respond to the overall decision process. If no data from both modalities, then Buffering technique helps in order to deal with asynchronous data such as shown in Table 1 as follows: S-Buffer and I-Buffer from Figure 4 receives data directly from RBFC and RSB respectively. When the data from any of these features are not available, then the buffering technique uses the means of previously seen samples every time a sample appears, up to some limited amount of time.

3.2 Feature Extraction

3.2.1 Visual analysis

Darwin et al. [1] stated that, the estimation of emotions by computer during human-computer interaction is possible mainly through facial expressions. Human articulates facial expressions through the existence of 53 facial muscles which are covered by skin with fat content. The skin tone of human face varies from one person to another and so does the fat content. Due to that, the facial expression of one emotion to every person cannot be same. In addition to that, human facial expressions are also cultures dependent in nature. Taking the above issues into consideration, this research work considered the full face of human instead of a few selected regions on the face. To do so, an experiment was conducted by observing the full face and masking a few regions on the human face using recorded video images. From this experiment, the observers experienced difficulty in order to identify emotions when some regions on the face is covered namely eye, mouth, eyelids and cheeks etc. Besides facial expression of a emotion, the color of the skin is also found to affect emotion detection; thus skin color becomes the extra dimension on order to predict emotions. In the experiment, the observer marked emotions randomly and achieved less accuracy. Thus, it is concluded that the estimation of human emotions through facial expression is better performed by using the full face.

Human emotions through facial expressions are fast in nature. The human face consists of more than 45 muscles for expressing emotions. Thus, instead of focusing on a few features like eyes, nose, mouth, eyebrows, etc., the proposed RSB [18] features focuses on the complete face. The basic steps for proposed facial feature extraction are as follows:

1) The middle frame is selected from 30 frames per second input video file.
2) The face region is detected and cropped using Viola-Jones algorithm.
3) Resize the cropped file.
4) RSB is extracted from the above region.

In order to estimate emotion through facial expressions, the proposed system needs to estimate the human emotion for each second. In order to do so, firstly the input is a video file which contains frame frequency of 30 frames per second. Viola-Jones algorithm is used to identify and crop the face region from the middle frame among every 30 frames of the file. The cropped region is then resized into 500 X 500. The resized region is again divided into 5 X 5 sub-images in order to extract the features; the steps are illustrated in Figure 5. Finally, the edge difference between every two sub-images through all possible combinations is done using Eq. (1).

$$Difference = [\frac{(X_2 - X_1)}{(X_1 + X_2)}] \quad (1)$$

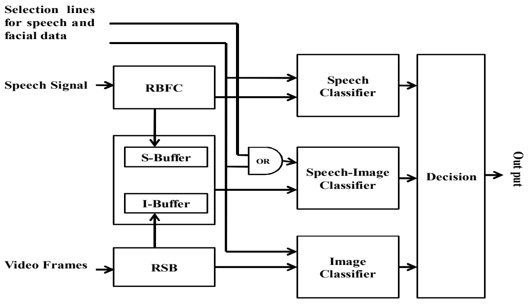
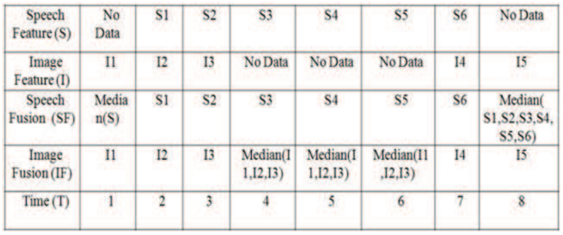


Figure 4. Proposed bi-modal emotion detection system

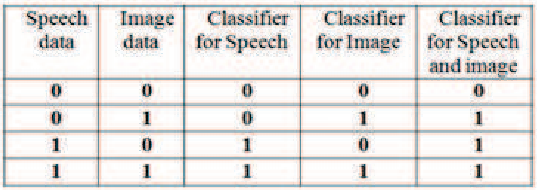
This output vector consists of 300 (25_{C_2}) elements. This vector will go to SVM classifier as an input for estimating emotions. These RBS feature contain less time complexity compared to the PCA, Gaussian wavelet etc. The time complexity of RSB feature at the worst case is O (MN). In this case, M and N represents rows and columns of the input image respectively.

Table 1. Buffering technique for asynchronous data



Speech Feature (S)	No Data	S1	S2	S3	S4	S5	S6	No Data
Image Feature (I)	I1	I2	I3	No Data	No Data	No Data	I4	I5
Speech Fusion (SF)	Median(S)	S1	S2	S3	S4	S5	S6	Median(S1,S2,S3,S4, S5,S6)
Image Fusion (IF)	I1	I2	I3	Median(I1,I2,I3)	Median(I1,I2,I3)	Median(I1,I2,I3)	I4	I5
Time (T)	1	2	3	4	5	6	7	8

Table 2. Effective usage of classifiers



Speech data	Image data	Classifier for Speech	Classifier for Image	Classifier for Speech and image
0	0	0	0	0
0	1	0	1	1
1	0	1	0	1
1	1	1	1	1

3.2.2 Speech analysis

The presence of emotional data in the speech file is possible in two levels, namely inter level and intra level. The inter level data means it exists between the variations of the pitch and energy contours, while the intra level data means it exists between the shape and duration of the glottal pulses and vocal fold. This research work is focused on both inter level and intra level subject variations to identify emotions from speech. Robust emotion detection through speech is only possible with frequency independent feature extraction. To do so, the proposed RBFC [19] is extracted as shown in Figure 6. The basic steps for proposed speech feature extraction are as follows:

1) Read speech file.
2) Pre-emphasize the signal.
3) Divide the signal into frames size of 20 ms and shift of 10 ms, and compute the frequency spectrum.
4) Compute RBFC from the frequency domain data.

In this experiment, the input signal is of fixed duration i.e. 1 second. Next, the given speech signal is converted into the form of frequency spectrum, and then the frequency spectrum is divided into 34 Bins as shown in Figure 6. In this case, each Bin size is fixed to 100 Hz. The next, average of each Bin is calculated. Then finally the relative difference between each Bin is calculated using Eq. (1). An example of the given signal is given in Figure 6. Here, each node represents an average of Bin in the frequency domain. The connection between each node represents the relative difference with other nodes. The total number of Bins should be 34 but the example given in Figure 7 is a model of 6 Bins.

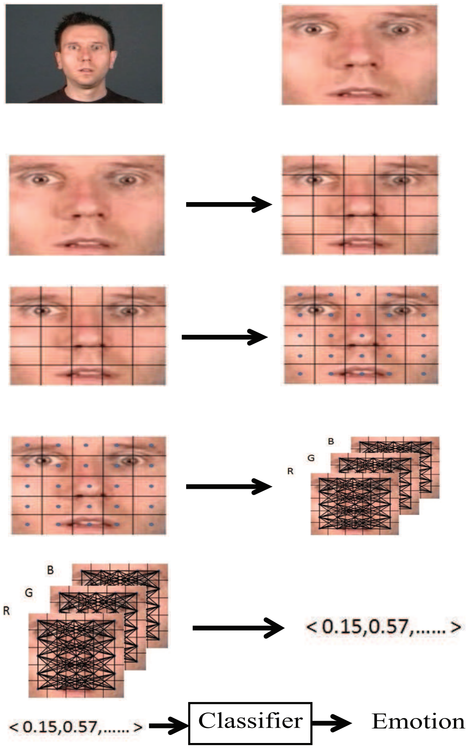
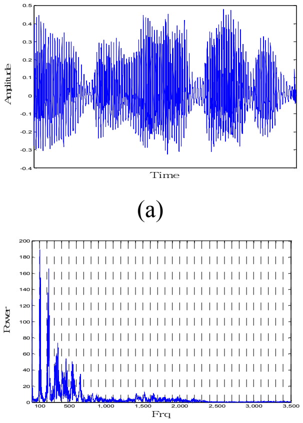
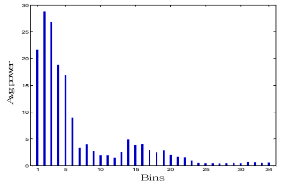


Figure 5. Relative Sub-imaged based features



(a)

(b)



(c)

Figure 6. (a) Preprocessed speech signal (b) Power spectrum with frequency bin sized 100 Hz (Frq range 100Hz-3500 Hz) (c) Avg of bin (total of 34 bins)

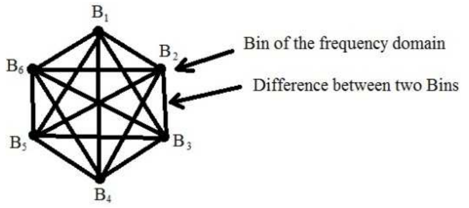


Figure 7. Normalized model for the given signal using 6 Bins

A unique characteristic of this proposed feature is that it provides resistance to intra- and inter-subject pitch variation to the speech signal. This is a frequency independent feature extraction, so that it is able to create a normalized modal for short input signal, and it is easy to implement. This feature helps to create a compact model with less time complexity and there is no loss of data. The time complexity of the RBFC at the worst case is O (NlogN). In this case, N represents the number of samples in frequency domain.

3.3 Emotion Database

There are three methods to design database namely, self reporting design, judgment design and labeling design. In the self reporting approach, the participant needs to report his feelings. However, human emotions are easy to feel but, hard to define. Thus, self reporting becomes error prone. The next is, judgment approach. This judgment depends on the specific cues from professional observers. This judgment approach describes the appearance rather than meaning of the emotion. Finally, labeling approach made from ordinary individuals. In this labeling approach, the process has matched with the experience of daily human life. In this approach, the particular database is again divided into three types, namely induced, acted and naturalist. Naturalistic type database is the best one among all. However, creating this naturalistic database is more challenging due to its many requirements such as expensive equipment is needed and copy rights problem etc. On the other hand, induced type database mostly suffers due to the lack of certainty towards specific emotion. For those reason, most researchers prefer acted databases with the help of professional actors to evaluate their system performance.

In this research work, eNterface ’05 [22] is used which was created by professional actors. The eNterface’05 audio-visual database has 6 basic emotional classes. This database contains the video clips of 10 to 45 seconds duration were taken. The video clips were recorded by 800,000 pixel mini-DV camera and the audio clips were recorded by the directional HF microphone to avoid noise. In this database, each video clip has a size of 30 frames per second. Each frame is of JPEG format type. The size of each frame in the video is 640X480 pixels. Each audio file is in WAV format. Each audio file has 16 KHz sampling frequency. eNterface’05 is an English version database. This English database contains 42 subjects from 14 different nationalities. In the database the actor’s age is varies between 25 to 35 years old. This database is comprised of 81% male and the remaining 19% is female actors. 31% of the actors wore glasses and remaining 17% of the actors have beards. Similarly DaFEx emotional database [21] is also used for testing. This emotional database contains 1000 video files of JPEG format. The time duration of each video file is 8 to 10 seconds. A total of 8 Italian actors (4 males and 4 females) were participated in the making of the video.

3.4 Classification

Support Vector Machine (SVM) [20] is a popular classifier, which not only works well with known samples, but also able to work well on unknown input samples. The main idea behind this SVM is to consider a few class means as a typical model, and assess the contribution to split the cases as far away as possible. This is possible through a training set. According to the statistical learning theory [20], larger margin always shows a better ability to separate two classes in the SVM. In addition to that, the radial basis kernel is used to keep complex data in higher dimension space. Furthermore, the performance of the SVM classifier depends on the type of kernel used.

3.5 Decision

The final decision is made with the help of rule based approach which is shown in Figure 8. In the decision level fusion technique, the final classification scores are combined together and the maximum score is considered as the final output of the emotion detection system. In this case, $[S_1, S_2, S_3,..S_6], [C_1, C_2, C_3,..C_6]$ and $[E_1, E_2, E_3,..E_6]$ are the individual scores for each input from the speech-classifier, image-speech classifier and image classifier respectively. $ADD_1, ADD_2, ADD_3,...ADD_6$ gives the total scores of all basic emotional classes namely, angry, disgust, fear, sad, happy and surprise for each input signal from all the above classifiers. Finally the one which has the highest score among all responses is considered as the final emotion or feeling.

4. RESULTS

4.1 Development of Emotional Models

SVM with radial basis kernel is used to estimate emotions from facial expressions and speech. These models contain the information about the emotion of a person using feature vectors. The development of the models is as follows:

4.1.1 Emotional models through facial expressions

This was created for each emotional class separately with the video data from DaFEx database. In this data set is 4,800 ((8 actors x 10 samples x 3 mins x 60 seconds)/3 seconds) number of samples were used to training the system. These feature vectors were given as the input to the SVM for training. In this research, the performance rate of SVM is analyzed with the help of different kernels. Finally, it is observed that, SVM with radial basis kernel gives better emotion recognition accuracy compared to other kernels. Thus, this research work considered SVM with radial basis kernel for developing emotional modals.

4.1.2 Emotional models through speech

In this case, the speech emotional modal was created with the help of audio data from the DaFEx database. Here also, each feature vector was created by proposed RBFC features for every 1 second clip. This proposed feature helps to deal with both inter and intra subject variation in the given input speech signal. In this data set, a total of 4,800 ((8 actors x 10 samples x 3 mins x 60 seconds)/3) number of feature vectors were used to generate the modals. Furthermore, SVM with radial basis kernel was used for classification and found to perform better than the other kernels.

4.2 The Procedure of the Experiment

Step 1: Initially the training data set is for facial expressions were used as a set of annotated images, which was achieved by identifying and cropping the face part in the training set. This process can be done by either manually or by using Viola-Jones algorithm to create a basic modal to detect emotions from the input data.

Step 2: Then extract the training data set using proposed RSB and RBFC features from the facial expressions and speech respectively.

Step 3: Next, train the SVM with radial basis kernel using training samples from both modalities.

Step 4: Finally, classify the input video to recognize the emotions.

4.3 Evaluation of Emotional Models

The evaluation of the emotional modals from video database was as follows: the test set was segmented into 5 sessions, into pieces of 1,3,5,7 and 10 seconds duration. In this case, the total number of test data per emotional class is 3600 ((20 files x 3 min x 60 seconds)/ 1 second), 1200, 720, 514 and 360 for 1, 3, 5, -, and 10 seconds segments, respectively. From the above, it was observed that 3, 5, 7, and 10 seconds segments showed higher performance than the 1 second segment. Thus, in this research work, 3 seconds time duration is used for system evaluation studies. The experiment was conducted for both person dependent and person independent, because the identification of emotions of a known person is always practically not possible.

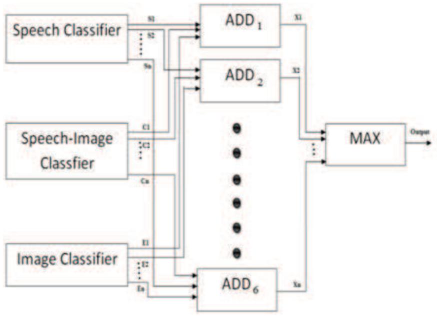


Figure 8. Decision level fusion method

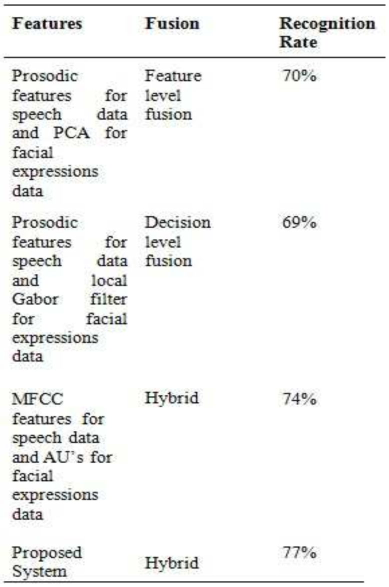
Table 3. List of Databases for Experiments

Experiment	Database	Samples
1	eNterface'05	600
2	DaFEx	600

Tables 4 and 5 show the performances for experiment 1 as person dependent and independent approach, respectively. eNterface'05 database is used to obtain the experimental results shown in

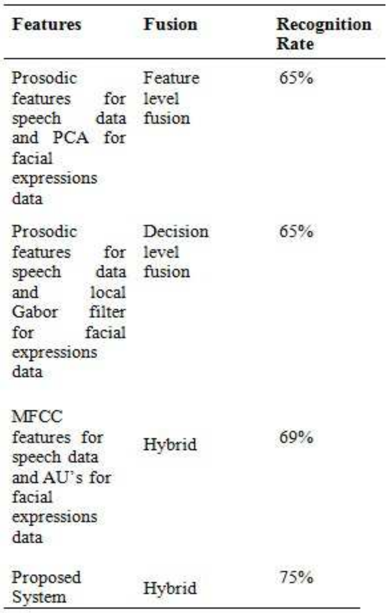
Table 4. A 10-fold cross validation is used for validation. 90% of the data is used for training and 10% for testing. All of the experiments were conducted in MATLAB 7.0. According to the experiment 1, the newly proposed emotion detection system achieved better accuracy and user response time compared to the existing systems.

Table 4. Experiment 1 as person dependent Test with eNterface '05



Features	Fusion	Recognition Rate
Prosodic features for speech data and PCA for facial expressions data	Feature level fusion	70%
Prosodic features for speech data and local Gabor filter for facial expressions data	Decision level fusion	69%
MFCC features for speech data and AU's for facial expressions data	Hybrid	74%
Proposed System	Hybrid	77%

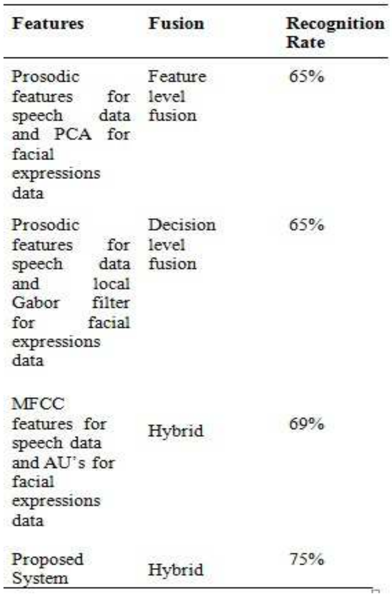
Table 5. Experiment 2 as person dependent test with DaFEx



Features	Fusion	Recognition Rate
Prosodic features for speech data and PCA for facial expressions data	Feature level fusion	65%
Prosodic features for speech data and local Gabor filter for facial expressions data	Decision level fusion	65%
MFCC features for speech data and AU's for facial expressions data	Hybrid	69%
Proposed System	Hybrid	75%

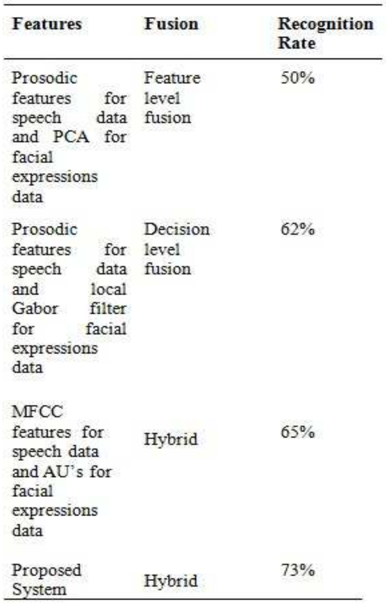
Similar to the previous experiment is as shown in Table 4, here also Table 6 and 7 show the results of experiment 2 as person dependent and independent test. The database, DaFEx is used to obtain the experimental results shown in Tables 6 and 7.

Table 6. Experiment 1 as person independent Test with eNterface'05



Features	Fusion	Recognition Rate
Prosodic features for speech data and PCA for facial expressions data	Feature level fusion	65%
Prosodic features for speech data and local Gabor filter for facial expressions data	Decision level fusion	65%
MFCC features for speech data and AU's for facial expressions data	Hybrid	69%
Proposed System	Hybrid	75%

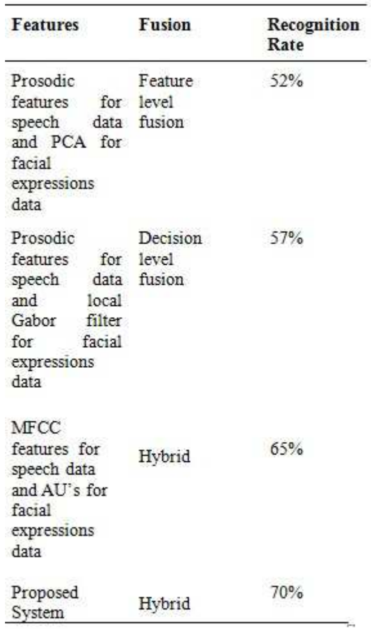
Table 7. Experiment 2 as person independent test with DaFEx



Features	Fusion	Recognition Rate
Prosodic features for speech data and PCA for facial expressions data	Feature level fusion	50%
Prosodic features for speech data and local Gabor filter for facial expressions data	Decision level fusion	62%
MFCC features for speech data and AU's for facial expressions data	Hybrid	65%
Proposed System	Hybrid	73%

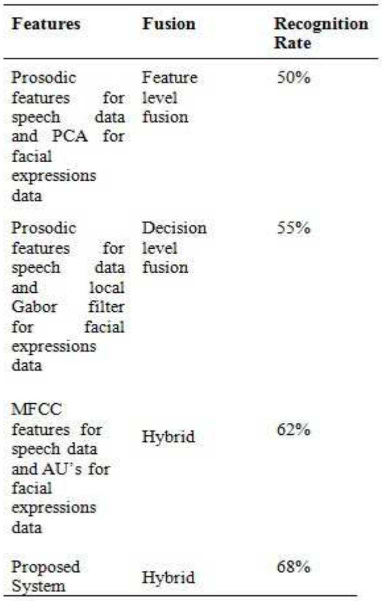
Emotion detection for real-time data is shown in Tables 8 and 9. From the results, it is observed that the proposed emotion detection system performed better even for real time data.

Table 8. Real time person dependent test



Features	Fusion	Recognition Rate
Prosodic features for speech data and PCA for facial expressions data	Feature level fusion	52%
Prosodic features for speech data and local Gabor filter for facial expressions data	Decision level fusion	57%
MFCC features for speech data and AU's for facial expressions data	Hybrid	65%
Proposed System	Hybrid	70%

Table 8. Real time person independent test



Features	Fusion	Recognition Rate
Prosodic features for speech data and PCA for facial expressions data	Feature level fusion	50%
Prosodic features for speech data and local Gabor filter for facial expressions data	Decision level fusion	55%
MFCC features for speech data and AU's for facial expressions data	Hybrid	62%
Proposed System	Hybrid	68%

A 10-fold cross validation is used for system validation. 90% of the data samples were used for training and 10% for testing. All of these experiments were conducted in MATLAB 7.0. According to experiment 2, the proposed emotion detection system achieved higher accuracy in comparison to the conventional systems. Figure 9 and 10 give compare the performance of the proposed emotion detection system against that of conventional unimodal, bimodal and hybrid for each class of emotion. From Figure 9 and 10, it is observed that the proposed hybrid system achieved the higher accuracy among all.

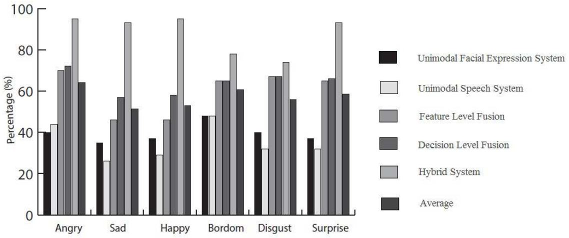


Figure 9. Results of eNterface '05

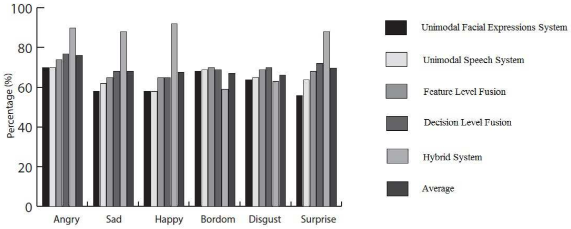


Figure 10. Results of DaFEx Database

Angry, Sad and Happy emotional classes achieved better recognition rate compared with the other classes. On the other hand, the performances of Disgust, Fear and Surprise changes between databases which is discussed in Section 5. Thus, more training samples are needed to increase the accuracy of the hybrid system.

5. DISCUSSION

In this research work, various conventional feature extraction methods have been investigated for estimating human emotion from facial expressions and speech data. The RSB feature is extracted on complete face instead of few regions. This proposed feature extraction does not involve complex transformations. Simple operations using pixel intensities are used to create the model for facial expression from facial muscles. The results show that the estimation of facial expression is better through full face instead of few regions. Based on the results, it is observed that angry, happy and sad emotion gives better accuracy. This indicates that proposed feature extraction of the full face helps to give a better description of human emotion through facial expressions. The literature review shows that facial features extraction has been done with the help of many conventional features, namely, PCA, Gaussian wavelet and Gabor energy filter. All the above conventional features are not able to deal with adjacent emotions. The advantage of the proposed RSB features are simple in nature, easy to extract and gives higher performance when compared to the other conventional features.

Emotion detection through speech is possible with the help of the spectral and temporal features. The RBFC is based on spectral feature. This spectral feature shows the shapes of the vocal track for each emotional class. Temporal feature is useful for time domain processing. Similarly, emotion detection is also possible through prosodic features. Prosodic features need long input signal which helps to focus on local and global level for a word and sentence, respectively. This provides language dependent feature extraction. In this research work, prosodic feature extraction is not used due to its large dependent nature. The proposed RBFC features enable the extraction of inter and intra subject variations within a short amount of time. Inter and intra subject variation of the speech signal helps to detect specific group of emotions. Due to that, the proposed RBFC extracts both inter and intra subject variations together to estimate emotions. From the experimental results, it is observed that, the performance of the proposed RBFC is better than the other conventional features.

For real-time conditions, the proposed emotion detection system is suitable due to its less computational time. Due to the proposed system with RBFC and RSB through speech and facial expression data, the synchronization between the above modalities becomes possible. So, the fusion between the modalities also takes place within time.

The use of spectral type feature does not reduce the recognition accuracy of the proposed emotion detection system which is attributed to the spectral level RBFC which is able to handle both inter and intra subject variations.

Finally, the fusion of above modalities namely, facial expressions and speech significantly improves the overall performance of the emotion detection system. In this research, 6 basic emotional classes are considered. But, in real-life condition, human emotions are not limited. They are complex in order to detect. They depend on culture and language. In addition to that, the identification of human emotions in the outside environments is not easy due to the environmental noise. Thus, the identification of human emotions remains as a challenging task.

6. CONCLUSION AND FUTURE WORK

A proposed bi-modal emotion detection system through speech and facial expressions has been developed. The proposed hybrid fusion through the RBFC and RSB of human speech and facial expressions has been shown to perform better for asynchronous data. The experimental results have shown improved system accuracy for emotion recognition and response time over the conventional emotion detection systems. The proposed emotion detection approach can be extended to applications such as a secured file handling system, multimedia game, identification of similar audio-video files in multimedia databases and so on.

REFRENCES:

[1]. C. Darwin, P. Ekman, and P. Prodger. (1998). The expression of the emotions in man and animals vol. 1. Oxford, USA: Oxford University Press.

[2]. R. W. Picard and R. Picard. (1997). Affective computing vol. 5. Cambridge, USA: MIT Press.

[3]. Abdullah M, Wazzan M, Bo-Saeed S. (2012). Optimizing Face Recognition Using PCA vol 3, International Journal of Artificial Intelligence and Applications, 23-31.

[4]. Verma GK, Singh BK. (2011). Emotion recognition based on texture analysis of facial expression. International Conference on Image Information Processing, India, 1-6.

[5]. Punitha A, Geetha MK. (2013). Texture based Emotion Recognition from Facial Expressions using Support Vector Machine vol. 80, International Journal of Computer Applications, 0975-8887.

[6]. Viola P, Jones MJ. (2004). Robust real-time face detection vol 57. International journal of computer vision, 137-154.

[7]. Padgett C, Cottrell GW. (1997). Representing face images for emotion classification. Advances in neural information processing systems, 894-900.

[8]. Murthy G, Jadon R. (2009). Effectiveness of eigenspaces for facial expressions recognition vol 1. International Journal of Computer Theory and Engineering, 1793-8201.

[9]. Pantic M, Patras I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. Systems, Man, and Cybernetics vol 36, Part B: Cybernetics, IEEE Transactions on, 433-449.

[10]. Pandzic IS, Forchheimer R. (2002). MPEG-4 facial animation. The standard, implementation and applications. Chichester, England: John Wiley&Sons.

[11]. Ekman P, Friesen WV. (1986). A new pan-cultural facial expression of emotion vol 10. Motiv Emotion, 159-168.

[12]. Tu B, Yu F. (2012). Bimodal emotion recognition based on speech signals and facial expression. Foundations of Intelligent Systems. 691-696.

[13]. Busso C, Lee S, Narayanan SS. (2007). Using neutral speech models for emotional speech analysis. INTERSPEECH, 2225-2228.

[14]. Banse R, Scherer KR. (1996). Acoustic profiles in vocal emotion expression vol 70. Journal of Personality and Social Psychology, 614-636.

[15]. Snoek CG, Worring M. (2005). Multimodal video indexing vol 25, A review of the state-of-the-art. Multimedia Tools Appl, 5-35.

[16]. Atrey PK, El Saddik A. (2011). Kankanhalli MS. Effective multimedia surveillance using a human-centric approach vol 51. Multimedia Tools Appl, 697-721.

[17]. Xu H, and Chua T. (2004). The fusion of audio-visual features and external knowledge for event detection in team sports video. Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval, USA, 127-134.

[18]. Kudiri KM, Said AM, Nayan MY. (2012). Emotion detection using sub-image based features through human facial expressions. International Conference on Computer & Information Science, Malaysia, 332-335.

[19]. Kudiri KM, Said AM, Nayan MY. (2012). Emotion detection using relative amplitude-based features through speech. International Conference on Computer & Information Science, Malaysia, 522-525.

[20]. Shen P, Changjun Z, Chen X. (2011). Automatic speech emotion recognition using support vector machine. International Conference on Electronic and Mechanical Engineering and Information Technology, China, 621-625.

[21]. Battocchi, Pianesi F, Goren-Bar D. (2005). A first evaluation study of a database of kinetic facial expressions (dafex). Proceedings of the 7th international conference on Multimodal interfaces, USA, 214-221.

[22]. O. Martin, I. Kotsia, B. Macq and I. Pitas. (2006). The eNterface'05 Audio-Visual Emotion Database. Proceedings of the first IEEE workshop on multimedia database management, Atlanta, 1-8.