A Multimodal Emotion Recognition System from Video

Thushara S
Department of Electronics and Communication,
Amrita Vishwa Vidyapeetham University
Amrita School of Engineering
Coimbatore, Tamil Nadu, India
[1]thushara.86@gmail.com

Dr. S Veni
Department of Electronics and Communication,
Amrita Vishwa Vidyapeetham University
Amrita School of Engineering
Coimbatore, Tamil Nadu, India
s_veni@cb.amrita.edu

***Abstract* - Emotion recognition (ER) systems finds applications in many fields like call centres, humanoid Roberts and robotic pets, telecommunication, psychiatry, behavioral science, educational softwares, etc., In this work, the speech and facial features extracted from the video data is explored to recognize the emotions. Since both these features are compliment to each other, on combining them will result in higher performance. The features used for emotion recognition from video data are geometric and appearance based while prosodic and spectral features are employed for speech signal. Support Vector Machine (SVM) classifier is used to capture the emotion specific information. The basic aim of this work is to explore the capability of speech and facial features to provide the emotion specific information.**

Keywords— Emotion recognition; Support Vector Machine; Spectral and prosodic features; Facialfeatures; Acoustic features.

I. INTRODUCTION

Emotions are extensively exploited by the human beings for conveying messages. Emotion recognition is a painless task for human beings but for the machine to identify the emotion is challenging. It offers a natural interface between machines and humans, by which the system can understand, interpret and respond to the human emotions. According to physiological theory six widely accepted emotions are happy, sad, surprise, disgust, fear and anger. Emotions can considerably change the sense of the message. From the studies, we came to know that 7% of message is conveyed through spoken words, 38% is conveyed through voice intonation and 55% by facial expressions [1]. When machine recognize the emotion either by facial expression or by speech, they could be used to provide help to the diseased or handicapped people. Automatic emotion recognition systems are used in call centers, where machines are used to handle the customer's queries based on their mood. In investigation department they use this emotion recognition system (ERS) to predict the activities of the criminals by analyzing the taped conversations. It is also used in humanoid Roberts and robotic pets for making them to react similar to human beings [2][3]. ER is also used in telecommunication, psychiatry, behavioral science, educational softwares, etc.

Automatic recognition system by using either facial or acoustic features are explored to maximum, but research on combining these two modalities is comparatively very few. The proposed system processes each model separately and then combine the features at the later stage to get the multimodal. Combination of two modalities will results in better performance and robustness. Facial ERS is grouped as :

- Feature based technique: This technique uses the features like corners of mouth, corners of eyebrows, center of the mouth, center of eyes etc.
- Region based technique: In this method certain region of the face is only considered for emotion recognition

Similarly for audio modality the different feature extraction techniques used are:

- Frame level modelling: In this technique, feature vectors are extracted from the overlapping frames
- Supra-segmental modelling: Here duration sequence, pitch, energy, formant like features are used to extract the emotion specific information

In this paper, we build a multimodal ERS by exploiting the facial and acoustic information. The spectral and prosodic features are exploited to extract the emotion specific knowledge from speech. Variation of vocal tract shape is mainly represented by the spectral features. In this work, Mel Frequency Cepstral Coefficients are used to study the spectral features [4]. Information embedded in pitch, energy and formants are used as prosodic features of speech. Prosodic features used in this work will give information about the duration of the vocal folds and shape of the glottal pulse. Here, the features derived from left eye, right eye and mouth is used to explore the recognizing emotions from face. Also this work deals with the contribution of each feature to recognize emotion from the video data. Final stage of every ERS is the classification part. Among different classifiers Support Vector Machine (SVM) is employed in our work. Simplicity, resistance to noise and flexibility are the attracting factors of using SVM. For estimating the intensity of each emotion we are also using Support Vector Regression (SVR). The rest of the paper is organized as follows. Section II presents a review of related works. The proposed work is presented in Section III. Results and discussion are provided in Section IV and Section V concludes the paper.

978-1-5090-1277-0/16/$31.00 ©2016 IEEE

II. RELATED WORKS

Existing works on ER from video is discussed in this section. The section include following dimensions (i) facial features, (ii) acoustic features and (iii) combined facial and speech features.

A. Facial Features

Many researches are done in the area of emotion recognition from facial features. For recognizing emotions from face one of the most commonly used method is Facial Action Coding System (FACS) [5]. In this method, action units are used as features to get emotion specific information. Recognition accuracy of almost 87% is owned by this recognition system. Main two branches of ER from face are facial affect detection and facial muscular action detection. Geometric and appearance based approaches are the two methods used for analyzing the facial expressions [13]. In computer vision application of face recognition Gabor filter response are mainly explored. Pose-invariant face recognition system by coupled Gaussian is addressed in [7]. Different methods were proposed even to identify pain by ERS. Accuracy of the latter is found to be 81%. Apart from the pain laughter is also identified. By using the head movement and face recognition laughter is detected. This system is made with an overall precision rate of 85%. Facial emotion tracking system is also implemented in real time. Six different emotions are tracked and identified. Multilayer neural network with 3 hidden layers are used for the final detection process. At the initial stage of feature extraction, logarithmic Gabor filters are used. This real time system is having accuracy of 70%. Filko and Yegnanarayana propose a recognition system by employing the principal component analysis for analyzing the key facial regions. Here they used 15 neural networks, 1 network was used for region detection, 14 were used for recognizing 7 universally accepted emotions. On FEEDTUM database this system had an accuracy of 89%. Facial emotion analysis is also done by neuro-fussy system. Both continuous and discreet emotional space can be evaluated by the proposed technique. Efficiency is found to be 78%.

B. Acoustic Features

Recognizing emotions from speech is also a hot topic in computer vision application. A brief review of different techniques used in speech ERS is discussed in this section. Most studies were done by using prosodic features in extracting the emotion information. The proposed system have an average accuracy of about 77% for discriminating the neutral speech from emotional speech. Fundamental frequencies, duration and intensity of pauses are studied by Busso and Lee. 55% accuracy was obtained for recognizing 4 different emotions by using this method. Features used by Nicholson H and Pantic M was based on the phonetics. Performance of this system was only about 50%. Selvi M and Chan studies the Log Frequency Power spectral Coefficients (LFPC) by using discrete Hidden Markov Model to recognize 7 different emotions and the efficiency was found to be 80% [9]. Apart from MFCC, pitch information, formants and log energy are used as features for discriminating 6 emotions by using Fisher Discriminant Analysis. Accuracy between 50-60% were obtained by combining the glottal and spectral features [11]. Modulation spectral features are the other type of features in speech ER. Combining the modulated spectral features and the prosodic features results in an average efficiency of 91.6%. Neural networks were used for recognizing the emotions by Mina Navran and Nasrollah Moghadam Charkari. Here for each emotions they used separate networks [12]. Both prosodic and spectral features were used in this paper. Pitch and energy were used as prosodic features and the LFPC coefficients were used as spectral features. Here the recognition accuracy is 80%.

C. Facial and Acoustic Features

Comparing to the researches made on ERS using either facial features or acoustic features, the studies on multimodal system by combining both features are very few. For predicting the emotions effectively two unimodal system (system by using facial features or by using speech features) are combined at the higher dimension resulting in an accuracy of 87%, 67% and 90% using facial, acoustic, facial plus acoustic features respectively [14]. [15] Apart from these features, thermal variations of the infrared rays of the camera were also combined to get the efficiency increases to 91%. Some other studies were made on using the prosodic features from the speech and the movement of the facial muscles to get the emotion specific information. In this only the best features were used to make the multimodal system which results in the increase of accuracy to 94%. Pitch and energy information are used as speech features and is combined to the motion information from the face by using multistream fused Hidden Markove Model.

III. PROPOSED METHOD

In this work, a multimodal system for automatic ER is proposed. For this first two unimodal systems are made by exploring the facial features and speech features respectively. The most appropriate features of these unimodal systems are used in the proposed system so as to get a higher performance rate. The block diagram of the proposed system is given in fig 1. This section is divided into 3: facial features, acoustic features and multimodal system.

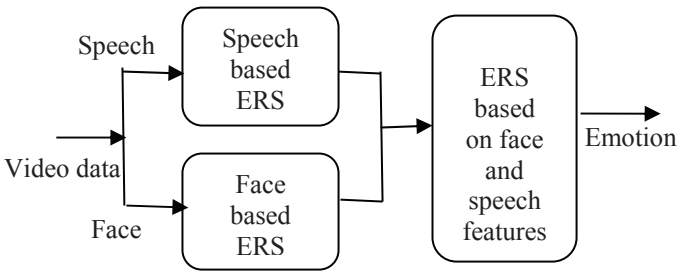
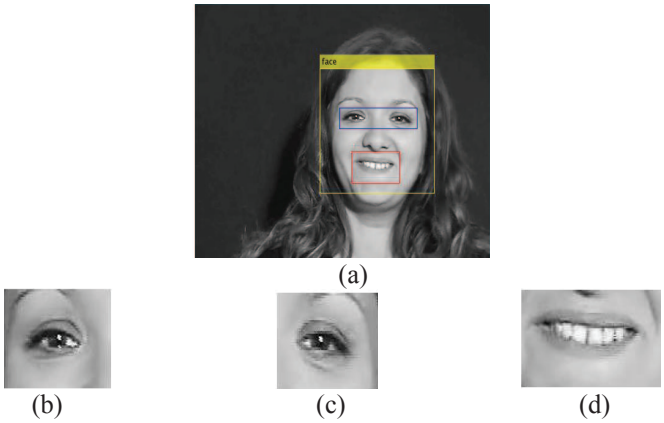


Figure 1: Block diagram of multimodal emotion recognition system

A. *Facial Emotion Recognition System*

Facial ERS is broadly divided into 3 stages: face detection and tracking, facial feature extraction and finally the classification stage. Before the first stage the video data has to be converted into frame. Face detection and tracking is done on that frames to identify the face from input data, align the face and to track the face region. Conventional face detection algorithm has many steps which are as follows: images captured in fixed pose are given as training images. Pre-processing is done over all these images to dominate the effect of light illumination. After pre-processing, extraction of some face samples is done. Mainly there are two type of extraction-knowledge based method and learning based method. In knowledge based method, face patterns are modelled by definitive rules like skin colour, face textures etc. while, face patterns are learned from the given data samples by discriminative functions in learning based method [13]. The face detector system then scans the entire image to locate and detect the faces. In our paper we use Viola-Jones Algorithm for automatic face recognition. Cascade classifiers and Adaboost learning method is employed in this algorithm. Viola-jones algorithm is applied over each key frame to detect the face region. After detecting the face region it is again used to get the eye and mouth region. We restrict our research to eyes and mouth region as it is the most expressive region on face and will contribute more to the ERS. The output of the face recognition is given in fig 2 (a). After detecting face, eyes and mouth these regions are cropped, which is shown in fig 2-(b), (c) and (d) respectively.



(a) (b) (c) (d)

Fig. 2. (a) Detected face, (b) Detected Right eye, (c) Detected left eye, (d) Detected mouth

Next stage in this system is the feature extraction which is the most crucial step in ERS. The features used here are geometric features and the appearance based features. In geometric features, the vectors are made by extracting the information of the shape and location of different face components. But this method strongly depends on the efficiency of the detection and tracking stage. Here, in our work we are measuring the size of the detected eyes and mouth region. For that first the corner points are detected and the Euclidian distance between these points are used as the feature. Apart from the corner points the central point of eyes and mouth are also used as features. Next type of features used are appearance based feature. In this approach, Gabor filters are applied over the key frame to extract the relevant features like texture, correlation. It can be applied over the whole frame or over specific region. In the proposed technique, Gabor filters are applied over the eyes and mouth region in order to reduce the dimensionality of feature vector. Filter used here have one scale spatial frequency with 9 different orientations. To reduce the dimensionality further we compute the statistical characteristics like mean, standard deviation for each responses of the Gabor filter. Next the final stage of emotion recognition system is the classification part. We use SVM classifiers because of its flexibility, simplicity and noise resistance. SVM is used to classify 6 different emotions: happy, sad, fear, surprise, anger, disgust. For estimating the intensity of each expression SVM regression can also be done.

B. *Acoustic Emotion Recognition System*

In the proposed method prosodic and spectral features are used to recognise the emotions from the speech. MFCC coefficients are used to study the spectral behaviour while pitch, energy and the formant information are used to get the prosodic features.

1) MFCC: MFCC is used to extract the speech information in the frequency domain. It uses the Mel filter bank to get the parameters in a similar fashion like humans hearing the speech. It will give the details of the changing shape of vocal tract. First the speech is divided into time frames. Overlapping is done on frames to get a smooth transitions. To eliminate the discontinuities at the edges, windowing is applied over the frames. For getting the frequency component of the given signal Fast Fourier Transform (FFT) is applied to the windowed frames. Then the output of FFT is passed through a Mel scaled filter bank. Finally Discrete Cosine Transform is calculated to get the MFCC coefficients. In our work we use frame duration of 20 ms with a frame shift of 10 ms. 13 MFCC coefficients are derived from 20 filter banks.

2) Pitch: Speech mainly includes voiced region, non-voiced region and silence. It is the fundamental frequency and is found only in the voice region of the speech. First to make the signal stationary framing is done. Because of the tapered edges of the hamming window, they are mainly used to eliminate the signal distortion. In order to get the pitch contour points, autocorrelation function is applied over the windowed frames so that it can give similarity between observations as a function of time lag between them. In voiced region of the speech the excitation of pitch contour is found to be periodic while it is random in unvoiced part. During silence period no excitation can be spotted off which is shown in fig 3. Because of this elimination of silence period can easily be done by eliminating the no excited area.

3) Energy Contour: For segmenting the speech signal it is multiplied with the hamming window function. In hamming window, center of the windows contribute more and the bandwidth is twice that of the rectangular window. Proper frame shift should also provide to get smoother transition. With the increase in the frame size, energy contour become smoother because of that distinction between voiced and unvoiced region reduces. After segmentation, squaring of amplitude is done. The signal gone through these step is plotted in fig 4. It is mainly used to eliminate the unvoiced region from speech.

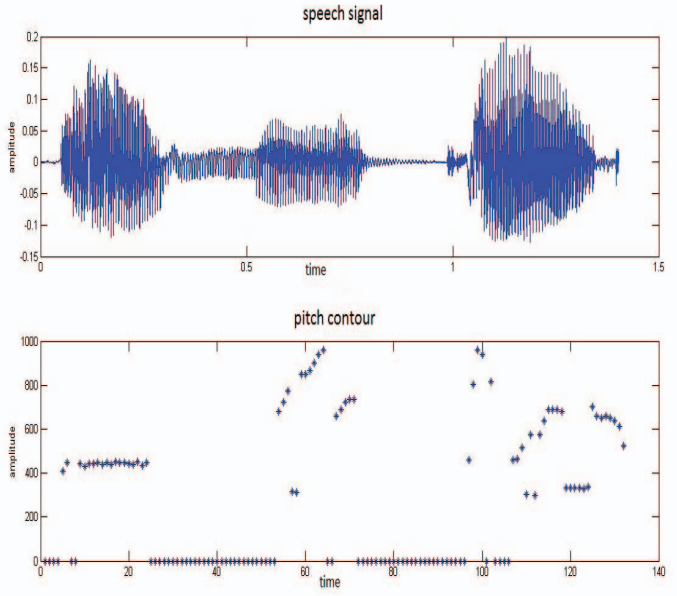


Fig. 3. Input speech and pitch contour

4) Formants: Formants are referred as the resonance frequency of vocal tract. The different steps in finding the formant frequency are given below

- Input speech is pre-emphasised and normalised to 1 and -1 amplitude
- Segmentation of speech signal is done by hamming window.
- Then it is passed through a Butterworth high pass filter to remove the DC values.
- Logarithm of the spectrum is computed so that both large and small values are visible.
- The peaks of the signal obtained will give the formant frequencies.

Fig 5 is the formant frequency plot of our input speech signal. The red cross is the resonance frequency of the signal. After extracting all the features the final step of classification is done. SVM is used for classification. Since the nonlinear characteristic of speech signal nonlinear SVM is employed. Proposed technique also uses a 6 class SVM to classify the 6 universal emotions.

C. Multi Modal System

At the later stages the best features of the two systems are fused to get a multi modal system. Since best features are employed multimodal system has a better accuracy [1]. Multi modal system can be built by two ways either by the fusion of the appropriate features or by taking the match score [7]. From our studies, we came to know that fusing best features outperform the matching score method.

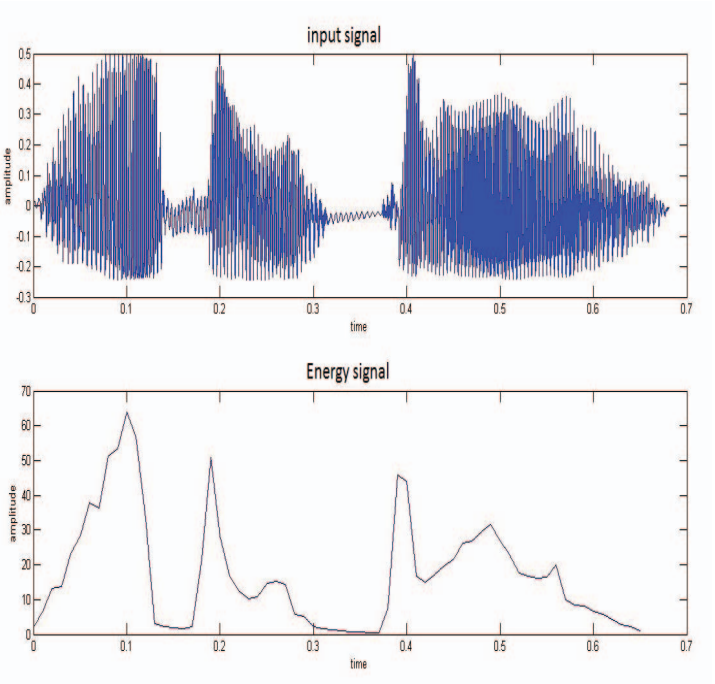


Fig. 4. Input speech signal and energy signal

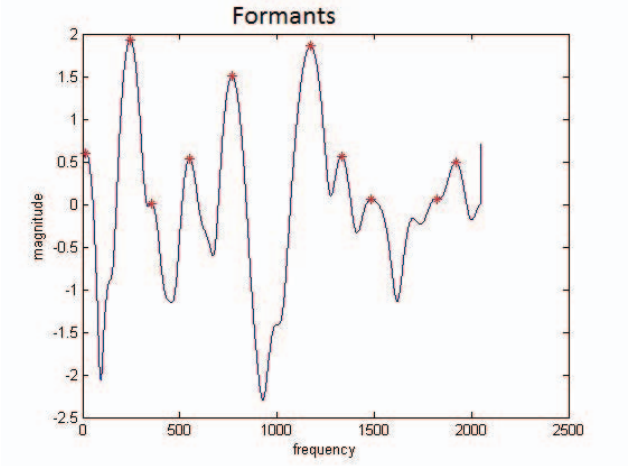


Fig 5. Formant frequency

IV. RESULTS AND DISCUSSIONS

Six emotions (happiness, sadness, disgust, fear, anger, surprise) are recognised by using three different systems based on facial expression, speech model and bimodal information. Emotional video samples from ENTERFACE'05 were given as input. It consist of 44 subjects with 6 different emotions. This system recognise the strength and weakness of the unimodal systems and fuse these two modalities to increase the performance. The facial ERS itself is a

combination of two features (appearance and geometric) similarly acoustic system is also made by combining prosodic and spectral features. Fusing of these features in the unimodal stage increase the overall performance of the system. Because of the nonlinear characteristics of the speech and image data a nonlinear multiclass SVM is used. Polynomial and Radial Basis Function (RBF) kernels gave better results than other kernels with a percentage of 94.765% and 95.415% respectively. 80% of scores were used for training and 20 % for testing. Training data was split into 10 set out of which 9 were used for training and 1 for testing. In the next iteration another 10 set were considered. After classification, intensity is also estimated to get the mental information of the subject. It can be computed by the Support Vector Regression (SVR).

V. CONCLUSION

A study was conducted to implement a bimodal system by employing both acoustic and facial features. The system classify six different emotions by taking the video data as the input. The information from the speech features and facial features are compliment to each other. By combining these two systems the efficiency and performance shoot up. For this model geometric and appearance based features are employed for face. Prosodic and spectral features are studied for speech signal. The classification is done by SVM classifiers because of its simplicity and flexibility.

REFERENCES

[1] K. Sreenivsa Rao and Shashidhar G. Koolagudi, "Recognition of emotions from video using acoustic and facial features", *International journal on Signal Image and Video Processing*, no. 9, pp. 1029-1045, July 2015.

[2] Simon Dobrišek1 and Aurobinda Routray, "Towards Efficient Multi-Modal Emotion Recognition", *Int J Adv Robotic Sy*, vol. 6, no. 1, January 2015.

[3] AlMejrad A. S.." Human emotions detection using brain wave signals: A challenging", *European Journal of Scientific Research*, vol. 44, pp. 640-659, 2010

[4] Jibi Raj and Sujith Kumar, "Gender based Affection Recognition of Speech Signals using Spectral & Prosodic Feature Extraction", *International Journal of Engineering Research and General Science*, vol. 3, March 2015.V.

[5] Pantic and Ptras, "Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences ", *IEEE Trans. Syst. Man Cybern*, vol. 17, 2009.

[6] Filko and Yegnanarayana, "Emotion recognition system by a neuralnetwork based facial expression analysis", *AUTOMATIKA* ***54***, no. 109, pp. 44-55, 2008.

[7] Lee C.M and Narayanan S.S, "Coupled Gaussian process regression for pose-invariant facial expression recognition*", Proceedings of 11th European Conference on Computer Vision* , March 2005.

[8] Busso M and Lee R, " Analysis of emotionally salient aspects of fundamental frequency for emotion detection*", EUROSPEECH* vol. 41, no. 3, pp. 1066–1082, Mar. 2008.

[9] Nicholson H and Pantic M., "Emotion recognition in speech using neural networks." *6th International Conference on Neural Information Processing (ICONIP),*2010.

[10] New O and Selvi M.," Speech emotion recognition using hiddenMarkovmodels*", IEEETrans. Speech Audio Process, vol.* 18, pp. 965–973, 2010.

[11] Falk O, Chan I and Pantic M.," Automatic speech emotion recognition using modulation spectral features. Speech Commun*." In:Proceedings of 11th European Conference on Computer Vision(ECCV)*, 2010.

[12] Nicolson L and Kessous, "Emotion recognition in speech using neural networks.", *2nd InternationalConference on Affective Computing and IntelligentInteraction,* Lisbon, September 2013.

[13] Mina Navran and Nasrollah Moghadam Charkari, "Fusion of Feature Sets for Facial Expression Recognition", *IEEE Transactions on Telecommunication,* vol. 62, no. 6, July 2014.

[14] Yoshitomii H, Kim P, Suteau A, and Allain P, "Effect of sensor fusion for recognition of emotional states using voice, face imageand thermal image of face.", *2012 IEEE International Conference,* pp. 1-8, 2010.

[15] L. Huang, F. Thawn, and L. Didaci, "Bimodal emotion recognition byman andmachine," *Pattern Recognition.*, vol. 42, no. 11, pp. 2807–2817, Nov. 2009.

[16] Vuppala and Yadav, "Vowel Onset Point Detection for Low Bit Rate Coded Speech", *IEEE Trans Audio Speech and Language Processing,* no. 20, pp. 1894-1903, August 2012.

[17] Datcu D, Rothkrantz L," Semantic audio-visual data fusionfor automatic emotion recognition". *In: Proceedings of Euromedia*, 2008.

[18] Vapnik V, "The Nature of Statistical Learning Theory", *IEEE International Conference,* 1999.