# Emotion Recognition by Using Bimodal Fusion

Shruti V. Kulkarni
Electronics and telecommunication engineering,
Siddhant College of Engineering, Sudumbare
Pune, India
*mailshrutikulkarni@gmail.com*

Prof. Savita S. Raut
Electronics and telecommunication engineering,
Siddhant College of Engineering, Sudumbare
Pune, India
*krishivmanu@rediffmail.com*

***Abstract*—** In order to improve the single-mode emotion recognition rate, the bimodal fusion method based on speech and facial expression was proposed. Here emotion recognition rate can be defined as ratio of number of images properly recognized to the number of input images. Single mode emotion recognition term can be used either for emotion recognition through speech or through facial expression. To increase the rate we combine these two methods by using bimodal fusion. To do the emotion detection through facial expression we use adaptive sub layer compensation (ASLC) based facial edge detection method and for emotion detection through speech we use well known SVM. Then bimodal emotion detection is obtained by using probability analysis.

***Index Terms*—***Marr-Hildreth, (ASLC), Hidden Markov Model, MFCC, Spectral and prosodic features, bimodal fusion*

*****

## I. INTRODUCTION

Recognizing basic emotions through speech or facial expression is the process of recognizing the mental state. Facial Expressions are universal language of emotion, instantly conveying happiness, sadness, anger, fear, and much more. Emotion recognition through speech and facial expression is an area which increasingly attracting attention within the engineers in the field of pattern recognition. Emotions play an extremely important role in human life. It is important medium of expressing humans perspective or fillings and his or hers mental state to others. Humans have natural ability to recognize emotions through speech information. Affective computing has gained enormous research interest in the development of human computer interaction over the past ten years. With the increasing power of emotion recognition, an intelligent computer system can provide a more friendly and effective way to communicate with users in areas such as video surveillance, interactive entertainment, intelligent automobile system and medical diagnosis[4].

This paper presents a dual –mode recognition system based on speech and facial expression to increase the rate of single mode emotion recognition. Proposed system is shown in figure 1. This system takes audio and image input and extracts the features as per requirement and then use two different classifiers. After this the output of two classifiers is combined together and final result was displayed. For emotion detection through speech Hidden Markov model is used and for emotion recognition through facial expression.

Adaptive Sub layer Compensation Based (ASLC) Edge detection method is used

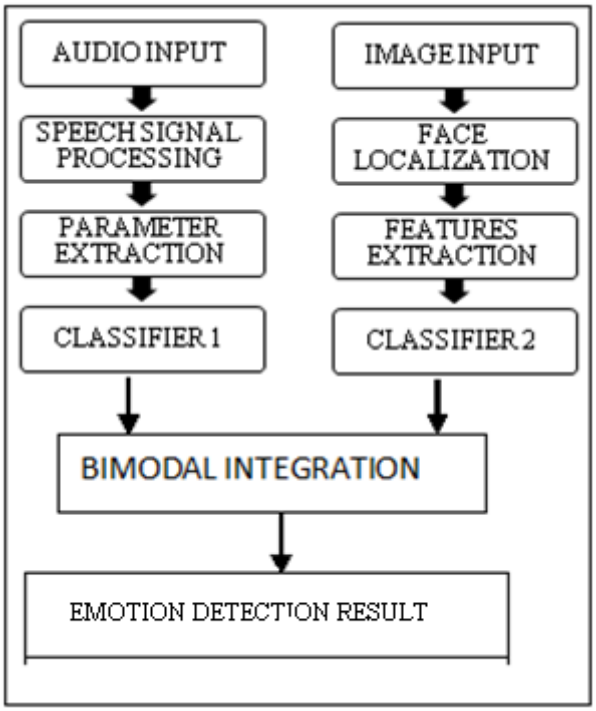


Fig. 1 flow chart of bimodal fusion for emotion detection

The remaining paper is organized as section two describes related work and section three describes methodology for facial expression based emotion detection . section four describes the methodology for speech based emotion detection. Section five gives information about database. section six describes the conclusion.

## II. RELATED WORK

Firstly we will focus on emotion detection through facial expressions. In this preferred method is edge detection. Edge detection is the name for a set of mathematical methods which aim at identifying points in a digital image at which the

image brightness changes sharply or, more formally, has discontinuities. In other words edge detection roughly can be understood as detecting the abrupt changes of gray level intensity in a digital image. There are three basic edge detectors Robert, Sobel and Prewitt But these three cannot give the best results under complex background[2].

Canny edge detector is most promising one and it has become industry standard. Canny edge detector is used to achieve maximum signal to noise ratio (SNR), good localization, and single response. But it has disadvantage that it is less sensitive to subtle facial lines. A second derivative based edge detecting mechanism called Marr-Hildreth detector has caught our attention due to its greater response to weak, non-rigid, deformable facial muscle movements. This edge detector is Laplacian based edge detector which has drawbacks such as higher response to unwanted high frequency details, the appearance of double edge and the so called 'spaghetti effect' .

In this paper we modify this Marr- Hildreth algorithm with Wiener filtering, Sub-Layer compensation and hysteresis analysis to compensate for negative effect of Laplacian of Gaussian operator.

Now we will see the emotion recognition through speech.

On the basis of some universal emotions which includes anger, happiness, sadness, surprise, neutral, disgust, fearful, stressed etc. for this different intelligent systems have been developed by researchers in last two decades. This different system also differs by different features extracted and classifiers used for classification. Prosodic features and spectral features can be used for emotion recognition from speech signal. Because both of these features contain large amount of emotional information. Pitch ,energy, Fundamental frequency, loudness, and speech intensity and glottal parameters are the prosodic features . some of the spectral features are Mel-frequency cepstrum coefficients (MFCC) and Linear predictive cepstral coefficients (LPCC). Also some of the linguistic and phonetic features also used for detecting emotions through speech. There are several types of classifiers are used for emotion recognition such as Hidden Markov Model (HMM), k-nearest neighbors (KNN), Artificial Neural Network (ANN) , GMM super vector based SVM classifier , Gaussian Mixtures Model (GMM) and Support Vector Machine (SVM). emotion classification using GMM has recognition rate of 81%. But this study was limited only on pitch and MFCC features using HMM and obtained the recognition rate of 84%[2]-[5].

The remaining paper is organized as section three describes methodology for facial expression based emotion detection. Section four describes the methodology for speech based emotion detection. Section five gives information about database. section six describes the conclusion.

## III. METHODOLOGY FOR FACIAL EXPRESSION BASED EMOTION DETECTION

Block diagram for proposed system is shown below.

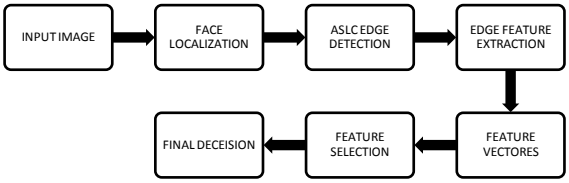


Fig. 2 proposed system for facial expression based emotion detection.

### *A. face localization*

When we capture the image or send the input image then that image is captured in complex background. To remove unwanted region we do the face localization that means unwanted part is removed and only facial part is extracted that process is called face localization. So With the local normalization based method, the proposed system can be more robust under different illumination conditions.

### *B. The Marr-Hildreth Algorithm*

The main part of this paper is to minimize the negative effects of the Marr-Hildreth edge detector, making it more suitable for the purpose of human emotion recognition. Marr-Hildreth edge detector is a two-step algorithm: LoG filtering and zero-crossing detection. Zero-crossing detection is the key part of any Laplacian based edge detection algorithm. Unlike gradient-based edge detection, edge locations in Laplacian based detector are indicated by the existence zero-crossings since a ramp edge or step edge produces a dual response.

### *C. Mechanism*

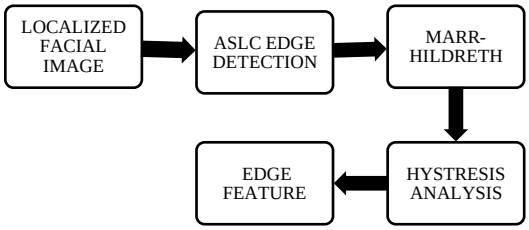


Fig. 3 flow chart of proposed system

When face localization is done our next step is Marr-Hildreth algorithm. This algorithm includes Log filtering and zero crossing detection. The Log is applied and then zero-crossing is detected. Zero-crossing in this stage is made adaptive by the following threshold.

$$referance = mean(f_{LOG}(x,y) + absolute(\min(f_{LOG}(x,y)))$$

$$LoG_{thresh} = S_{LoG} \times referance$$

$f_{Zc}(x,y) = Zerocrossing(f_{LOG}(x,y), LoG_{thresh})$ (1)

where reference represents the mean value LoG filtered image , offset by its minimum intensity. is the adaptive threshold value used in zero-crossing detection, and is the adjustable sensitivity parameter chosen to be 0.015.

*1) Adaptive Sub Layer Compensation(ASLC)*

To overcome all the issues or disadvantages of Marr-Hildreth Algorithm we use Adaptive Sub Layer Compensation based Edge detection. This method filters the original frame with ordinary Sobel Operator[1]. The magnitude of the gradient is expressed as:

$M_{Sobel}(x,y) = |(p_7 + 2p_8 + p_9) - (p_1 + 2p_2 + p_3)| + |(p_3 + 2p_6 + p_9) - (p_1 + 2p_4 + p_7)|$ (2)

Then we threshold the magnitude as:

$M_{Sobel-t}(x,y) = M(x,y) > S \times mean(M(x,y))$ (3)

Where the S=0.8 adjustable parameter.

At this point, both Sobel filtered and zero - crossing detected image are in binary format Based on the fact that gradient operation yields a weaker response to noise while Laplacian operation yields a stronger response to fine details and noise, an intention to eliminate the noise response without solely relying on the Gaussian filter is reasonably possible. Assuming an 8 neighborhood we identify the unwanted details by comparing $f_{Zc}(x,y)$with $M_{Sobel-t}(x,y)$. We remove these unwanted details and find the final valid points.

**Case 1: P1=0 && P2=1** Check the neighbors of P2, if P2 is connected to at least one of its opposite neighbor, then p2 is a valid edge point, otherwise P2 is an isolated noise point.

**Case 2: P1=1 && P2=0** P2 is not a valid edge point. C**ase 3: P1=0 && P2=0** P2 is not a valid edge point.

**Case 4: P1=1 && P2=1** P2 is a valid edge point. We call the binary image after this stage

$f_{Zc\_ASLC}(x,y) = ASLC[f_{Zc}(x,y)]$ (4)

2) *Hysteresis analysis*

As the last step in our algorithm is Hysteresis analysis and this is used for a double thresholding to deal with disconnected edges. Thresholding the edge with a high value would result in many disconnected edges while thresholding the edge with a low value can introduce meaningless image. The key idea behind hysteresis is both high and low is required for this process.

The first threshold is $LoG_{thresh\_H} = LoG_{thresh}$ and second is $LoG_{thresh\_L} = 0.35 \times LoG_{thresh}$ now we can define second zero crossing detected image

$f_{Zc\,SLCH}(x,y) = ASLC[Zerocrossing(f_{LOG}(x,y), LoG_{thresh_L})]$

$f_{Zc\,SLCL}(x,y) = ASLC[Zerocrossing(f_{LOG}(x,y), LoG_{thresh_H})]$ (5)

Here

$$LoG_{thresh_H} > LoG_{thresh_L}$$

So to make these two binary images mutually exclusive we define the following

$f_{ZC\_ASLCL_{LNEW}}(x,y) = f_{Zc\,ASLC_L}(x,y) - f_{Zc\,ASLC_H}(x,y)$ (6)

Here

$f_{ZC\_ASLCL_{LNEW}}(x,y)$ is a weak edge response and

$f_{Zc\,ASLC_H}(x,y)$ is a strong edge response

So if we combine wear edge response to the strong edge response then we will get valid edge points so final image result is

$f_{final\,result}(x,y) = f_{Zc\,ASLC_H}(x,y) + [connectivity\,analysis\left(f_{ZC_{ASLCL_{LNEW}}}(x,y)\right)]$ (7)

*D.PCA based classification:*

This classification method is used for comparing the edge features with the database . other classifiers are also available for this like D-isomap ,Gaussian Mixture Model(GMM). The classification rate for this proposed system is shown below in table 1.

| SR NO. | EMOTIONS | RATE IN % |
|---|---|---|
| 1 | HAPPY | 76.05 |
| 2 | SAD | 70 |
| 3 | NUETRAL | 68.18 |
| 4 | SURPRISE | 70 |
| 5 | FEAR | 55 |

Table 1 Expected result classification rate for PCA classifier.

## IV. METHODOLOGY FOR SPEECH BASED EMOTION DETECTION

The block diagram of the emotion recognition system through speech considered in this study is illustrated in Figure 4. Emotion recognition system through speech is similar to the typical pattern recognition system. An important issue in evaluation of Emotion recognition system through speech is the degree of naturalness of the database used. Proposed system is based on prosodic and spectral features of speech. It consists of the emotional speech as input, feature extraction, classification of Emotional state using SVM classifier and detection of emotion as the output.

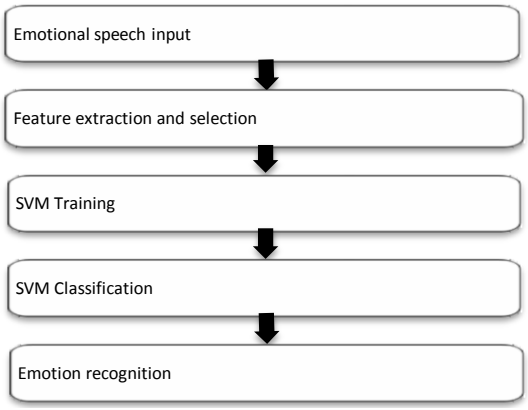

Fig. 4 Emotion recognition through speech

*A. Emotional speech input*

The emotional speech input to the system may contains the collection of the acted speech data the real world speech data. After collection of the database containing short Utterances of emotional speech sample which was considered as the training samples, proper and necessary features such as prosodic and spectral features were extracted from the speech signal. These feature values were provided to the SVM for training of the classifiers. Then recorded emotional speech samples presented to the classifier as a test input. Then classifier classifies the test sample into one of the emotion from the above mentioned five emotions and gives output as recognized emotion.

*B. features extraction and selection*

An important step in emotion recognition System through speech is to select a significant feature which carries large emotional information about the speech signal. Several researches have shown that effective parameters to distinguish a particular emotional states with potentially high efficiency are spectral features such as Mel frequency cepstrum coefficients (MFCC) and prosodic features such as pitch ,speech energy, speech rate ,fundamental frequency. Speech Feature extraction is based on smaller partitioning of speech signal into small intervals of 20 ms or 30 ms respectively known as frames[2]. Speech features basically extracted from vocal tract , excitation source or prosodic points of view to perform different speech tasks. In this work some prosodic and spectral feature has been extracted for emotion recognition. Speech energy is having more information about emotion in speech. The energy of the speech signal provides a representation that reflects these amplitude variations here short time energy features estimated energy of emotional state by using variation in the energy of speech signal. The analysis of energy is focused on short- term average amplitude and short-term energy. We implied short-term function to extract the value of energy in each speech frame to obtain the statistics of energy feature. Another important feature carries information about emotion in speech is pitch. The pitch signal is also called the glottal wave-form. The pitch signal produced due to the vibration of the vocal folds , tension of the vocal folds and the sub glottal air pressure. Vibration rate of vocal cords is also called as fundamental frequency . Another features considering is a simple measure of the frequency content of a signal which is the rate at which zero crossings occur. Zero-crossing rate is a measure of number of times in a given time interval/frame such that the amplitude of the speech signals passes through a value of zero .it is one of the important spectral feature.

The next important type of spectral speech features are Mel-frequency cepstrum coefficients (MFCC).

$$CC_{(n)} = FT^{-1}\{\text{Log } |\text{FT}\{y(n)\}|\} \quad (8)$$

Frequency components of voice signal containing pure tones never follow a linear scale. Therefore the actual frequency for each tone, F measured in Hz, a subjective pitch is measured on a scale which is referred as the "Mel" scale. The following equation shows the relation between real frequency and the Mel frequency is

$$F_{mel} = 3233 \log_{10}\left(1 + \frac{F_{HZ}}{1000}\right) \quad (9)$$

Thus the MFCC component can be obtained as shown in figure 5. while calculating MFCC firstly pre-emphasize of speech signal from constructed emotional database has been done .after this performed windowing over pre-emphasize signal to make frames of 20 sec then the Fourier transform is calculated to obtain spectrum of speech signal and this spectrum is filtered by a filter bank in the Mel domain. Then taking the logs of the powers at each of the Mel frequencies . Then the inverse Fourier transform is replaced by the cosine transform in order to simplify the computation and is used to obtain the Mel frequency cepstrum coefficients. Here we extract the first 13-order of the MFCC coefficients [2].

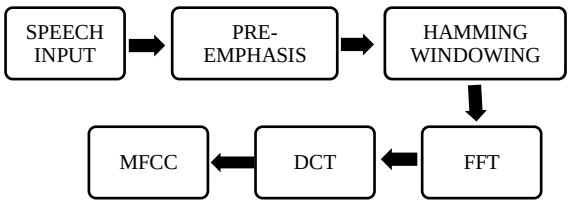

Fig 5. Block diagram of MFCC

C. *SVM classification*

The most important aspect of emotion recognition system through speech is classification of an emotion. The performance of the system influenced by the accuracy of classification, on the basis of different features extracted from the utterances of emotion speech samples emotions can be classified by providing significant features to the classifier. In introduction section describes many type of classifiers, out of which SVM is used in proposed system.

## V. DATABASE AND BIMODAL FUSION

For emotion detection based on facial expressions RML emotion database and Cohn-Kanade (CK) database were used. These database has 320 videos of eight subjects from the RML Emotion database, 360 image sequences of 90 subjects from CK database for the experiment[1]. For our project Chon-Kanade database is suitable. This database is modified by using face localization and then used for our experiment. Total 150 images are used as a database. For each emotion 30 images are used.

The most difficulty for motion detection through speech is the collection of database. The results of speech based emotion detection is good if our database is stronger. The Chinese database is available for this or we can create our own database[2]. After collection of database we compare the extracted features with the database and gives the output.

The expected results for facial expression based emotion detection is shown in table 1, the expected result for speech based emotion detection is shown below in table 2. But when we combine the two methods by using bimodal fusion then the classification rate will increase as shown in table 3 below.

| R NO. | EMOTIONS | RATE IN % |
|---|---|---|
| 1 | HAPPY | 68.0 |
| 2 | SAD | 55.23 |
| 3 | NUETRAL | 44.25 |
| 4 | SURPRISE | 38.2 |
| 5 | FEAR | 56.2 |

Table 2 Expected result classification rate for SVM classifier

| SR NO. | EMOTIONS | RATE IN % |
|---|---|---|
| 1 | HAPPY | 96.45 |
| 2 | SAD | 94.25 |
| 3 | NUETRAL | 93.2 |
| 4 | SURPRISE | 92.06 |
| 5 | FEAR | 90.6 |

Table 3 Expected result of Bimodal fusion.

## VI. CONCLUSION

In this paper we introduced the ASLC edge detection based method for human emotion recognition. We improved the existing Marr-Hildreth detector and identify the correct emotion from basic five emotions. The proposed system do the emotion detection through speech. As we can see from above table 2 the classification rate for speech based emotion detection is less to increase that rate we combine the two results by using Bimodal features and the classification rate is increased as shown in table 3.

## VII. REFERENCES

[1] Yi Huang, Yun Tie, " Human Emotion Recognition using the Adaptive Sub-Layer-Compensation Based Facial Edge Detection," 2013 IEEE.

[2] Akshay S. Utane, Dr. S.L.Nalbalwar "Emotion recognition through speech using Gaussian Mixure Model and Hidden Markov Model" international conference paper April - 2013, pp. 742-746

[3] Yutai Wang, Xinghai Yang, Jing Zou, Research of Emotion Recognition Based on Speech and Facial Expression TELKOMNIKA, Vol.11, No.1, January 2013, pp. 83~90

[4] A.C.Rafael, D. Sidney, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," TAC, l(1), 18-34, 2010

[5] Ashish B. Ingale, D. S. Chaudhari "Speech Emotion Recognition" International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012

[6] Ayadi M. E., Kamel M. S. and Karray F., „Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases, Pattern Recognition, 44 (16), 572-587, 2011.

[7] Chung-Hsien Wu, and Wei-Bin Liang "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels "Ieee Transactions On Affective Computing, Vol. 2, No. 1, January-March 2011.

[8] Zhou y., Sun Y., Zhang J, Yan Y., "Speech Emotion Recognition using Both Spectral and Prosodic Features", IEEE, 23(5), 545-549, 2009.

[9] Nitin Thapliyal, Gargi Amoli "Speech based Emotion Recognition with Gaussian Mixture Model" international Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 5, July 2012

[10] Y.Tie L. Guan "Human Emotion Recognition Using a Deformable 3D Facial Expression Model", ISCAS, 1115 - 1118, 2012

[11] Y. Wang, L. Guan, "Recognizing Human Emotional State from Audiovisual Signals",TMM, 10(5), 659 – 668, 2008

[12] Albornoz E. M., Crolla M. B. and Milone D. H. "Recognition of Emotions in Speech". Proceedings of 17th European Signal Processing Conference, 2009