ISSN:0975-9646

Emotion Recognition from Speech

K. Sreenivasa Rao*[1], Tummala Pavan Kumar[#2], Kusam Anusha[#2], Bathina Leela[#2], Ingilela Bhavana[#2] and Singavarapu V.S.K. Gowtham[#2]

[*]*School of Information Technology, Indian Institute of Technology Kharagpur*
Kharagpur-721302, Midnapore District, West Bengal, India.

[#]*Department of Information Science and Technology,*
Koneru Lakshmaiah College of Engineering
Green Fields, Vaddeswaram-522502, Guntur District, Andhra Pradesh, India.

***Abstract*-This paper proposes the classification of emotions based on spectral features using the Gaussian Mixture Model as the classifier. The performance of the Gaussian Mixture Model has been evaluated for two types of databases – acted and real-life speech corpuses. The model has also been evaluated for the variation in its performance based on the speaker, gender of the speaker and the number of the speakers.**

***Keywords*- emotion recognition system, Gaussian mixture model (GMM), mel frequency cepstral coefficients (MFCCs)**

I. INTRODUCTION

Human machine interfaces are commonly used nowadays in many applications. Most of them require the detection of emotion in the speech. But very few human machine interfaces being implemented currently are able to achieve that [1]. Therefore, there is a need to build a human machine interface that can detect emotions effectively and efficiently. Identification of emotions can be done using three factors - the content of the speech, facial expressions of the speaker or by the features extracted from the emotional speech [2]. This paper is confined to the recognition of emotion by making use of the features extracted from the speech.

Usually human beings can easily detect various kinds of emotions. This can be achieved by the human mind through years of practice and observation. The human mind captures all kinds of emotions since childhood and is taught to differentiate between the emotions based on its observations. For instance, when a person is angry, his tone raises, his expression becomes stern and the content of his speech no longer remains pleasant [3]. Similarly, when a person is happy, he speaks in a musical tone, there is a look of glee on his face and the content of his speech is rather pleasant and joyous. Based on these observations, a person can quickly identify the state of the speaker – whether he is happy, sad, angry, depressed, disgusted etc.

A human machine interface that can process speech having emotional content makes use of a similar concept – training and then testing [4]. In the training phase, the interface is fed with samples of each emotion. The classifier used in the interface extracts features from all the samples and forms a mixture for each emotion. In the testing phase, emotional speech is given as input to the classifier. The classifier extracts the features from the input and compares it to all the mixtures. The input is classified into that emotion to which it is closest. In other words, the input file will be classified into that emotion whose features are the most similar to that of the input file. There are a number of features and classifiers that can be used for the purpose of emotion detection. However, it is difficult to identify the best model among these since the selection of the feature set and the classifier depends on the problem [5].

II. DATABASES

Generally, there are two types of databases that are used in emotion recognition – acted and real. As the name suggests, in acted emotional speech corpus, a professional actor is asked to speak in a certain emotion. In real databases, speech databases for each emotion are obtained by recording conversations in real-life situations such as call centers and talk shows [6]. But it has been observed that there is a difference in the features of acted and real emotional speeches. This is because acted emotions are not felt while speaking and thus come out more strongly [7]. In this work, both acted as well real emotional speech have been considered for classification.

A. Acted Emotion Speech

The acted emotion speech corpus includes the voices of two male and two female actors. A total of four emotions – anger, happy, sad and neutral have been recorded by each one of them. Each actor was made to speak fifteen sentences in the Telugu language in each of the four emotions. The sentences used in daily conversations were chosen for recording. The speech samples are recorded at 16 Hz and are represented by 16 bit numbers.

B. Real Emotion Speech

In this work, voice samples of a single male actor were collected from Telugu movies to form the real emotion speech corpus. Speech samples of the actor in four different emotions were collected. The emotions that were considered for performing the classification in the real emotion speech corpus were – anger, sad, happy and neutral. The speech samples are normalized before feature extraction since they are collected from different sources. 75% of the data was used for training and the rest was used for testing.

III. FEATURES OF SPEECH

Speech signals are produced as a result of excitation in the vocal tract by the source signal. Speech features can therefore be found both in vocal tract as well as the excitation source signal. Features that are extracted from the vocal tract system are called system features or spectral features [8]. The most popular spectral features are Mel frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs) and Perceptual linear prediction coefficients (PLPCs). The features extracted from the excitation source signal are called source features. Linear prediction (LP) and glottal volume velocity (GVV) are some source features. Prosodic features are those features which are extracted from long segments of speech such as sentences, words and syllables. They are also known as supra-segmental features [9]. They contain speech properties such as rhythm, intonation, stress, volume and duration. The acoustic properties of the prosodic features are pitch, energy, duration and their derivatives. The pitch signal is produced when vocal folds vibrate [10]. Pitch frequency and glottal air velocity are the features related to pitch signal. Speech energy is useful because it is related to arousal levels of the emotion. The prosodic features are used to extract emotional expression or excited behavior of articulators. Glottal activity characteristics are evaluated using source features [11]. Spectral features are used to capture the information regarding the movement of articulators and the shape and size of vocal tract which produces different sounds. Articulator is the part of vocal organs that helps form speech sounds. Active articulators are organs such as pharynx, soft palate, lips and tongue. Upper teeth, alveolar ridge and hard palate are passive articulators [12]. The work in this paper is based on the MFCC spectral features.

Mel frequency cepstrum coefficients (MFCCs) are coefficients of mel frequency cepstrum (MFC) which is in turn derived from power cepstrum. Cepstrum is derived from the word ‘spectrum’ by swapping the first half of the word with the second half [13]. A cepstrum is obtained by computing the Fourier Transform of the logarithm of the spectrum of a signal. There are different kinds of cepstrum such as complex cepstrum, real cepstrum, phase cepstrum and power cepstrum. The power cepstrum is used in speech synthesis applications. The cepstrum are linearly spaced frequency bands whereas MFC are equally spaced [14]. Hence, MFCs can provide a better approximation of the speech. The details of computation are not taken up in this work. A total of 21 MFCC features are extracted using the MATLAB software. But before the extraction of features is done, it is necessary that background noises or any other noises are removed. This is because noise or disturbances in the speech interfere with the characteristics of actual speech and the features get altered. Fig. 1 (a) shows the speech signal waveforms of clean speech while (b) shows speech with disturbances where two or more speech signals overlap.

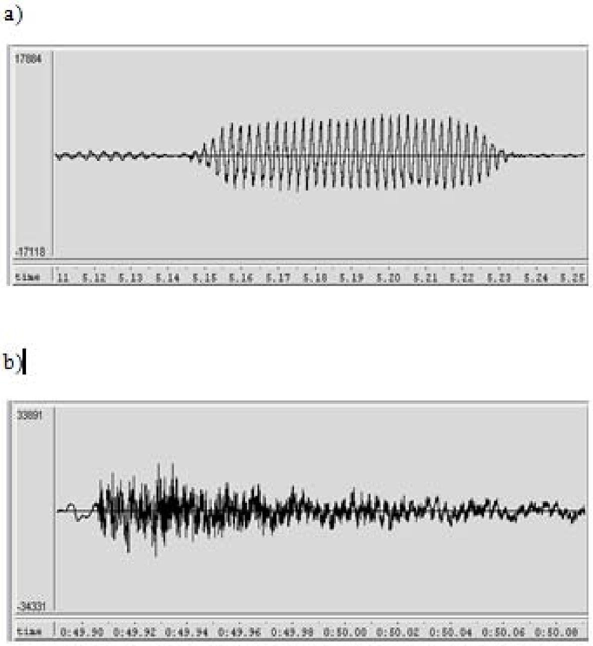

Fig. 1. Speech waveform viewed in wavesurfer: (a) speech without any noise, (b) speech having disturbances

IV. THE CLASSIFIER

The various classifiers that are currently being used are Artificial Neural Networks (ANNs), Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs), k-nearest neighbors and Support Vector Machines (SVMs) [15]. The classification techniques can be divided into two categories – those that make use of the timing information and those which do not [16]. Techniques based on HMMs and ANNs retain the timing information whereas classification techniques based on SVMs and Bayes classifier lose the timing information. One positive aspect of the techniques that retain timing information is that they can be used for speech recognition applications in addition to emotion recognition [17].

Gaussian Mixture Model is used as the model for classification in this work. Gaussian Mixture Models (GMMs) are considered good for evaluating density and for performing clustering [18]. The expectation-maximization algorithm is used for this purpose. GMMs are comprised of component functions called Gausses. The number of these Gausses in the mixture model is also referred to as the number of components. The total number of components can be altered based on the count of training data points. However, the model becomes more complex with the increase in the number of components. In this work, a GMM with 16 components is created for each emotion and these are iterated 30 times. The error count decreases with each iteration and finally reaches a constant. Fig. 2 shows the relationship between error count and number of iterations. The parameters of the model are stored in a diagonal covariance matrix.

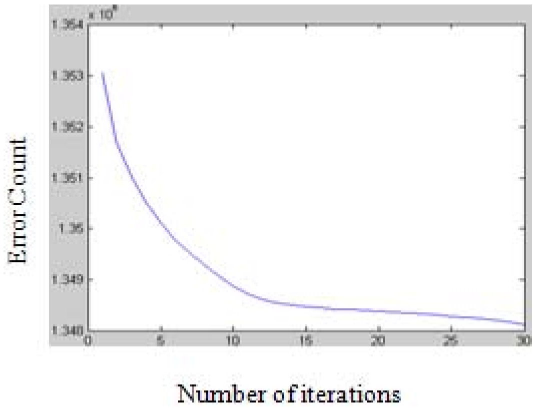

Fig. 2. Error count as a function of the number of iterations for the anger emotion

V. RESULTS

In this work, MFCC features are extracted from the speech signal to perform emotion recognition. A total of 21 features are extracted and the GMMs are constructed using 16 components. Various emotion recognition systems (ERS) are developed for different speakers and combination of speakers.

A. Acted Emotion Speech Corpus

ERS1 is developed using speech samples of female speaker 1. The corresponding percentage confusion matrix of ERS1 is shown in Table I. The values on the diagonal represent the percentage of correctly classified samples whereas the rest of the values represent the percentage of wrongly classified ones. The average correct classification of ERS1 is 95%.

TABLE I
PERCENTAGE CONFUSION MATRIX FOR ERS1
AVERAGE: 95%

	Anger	Happy	Neutral	Sad
Anger	100	0	0	0
Happy	0	97	0	3
Neutral	0	0	100	0
Sad	0	3	13	84

ERS2 is developed using the speech samples of female speaker 2. The corresponding percentage confusion matrix of ERS2 is shown in Table II. The average correct classification rate of ERS2 is 95.83%.

TABLE II
PERCENTAGE CONFUSION MATRIX FOR ERS2
AVERAGE: 95.83%

	Anger	Happy	Neutral	Sad
Anger	100	0	0	0
Happy	0	83	17	0
Neutral	0	0	100	0
Sad	0	0	0	100

ERS3 is developed using speech samples of male speaker 1. The corresponding percentage confusion matrix of ERS3 is shown in Table III. The average correct classification rate of ERS3 is 93.33%.

TABLE III
PERCENTAGE CONFUSION MATRIX FOR ERS3
AVERAGE: 93.33%

	Anger	Happy	Neutral	Sad
Anger	100	0	0	0
Happy	0	90	3	7
Neutral	0	7	87	7
Sad	0	0	3	97

ERS4 is developed using speech samples of male speaker 2. The corresponding percentage confusion matrix of ERS4 is shown in Table IV. The average correct classification rate of ERS4 is 92.5%.

TABLE IV
PERCENTAGE CONFUSION MATRIX FOR ERS4
AVERAGE: 92.5%

	Anger	Happy	Neutral	Sad
Anger	97	3	0	0
Happy	0	87	7	7
Neutral	0	3	94	3
Sad	0	0	7	93

ERS5 is developed using the speech samples of both the female speakers 1 and 2. The speech samples of both female speaker 1 and 2 are used for training and those of female speaker 2 are used for testing. The corresponding percentage confusion matrix of ERS5 is shown in Table V. The average correct classification rate of ERS5 is 68.33%.

TABLE V
PERCENTAGE CONFUSION MATRIX FOR ERS5
AVERAGE: 68.33%

	Anger	Happy	Neutral	Sad
Anger	73	0	27	0
Happy	0	53	47	0
Neutral	0	30	70	0
Sad	0	0	23	77

ERS6 is developed using the speech samples of both the male speakers 1 and 2. The speech samples of male speakers 1 and 2 are used for training and those of male speaker 2 are used for testing. Similar to ERS5, this emotion recognition system has been developed to find out whether the features on speech are dependent on the speaker. The corresponding percentage confusion matrix of emotion recognition system ERS6 is shown in Table VI. The average correct classification rate of ERS6 is 74.17%.

TABLE VI
PERCENTAGE CONFUSION MATRIX FOR ERS6
AVERAGE: 74.17%

	Anger	Happy	Neutral	Sad
Anger	90	0	10	0
Happy	0	97	3	0
Neutral	0	30	70	0
Sad	0	13	47	40

The ERS7 is developed using the speech samples of both the female speaker 1 and male speaker 1. The speech samples of male speaker 1 are used for training and those of female speaker 1 are used for testing. This emotion recognition system has been developed to find whether the features of speech depend on the gender of the speaker. The corresponding percentage confusion matrix of ERS7 is shown in Table VII. The average correct classification rate of ERS7 is 80.83%.

TABLE VII
PERCENTAGE CONFUSION MATRIX FOR ERS7
AVERAGE: 80.83%

	Anger	Happy	Neutral	Sad
Anger	93	7	0	0
Happy	3	90	7	0
Neutral	0	3	94	3
Sad	10	27	17	47

B. Real Emotion Speech Corpus

ERS8 is developed using speech samples of a Telugu male actor. The speech samples of the actor have been labeled as that of male speaker 3. The corresponding percentage confusion matrix of ERS8 is shown in Table VIII. The average correct classification rate of ERS8 is 85 %.

TABLE VIII
PERCENTAGE CONFUSION MATRIX FOR ERS8
AVERAGE: 85 %

	Anger	Happy	Neutral	Sad
Anger	82	11	0	7
Happy	13	87	0	0
Neutral	0	7	90	3
Sad	6	0	13	81

This emotion recognition system has been developed to compare the performance of the model for real emotions against the acted ones.

VI. CONCLUSION

Table IX summarizes the performance of all the emotion recognition systems developed in this study. Three observations can be made from table IX. First, the performance of the emotion recognition systems decreases when the voices of two males or two females are mixed. This indicates that the systems are dependent on the speaker. Second, the performance also decreases when the voices of male and female are mixed, which shows that the system is dependent on gender of the speaker. Last of all, a decrease in the performance of the real emotion recognition systems has been observed in comparison to acted emotion recognition systems.

TABLE IX
SUMMARY OF THE PERFORMANCE OF ALL THE EMOTION RECOGNITION SYSTEMS DEVELOPED IN THIS STUDY

Emotion Recognition System	Average (%)
ERS1 (acted, single female)	95
ERS2 (acted, single female)	95.83
ERS3 (acted, single male)	93.33
ERS4 (acted, single male)	92.5
ERS5 (acted, two females)	68.33
ERS6 (acted, two males)	74.17
ERS7 (acted, one male one female)	80.83
ERS8 (real-life, single male)	85

ACKNOWLEDGEMENT

This work was initiated by K. Sreenivasa Rao, who also provided the acted emotion speech corpus. The quality of this work was enhanced by the valuable guidance of Tummala Pavan Kumar.

REFERENCES

[1] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, *IITKGP-SESC : Speech Database for Emotion Analysis*. Communications in Computer and Information Science, JIIT University, Noida, India: Springer, issn: 1865-0929 ed., August 17-19 2009.

[2] S. G. Koolagudi and K. S. Rao, “Exploring speech features for classifying emotions along valence dimension,” in *The 3rd international Conference on Pattern Recognition and Machine Intelligence (PReMI-09), Springer LNCS* (S. C. et al., ed.), (IIT Delhi), pp. 537–542, Springer-verlag, Heidelberg, Germany, December 2009.

[3] S. G. Koolagudi, S. ray, and K. S. Rao, “Emotion classification based on speaking rate,” in *The 3rd International Conference on Contemporary Computing*, (Noida, India), JIIT university and University of Florida, August 2010.

[4] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersy: Prentice-Hall, 1993.

[5] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” *Speech Communication*, vol. 48, p. 11621181, 2006.

[6] S. R. M. Kodukula, *Significance of Eecitation Source Information for Speech Analysis*. PhD thesis, Dept. of Computer Science, IIT, Madras, March 2009.

[7] T. L. Pao, Y. T. Chen, J. H. Yeh, and W. Y. Liao, “Combining acoustic features for improved emotion recognition in mandarin speech,” in *ACII* (J. Tao, T. Tan, and R. Picard, eds.), (LNCS 3784), pp. 279–285, Springer-Verlag Berlin Heidelberg, 2005.

[8] T. L. Pao, Y. T. Chen, J. H. Yeh, Y. M. Cheng, and C. S. Chien, *Feature Combination for Better Differentiating Anger from Neutral in Mandarin Emotional Speech*. LNCS 4738, ACII 2007: Springer-Verlag Berlin Heidelberg, 2007.

[9] S. Wu, T. H. Falk, and W. Y. Chan, “Automatic recognition of speech emotion using long-term spectro-temporal features,” in *16th International Conference on Digital Signal Processing*, (Santorini-Hellas), pp. 1–6, IEEE, 5-7 July 2009. DOI: 10.1109/ICDSP.2009.5201047.

[10] D. Neiberg, K. Elenius, and K. Laskowski, “Emotion recognition in spontaneous speech using gmms,” in *INTERSPEECH 2006 - ICSLP*, (Pittsburgh, Pennsylvania), pp. 809–812, 17-19 September 2006.

[11] D. Bitouk, R. Verma, and A. Nenkova, “Class-level spectral features for emotion recognition,” *Speech Communication*, 2010. Article in press.

[12] M. Sigmund, “Spectral analysis of speech under stress,” *IJCSNS International Journal of Computer Science and Network Security*, vol. 7, pp. 170–172, April 2007.

[13] R. Nakatsu, J. Nicholson, and N. Tosa, “Emotion recognition and its application to computer agents with spontaneous interactive capabilities,” *Knowledge-Based Systems*, vol. 13, pp. 497– 504, December 2000.

[14] V. A. Petrushin, “Emotion in speech:recognition and application to call centers,” Proceedings of the 1999 Conference on Artificial Neural Networks in Engineering (ANNIE 99), 1999.

[15] E. Bozkurt, E. Erzin, C. E. Erdem, and A. T. Erdem, “Improving automatic emotion recognition from speech signals,” in *10th Annual Conference of the International Speech Communication Association (Interspeech 2009)*, (Brighton, UK), pp. 324–327, 6-10 September 2009 2009.

[16] N. Kamaruddin and A. Wahab, “Features extraction for speech emotion,” *Journal of Computational Methods in Science and Engineering*, vol. 9, no. 9, pp. 1–12, 2009. ISSN:1472-7978 (Print) 1875-8983 (Online).

[17] S. McGilloway, R. Cowie, E. Douglas-Cowie, S. Gielen, M. Westerdijk, and S. Stroeve, “Approaching automatic recognition of emotion from voice: A rough benchmark,” (Belfast), 2000.

[18] F. Dellert, T. Polzin, and A. Waibel, “Recognizing emotion in speech,” (Philadelphia, PA, USA), pp. 1970–1973, 4th International Conference on Spoken Language Processing, October 3-6 1996.

[19] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification,” pp. I593–I596, ICASSP 2004, IEEE, 2004.

[20] J. Nicholson, K. Takahashi, and R.Nakatsu, “Emotion recognition in speech using neural networks,” in *6th International Conference on Neural Information Processing*, pp. 495–501, ICONIP-99, 1999.

[21] O. Kwon, K. Chan, J. Hao, and T. Lee, “Emotion recognition by speech signals,” (Geneva), pp. 125–128, Eurospeech, 2003.

[22] A. I. Iliev, M. S. Scordilis, J. P. Papa, and A. X. Falco, “Spoken emotion recognition through optimum-path forest classification using glottal features,” *CSL*, vol. in press, 2009.

[23] B. Schuller, G. Rigoll, and M. Lang, “Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture,” (ISBN: 0-7803-8484-9), pp. I– 577–80, IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP ’04), May 17-21 2004.

[24] C. M. Lee and S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEEAUP*, vol. 13, pp. 293–303, March 2005.

[25] S. V. Gangashetty, C. C. Sekhar, and B. Yegnanarayana, “Spotting multilingual consonant-vowel units of speech using neural network models,” in *NOLISP* (M. Faundez-Zanuy, ed.), p. 303317, Springer-Verlag Berlin Heidelberg, 2005.

[26] S. R. M. Prasanna, B. V. S. Reddy, and P. Krishnamoorthy, “Vowel onset point detection using source, spectral peaks, and modulation spectrum energies,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, pp. 556–565, May 2009.