Journal of Theoretical and Applied Information Technology
15th July 2018. Vol.96. No 13
© 2005 – ongoing JATIT & LLS

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

PERFORMANCE IMPROVEMENT AND ANALYSIS OF SPEAKER INDEPENDENT EMOTION RECOGNITION SYSTEM USING I-VECTORS

[1]NAGA PADMAJA JAGINI, [2]RAJESWAR RAO. R
[1]Assistant Professor, JNTUH, Department of CSE, INDIA.
[2]Professor, JNTUK, Department of CSE, INDIA.
E-mail: [1]srija26@gmail.com, [2]rajaraob4u@gmail.com

ABSTRACT

Emotion recognition from speech is experiencing different research applications. It is becoming one of the tool for analysis of health condition of the speaker. In this work, the emotions such as anger, fear, happy, neutral are considered for speech emotion algorithm design. A database built by IITKGP is used for emotion recognition. For any recognition, feature extraction and pattern classification are the important tasks. In this work the features considered are Mel Frequency Cepstral Coefficients (MFCC), Pitch chroma, prosodic features are used and i-vectors are used to identify the emotions. in this research work, the database considered for emotion recognition is taken in different combinations such as male training-female testing, male training-male testing, female training- female testing, female training-male testing. In almost all the cases, i-vector method has shown an improvement in recognition accuracy than the method employed principle component analysis.

Keywords: *Emotion Specific, I-Vector, Gaussian Mixture Models, Prosody Features, Spectral Features, PCA*

1. INTRODUCTION

Distinctive data, for example, emotion, speaker, gender, dialect and so forth can be seen from speech signal. Increment in the human PC collaboration requests more research in speech handling, for example, speaker ID, emotion ID, Speech recognition and so on. The execution of the a speech recognition or a speaker identification framework [1, 2] relies upon the extraction of appropriate features from the speech signals. Emotion, on the flip side is an individual mental state. There are various emotion which effect the flow of the speech [1]. Emotion recognition is crucial in making the speech recognition efficient

A fundamental challenge for current research in the area of speech science and technology is understanding and modelling individual variations in spoken language [2]. Individuals have their own style of speaking, depending on many factors like emotion, dialect and accent of the speech. Due to prosody of the language, speaker identification also becomes challenging task [3]. In this present work, emotion recognition is performed for various blends of training and testing utterances. The emotions like happy, Fear, Anger and neutral are considered in this work.

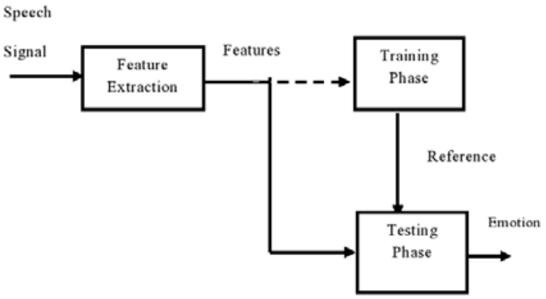

Figure 1: Block diagram representing Emotion Recognition task.

The discourse emotion recognition falls under the classification of example grouping and example recognition as appeared in Fig.1 which includes highlight extraction for preparing and testing emotion talks and determination of the best appropriate highlights is a critical aspect[3, 4]. Diverse highlights, for example, prosody and unearthly highlights are utilized as a part of this component. There are distinctive ghastly and prosodic highlights, in which few of them are appropriate for given discourse handling application. Keeping the above reality in see, the emotion recognition execution is investigated with a blend of prosodic and ghostly highlights.

2. RELATED WORK

Kasiprasad mannepalli et al have worked on emotion recognition system [5] that utilized MFCC features, pitch chroma, spectral flux and tonal power ratio. The classification methodology is improved by using fractional calculus along with Deep Belief networks. The performance of the system is considerably.

Laura Caponetti,Cosimo Alessandro Buscicchio and Giovanna Castellano [6] have taken a shot at emotion recogntion framework configuration in light of MFCC and Lyon cochlear models. The analyses were performed on SUSAS corpus. From perceptions made by them, thecombinations fo the Long Short Term Memory (LSTM) classifier with the Lyon cochlear portrayal gave better recognition performance when contrasted and consolidating a similar classifier with the customary MFCC portrayal.

Yazid Attabi and Pierre Dumouchel [7] have identified that, for the highly unbalanced classes of speech data, the back-end system increases the recognition accuracy obtained by using Gaussian Mixture Models with the application of proper sampling. The experiments demonstrated that anchor models such as Euclidean distance improved the performance of GMMs by 6.2 percent.

L. Zao [8] proposed the pH time-frequency vocal source feature for multi-style emotion classification in which binary acoustic mask also considered to increase the accuracy of emotion classification. Two different databases were used to test this methodology, the Berlin EMO-DB and SUSAS. The MFCC and the pH feature have shown improvement in accuracy with acoustic mask.

In the design of emotion recognition system [9] different features Energy, different Formants, and Zero Crossing Rate (ZCR) have been used for the samples of different range (50 or 100) for classification in the work “Realization of emotions in speech using prosodic and articulation features”. The use of energy feature alone gave a recognition rate of 58% only. By adding formants and zero crossing rate, the recognition accuracy was increased to 78%.

Mayank Bhargava and Tim Polzehl [10] have proposed their work on emotion recognition using rhythm and temporal features Apart from the features like MFCC’s, pitch and energy/ intensity. The recognition accuracy of 80.60 % on Berlin emotion database with 7 different emotions with speaker dependent framework was obtained.

Yasmine Maximos and David Suendermann- Oeft [11] have explored the recognition of emotions in speech via class of the FAU Aibo Corpus for the 2-class task. The database includes German speech recordings of fifty one children. They have examined special spectral and prosodic feature combination. They have taken into consideration 110 features and a feature size reduction technique was used. The parameter optimized sequential minimal optimization set of rules with Unweight averages recogntion (UAR) gave a recognition accuracy of 69.39%. The Successive minimum Optimization (SMO) was said as better optimizer. The original overall performance of UAR turned into 63.8% and after adding, attributes selection and parameter optimization a growth in recognition accuracy by means of 5.6% is observed.

Ankur sapra, nikhil panwar and sohan panwar [12] have mentioned specific strategies for detection of human feelings is discoursed based totally at the acoustic features like pitch, energy and so on. The proposed gadget is the usage of the traditional MFCC technique after which the usage of nearest neighbor for the classification. The MFCC capabilities were extracted from the speech samples after the speeches were separated for male and woman. The recognition accuracy obtained by them in this proposed technique become 90%.

Vaishali M. Chavan and V.V. Gohokar [13] used Support Vector Machine (SVM), to categorize five emotional states i.e anger, happiness, disappointment, surprise and a neutral emotion of human speech. The functions taken into consideration in their work have been Mel Frequency Cepstrum Coefficients. The database used was the Danish Emotion Speech (DES) Database. The accuracy acquired through using SVM classifier have been 68 % for Linear regression version, 60 % for polynomial regression, 55.40 % for RBF regression and 60 % with sigmoid regression. Overall time taken through SVM changed into 15 seconds for testing records set and for the take a look at pattern it turned into 43 seconds to test 88 values.

From the literature review, it is unmistakable that, most of the emotion studies using speech are conceded using prosodic features [14]. pitch, duration and energy and their derivatives related Prosodic features are treated as high correlates of emotions [15].

Fundamental Frequency (F0), raise and falls of F0 contours measures and characterizes the emotions [16]. n [17][18] completely different emotions and there importance within the context of delivery options is mentioned. In varied dynamic nature of delivery contours [19] clearly provides varied expressions associated with emotions. Perception of high arousal emotions and low arousal emotions is achieved victimization MFCC, ZCR and Pitch [20].

Diana Torres-Boza et al [21] proposed method Hierarchical Sparse Coding (HSC) and Global Descriptor Extraction Layer (GDEL) is a combinational technique of Transfer Principle Component Analysis (TPCA) and Sparse Coding Learning for emotion recognition. From the labeled and unlabeled data sparse coding stage set to learn robust feature representation. The experimental results of the proposed method gives 73% of accuracy in emotion recognition and by using the combinational method valence and arousal features emotion prediction is enhanced to 78% of accuracy.

Jainath Yadav, ,Md. Shah Fahad et al [22]. Due to rapid changes of pitch period in emotional speech performance degradation is more in the existing epoch estimation algorithm. In the proposed work zero time windowing method is utilized to get the spectral information of each sample point instantaneously. Compared to the neighboring sample points the amplitude of instant spectral of windowing technique is higher. The proposed method uses the sum of three prominent spectral peaks at each sampling instant of the Hilbert envelope of Numerator Group Delay (HNGD) spectrum, for accurate detection of epochs in the emotional speech. The identification rate (IDR) of epoch extraction is significantly high in the proposed method. The average values of IDR, RMSE, and IDA of detected epochs using proposed method are 93.69%, 0.42 ms, and 0.30 ms, respectively for the emotional speech.

Zhen-Tao Liu, et al [23] .Feature vectors are set by extracting the test samples of speech emotional data., to realize the classification experiment these set are respectively input into the ELM decision tree and SVM decision tree. The time of accomplish speech emotion recognition and the accuracy of the two groups of experiment, in which the recognition time is the time from the input of the selected feature vector set to the output of the result. When using the feature set without feature selection, the average recognition accuracy of ELM decision tree is up to 88.251%, which is about 1.2% higher than SVM decision tree. And recognition time of ELM decision tree is about 1.2 s less than SVM decision.

ShaolingJing et al [24]. This research proposes a novel type of feature related to prominence, which, together with traditional acoustic features, are used to classify seven typical different emotional states. A Chinese Dual-mode Emotional Speech Database (CDESD), which contains additional prominence and paralinguistic annotation information. Then, a consistency assessment algorithm is presented to validate the reliability of the annotation information of this database. The results show that the annotation consistency on prominence reaches more than 60% on average. The proposed prominence features are validated on CDESD through speaker-dependent and speaker-independent experiments with four commonly used classifiers. Consistency increases with increasing tolerance. For a time difference tolerance of 5ms, annotators is higher than 60% on average. When the time difference tolerance is increased to 8 ms and 10 ms, the consistency increases by 6.5% and 10.5%, respectively. 70% the low performance achieved in categorical emotion recognition is probably the monetization of emotion as distinct and independent affective states.

M. Srikanth, et al [25]. In this model Mel frequency Cepstral Coefficient (MFCC) and Gaussian Mixture Model (GMM) is used to extract the emotion independently. The proposed model GMM-DBN system by testing on the Classical German database (EmodB). Bag of acoustic features (BoF) is derived by identifying the minimum distance distribution for individual utterance regarding emotion and histogram plotting count gives the feature distribution that is near to each emotion model. The performance rate increase by 5% than the conventional MFCC-GMM system by empirical observation. Further testing of the proposed system over the recently developed simulated speech emotion database for Tamil language gives a comparable result shows 90% accuracy for the emotion recognition.

Dai et al. [26] proposed a novel approach for emotional speech recognition and to analyze the voice content posted in social media like Wechat. By dynamic and emotional variations in Position-arousal-dominance (PAD) 25 acoustics voice signals are extracted and least squares-support vector regression (LV-SVR) model is used for classification of speech emotion. From the experimental results the recognition range for different emotions varies and the final average rate of recognition achieved is around 83%. Which has the high accuracy rate than previous models.

Cao et al. [27] in order to reduce the error in binary classification, SVM ranking model is established to synthesize information of emotion speech recognition. To apply muti-class prediction each utterer is considered as individual query and then gathering all predictions from rankers for particular emotion. The major advantage of Ranking SVM method is to obtain speaker specific data for training and testing steps in speaker – independent and each speaker can express mixed emotion to recognize the dominant emotion. The proposed method attains gain in terms of accuracy when compared with conventional SVM. Unweight average (UA) or Balance accuracy achieved is 44.4%.

3. METHODOLOGY OF THE PROPOSED APPROACH

Basically in view of the investigations completed in going before stage, it's miles introduced that the spectral features and prosodic features have two exceptional discriminative qualities. Along these lines to find the novel emotion specific data, it's far proposed to combine both spectral and prosodic features. For spectral features MFCC, Pitch Chroma, Spectral Centroid, and Spectral Skewness are viewed as and for prosodic features Pitch, Formants, Zero Crossing Rate, Root Mean Square and Energy are considered. The philosophy of the proposed approach is appeared in Fig 2. The database considered in this paper is IITKGP database. The speeches were separated for both the training and testing sets in the emotions of anger, Fear, Happy and Neutral. Both spectral and prosodic features are extracted and these features are concatenated. I-vectors extracted from the means of Gaussian Mixture Model (GMM) and are applied for the recognition of emotions for different cases.

3.1 Procedure For Feature Extraction

The important task in emotion recognition system is the extraction of best suitable features which represents acoustic data.

3.1.1 Silence Removal

The speech signal consists of both silence and voiced regions. The silent regions are identified and removed to obtain the exact feature from emotion speech database

3.1.2 Pre Processing

The speech signal identified with every emotion is preprocessed before demonstrating. Here, the given emotion speech test is pre-emphasized, blocked, windowed and examined at 16000 samples/sec. a frame size of 25 ms is considered with a shift of 10ms. Amplitude check is ascertained for the given speech signal to see if features are to be extracted or not. The speech signal is pre-emphasized and the windowing is performed using hamming window.

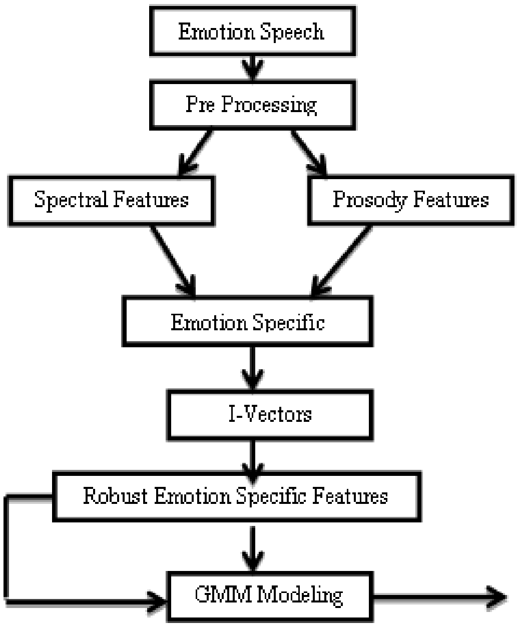

Figure 2: Emotion Recognition System.

The pre-emphasize filter can be given by equation 1.

$$s'(t) = s(t) + \alpha * s(t-1) \qquad \ldots\ldots\ldots\ldots\ldots(1)$$

The typical α value is taken between 0 and 1. Generally hamming window is used for segmenting and selecting a part of speech signal. The Hamming window can be represented mathematically as follows in the equation 2.

$$w(l) = (1-a) - a\cos\left[\frac{2\pi l}{N-1}\right] \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(2)$$

Here a = 0.46164 and where N is the length of window.

3.1.3. Pitch Chroma

The extraction of pitch chroma is the powerful tool to categorize the pitches significantly. On an equally tempered scale , the frequency of the Fourier transform is drawn with the 12 semi-tones pitch class C. Due to this, the transients are reduced and noise is minimized [21]. The pitch chroma is obtained as shown in the equation 3.

$$\left[C(l) = \sum_{n';c(n'=l)} \sum_{f} h_{n'}(f) X(f);\, l \in [0,12] \qquad .(3) \right]$$

Where, $X(f)$ is the Fourier transform of input speech signal. The filter hn(f) is defined by the following equation. separation between spectrum frequency and centre of the filter is represented by parameter x

$$h_{n'}(f) = \frac{1}{2}\tanh\left(\pi(1-2x)\right) + \frac{1}{2} \qquad .....(4)$$

Where,

$$x = R\left|n' - n(f)\right|$$

3.1.4. Mel Frequency Cepstral Coefficients

The human audio frequency bands can be extracted by Mel Frequency Cepstral Coefficient (MFCC) [22, 23]. The DFT is computed for the speech segments using the window function defined in earlier section and thus the power spectrum is found. The power spectrum is converted into Mel spectrum by triangular filter banks called as Mel filter banks. Further, the energy G is calculated by multiplication of each filter bank with power spectrum and the logarithm is applied. MFCC features are obtained by application of the Discrete Cosine Transform (DCT) on the filter bank energies $G(p)$ for the p log filter bank energies is calculated as shown in equation 5.

$$M(p) = \frac{1}{2}(G_0 + (-1)^k G_{p-1} + \sum_{q=1}^{M-2} G_q \cos\left[\frac{\pi}{m-1} qp\right]$$

$$m = 0,\ldots,M-1 \qquad (5)$$

Where, $M(m)$ is the desired p cepstral Coefficient.

3.1.5. PROSODIC FEATURES

Prosodic features are supra segmental in nature which can happens in some top level of an speech utterance. These prosodic features require not identify with linguistic units, for example, phrases and clauses. Prosodic units are set apart by phonetic prompts which incorporate parts of prosody like Pitch, and Accents, that must be broke down in setting, or in contrast with different parts of a sentence. For instance, Pitch can change over the span of sentence, falling sounds. Prosody helps in settling sentence vagueness. The pitch related prosody is gotten via autocorrelation technique.

3.1.6. FORMANTS

The spectrum peak of the speech is called as the formant. The frequency components of human speech formants with F1, F2, F3. The arrangement of formants is from F1 to F3 in an increasing order. The formants generally distinguish the vowels. The LPC model is used to find the formants of a given speech signal in this research work. Linear Prediction Coding (LPC) method is used in estimation of formant estimation in light of the fact that the determination can be set by windowing the speech signal and obtaining the LPC coefficients

The mathematical expression for LPC can be shown in the z- domain with the equation 3.

$$H(Z) = \frac{1}{1 - \sum_{K=1}^{P} a_k Z^{-k}} \qquad ....(3)$$

The poles of the vocal tract can be given by the roots of linear predication coding and the formants are associated with respective vocal tract poles.

3.2. I-vector

Speech can be represented by i-vectors in a low dimension. I-vector is the short-form of identity vectors. These i-vectors capture the speaker and channel characteristics. The i-vector extraction of the emotion model is obtained by Joint factor Analysis (JFA) in this model. In this technique the means obtained for Gaussian Mixture model (GMM) are concatenated as follows.

$$M_c = \{\mu_1, \mu_2, \mu_3 ....\mu_k \qquad ....6)$$

The joint factor analysis is performed by decomposition of into Eigen voice and Eigen channel
That can be given by

$$M_c = \{m + V_X + U_Y + D_Z \qquad ....7)$$

Where m is emotion independent and session independent super vector, D is diagonal residual matrix, V represents the Eigen voice residual matrix. Both V and D matrixes represent speaker subspace.

The figure 3 shows i-vector extraction method for the emotion speaker identification.

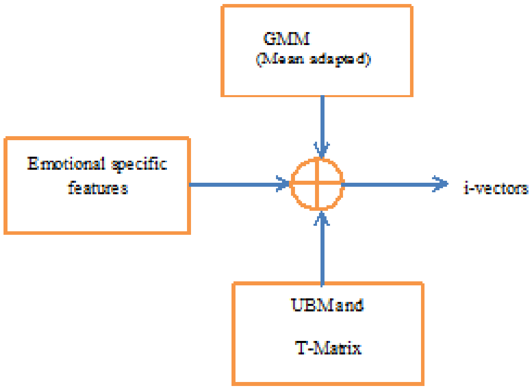

Figure 3: Extraction of I-Vectors.

Initially, the UBM is adapted to form the GMM super vector and in the total variability space Eigen vectors are extracted to get i vector values.

The steps involved in the i vector extraction are given below.

Step 1:

In UBM with a C component $c = \{1,2,3,4,.........,C\}$, the Bam-Welch sutatistics [28] of zeroth order and first order are calculated as shown in the equation,

$$I_c = \sum_{t=1}^{L} p(c/x_t) \qquad \ldots\ldots\ldots\ldots\ldots(8)$$

Where c is the mixture component, x_t is the feature frames, $p\left(c/x_t\right)$ is the probabilistic alignment (Posterior probability of the cth component for the tth observation)

Step 2:

Expectation maximization algorithm is used for i-vector is extraction using the expectation maximization algorithm. The mean of the posterior components of the UBM delivers the corresponding vector.

The value for the corresponding feature vector is calculated using the probability function.

$$P(x) = \sum_{r=1}^{c} \lambda_i N(x, \mu_i, \Sigma_i) \qquad \ldots\ldots\ldots\ldots\ldots\ldots(9)$$

The element λ_i is calculated using the probabilistic alignment of the statistical values.

3.3 Gaussian Mixture Model (GMM)

GMM is considered to be one of the best parametric approaches for modelling the spectral shapes of the emotional speech signal [23]. For the given emotion model, the probability distribution is represented by the following equation (8) and equation (9)

$$p(\bar{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots 1(0)$$

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2} \left|\Sigma_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x}-\bar{\mu}_i)^T \Sigma_i^{-1} (\bar{x}-\bar{\mu}_i)\right\} \qquad (11)$$

The mean vector $\bar{\mu}_i$ and covariance matrix Σ_i represents the Gaussian $b_i(\bar{x})$ by satisfying the condition $\sum_{i=1}^{M} p_i = 1$. The each emotion is represented by Gaussian model with these parameters as $\lambda = \{p_i, \bar{\mu}_i, \Sigma_i\}$ for i = 1, … M.

4. *FEATURE EXTRACTION OUTPUT*

The experimentations were done by combining both spectral and prosody features.

4.1 EXPERIMENTAL SETUP

The proposed Mixture models (GMMs) for modeling. 30 seconds of speech duration are considered for training and ten samples each of 3 seconds of speech duration is considered for testing. In this work the number of Gaussian Mixtures considered were 32. The speech signal of anger emotion is shown in Fig.4.

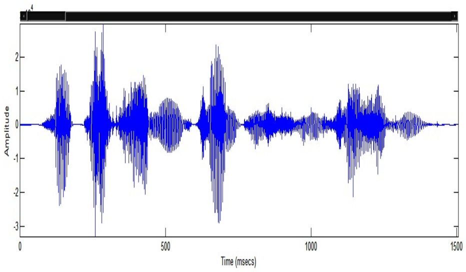

Fig 4.a: the speech signal of Anger emotion

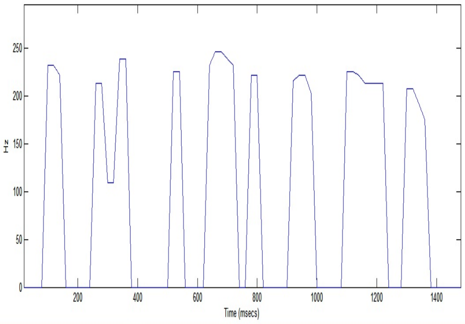

Fig 4.B: The Pitch Graph And Related Prosody For The Signal In Fig4.A

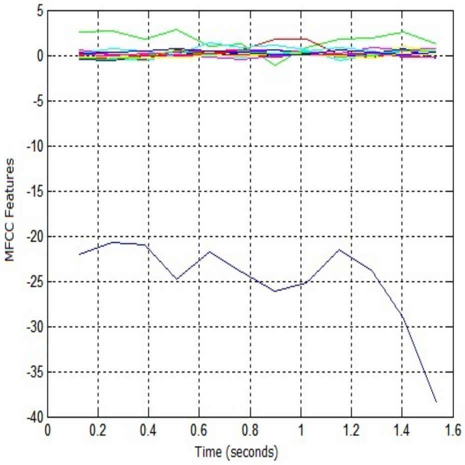

Fig 4.C: The MFCC Feature Graph For The Signal In Fig4.A

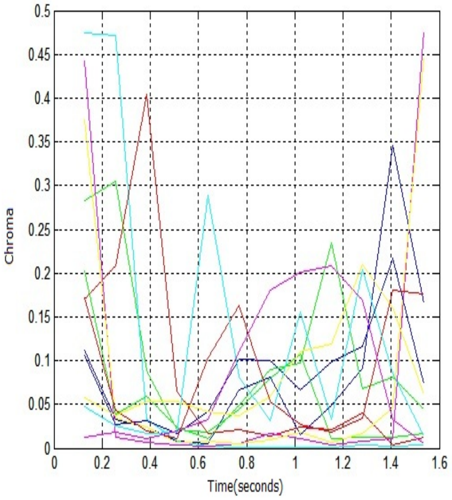

Fig 4.D: The Pitch Chroma For The Signal In Fig4.A

5. DISCUSSION OF RESULTS

In this research work, four emotions are considered for the analysis of the algorithm, those are anger, fear, happy and neutral. The research is carried out by taking the emotion speech sample in three different combinations of training and testing samples. The combinations are training with Male speech samples and testing with Male speeches termed as M-M, The training with Male speech samples and testing with Female speeches termed as F-M. The training with Female speech samples and testing with Male speeches termed as M-F. The training with Female speech samples and testing with Female speeches termed as F-F. In work i-vector are generated by joint feature analysis, by taking the means of Gaussian Mixtures.

The result is compared with PCA feature dimensionality reduction technique. When M-M case is considered, the emotion recognition of anger has improved from 80% to 90%, happy is improved from 70% to 80%, neutral has improved from 80% to 90%,whereas recognition accuracy of fear id in line with the PCA.

In the case M-M is considered, the emotion recognition accuracy of anger has improved from 30% to 40%, happy is improved from 30% to 40%, and happy has improved from 40% to 50%, whereas recognition accuracy of neutral is 50% which is in line with the PCA.

Table.1: Comparison Of Recognition Accuracies M-M Case

	Anger	Fear	Happy	Neutral
PCA Approach	80	90	70	80
i-vector Approach	90	90	80	90

The graphical representation of the comparison of emotion recognition system for M-M case is depicted in fig.5

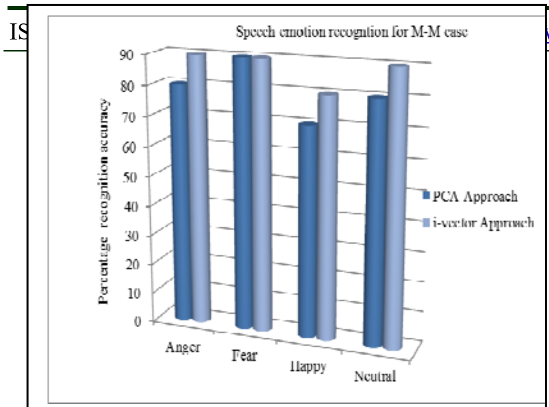

Fig.5 comparison of percentage accuracies for M-M case

In the case M-F, as shown in Table2, the emotion recognition accuracy of anger has improved from 30% to 40%, happy is improved from 30% to 40%, and happy has improved from 40% to 50%, whereas recognition accuracy of neutral is 50% which is in line with the PCA.

Table.2: Comparison Of Recognition Accuracies M-F Case

	Anger	Fear	Happy	Neutral
PCA Approach	30	30	40	50
i-vector Approach	40	40	50	50

The graphical representation of the comparison of emotion recognition system for M-F case is depicted in fig.6

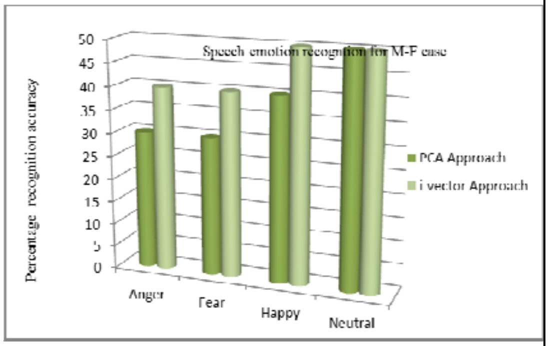

Fig.6 Comparison Of Percentage Accuracies For M-F Case

In the third case F-M,, the emotion recognition accuracy of anger has improved from 40% to 60%, fear is decreased from 60% to 50%, and happy has improved from 40% to 50%, whereas recognition accuracy of neutral is 60% which is in line with the PCA and this is shown in Table.3

Table.3: Comparison Of Recognition Accuracies F-M Case

	Anger	Fear	Happy	Neutral
PCA Approach	40	60	40	60
i-vector Approach	60	50	50	60

The comparison also shown graphically in the fig.7 for the F-M case.

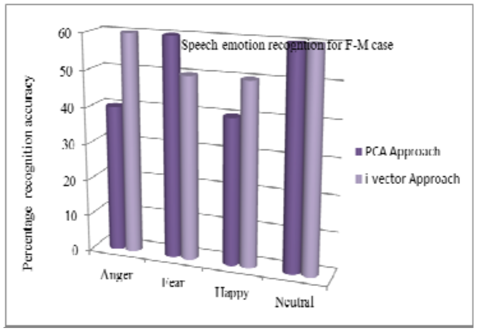

Fig.7 Comparison Of Percentage Accuracies For M-F Case

Table.4 The emotion recognition accuracy of happy is improved from 90% to 100%, neutral has improved from 60% to 80%, whereas recognition accuracy of anger and fear have shown an accuracy of 100% and 90% respectively which are in line with the PCA.

Table.4: Comparison Of Recognition Accuracies F-F Case

	Anger	Fear	Happy	Neutral
PCA Approach	100	90	90	60
i-vector Approach	100	90	100	80

The graphical representation of the comparison of emotion recognition system for M-F case is depicted in fig.5

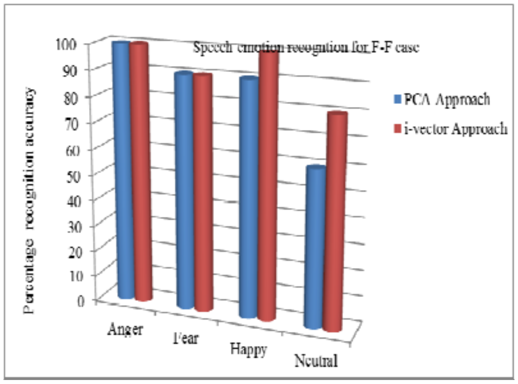

Fig.7 comparison of percentage accuracies for F-F case

The performance PCA based emotion recognition system and I-V are compared with respect to emotions in the ,fig 8,fig 9,fig 10 and fig 11..

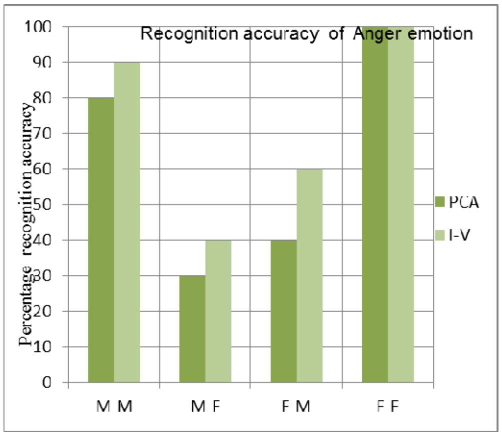

Fig 8: performance of PCA method and I-Vector method for anger emotion for different test and train cases

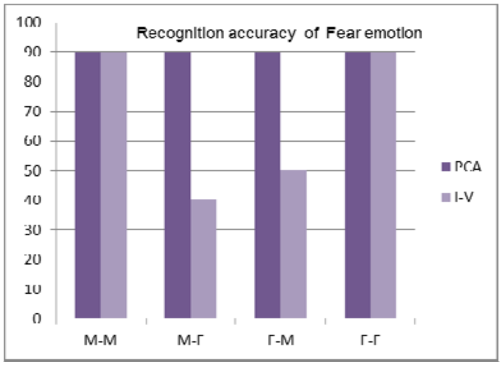

Fig 9: performance of PCA method and I-Vector method for Fear emotion for different test and train cases

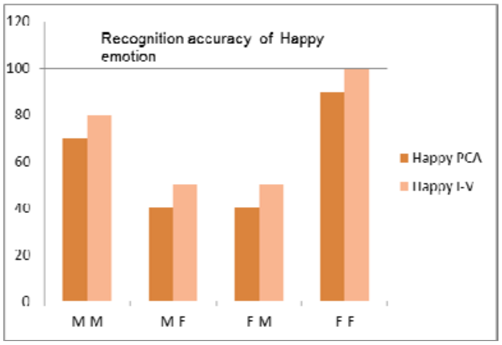

Fig 10: performance of PCA method and I-Vector method for anger emotion for different test and train cases

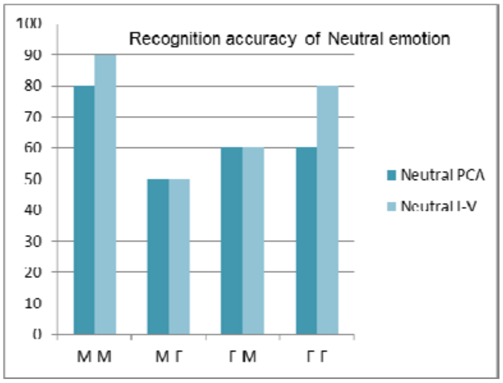

Fig 11: performance of PCA method and I-Vector method for anger emotion for different test and train cases

From the analysis given in the above figures the overall performance i-vector method has shown an improvement of 10% in most of the cases than the PCA method.

The recognition performance of the system is analyzed for the emotion vs the methodology.

6. CONCLUSION

1. The Emotion specific features of the emotions Anger, Happy, Fear and Neutral are from each emotion of the training samples of duration of 30 seconds and also for testing samples of 3 seconds each.

2. I-vectors are obtained from joint feature analysis of means.

3. There is a good improvement in the accuracies for the emotion recognition system for Gender

dependent and Gender independent experiments by using i-vectors at least with an improved accuracy of 10% over PCA method.

4. *From the experiments, it is observed that, the training database should contain both gender speeches to obtain better recognition accuracy.*

REFERENCES

[1] Athanaselis et al.2005, "ASR for Emotional Speech: clarifying the issues and enhancing performance", Neural Netw, 18:437-444.

[2] Iker Luengo et al., 2008, "Text independent speaker identification in multilingual environments", lrec2008, pp:1814-1817.

[3] Ch.Srinivasa Kumar et al. "Design Of An Automatic Speaker Recognition System Using MFCC, Vector Quantization And LBG Algorithm", Vol.3, IJCSE,2011, pp:2942-2954.

[4] Prerna Puri, "Detailed analysis of Speaker Recognition System and use of MFCCs for recognition",Vol.3,2013, IOSR journal of Engineering, pp:32-36.

[5] Mannepalli, K., Sastry, P.N. & Suman, M. Int J Speech Technol (2016) 19: 779. https://doi.org/10.1007/s10772-016-9368-y

[6] Laura Caponetti,Cosimo Alessandro Buscicchio and Giovanna Castellano "Biologically inspired emotion recognition from Speech" Journal on Advances in Signal Processing, 2011.

[7] Yazid Attabi and Pierre Dumouchel, "Anchor Models for Emotion Recognition from Speech" IEEE Transaction on Affective Computing, Vol. 4, No. 3, pp: 280-290, 2013.

[8] L. Zão, "Time-Frequency Feature and AMS-GMM Mask for Acoustic Emotion Classification", IEEE Signal Processing Letters, Vol. 21, No. 5, pp: 620-624,May, 2014.

[9] S. Selva Nidhyananthan, R. Shantha Selva Kumari, L. Bala Manikandan & P. Suresh, "Realiztion of emotions in speech using prosodic and articulation features" International Journal of Advanced Electrical and Electronics Engineering.Vol.2 Issue no.2, pp.83-86, 2013.

[10] MayankBhargava and Tim Polzehl "Improving Automatic Emotion Recognition from speech using Rhythm and Temporal feature" ICECIT published by Elsevier. Vol. 2 Issue no.2, pp.139-147, 2012.

[11] Yasmine Maximos and David Suendermann-Oeft,"Emotion recognition from children's speech" Speaker Emotion Challenge at Interspeech2013.

[12] Ankur Sapra, Nikhil Panwar and Sohan Panwar, "emotion recognition from speech",international journal of emerging technology and advanced engineering, volume 3, issue 2, pp. 341-345 February 2013.

[13] Vaishali M. Chavan and V.V. Gohokar "Speech Emotion Recognition by using SVM classifier" International Journal of Engineering and Advanced Technology (IJEAT), pp: 2249 – 8958,Volume-1, Issue-5, June 2012.

[14] Dellaert et al., " Recognizing Emotion in Speech", ICSLP, 1996.

[15] Lee CM, Narayanan, " Toward detecting emotion in spoken Dialogs", IEEE Trans Speech Audio Process, 13 (2): 293-303.

[16] Murray, Arnott, "Implementation and Testing of a system by producing emotion by a rule in synthetic speech", Speech Communication,16, 369-390.

[17] McGilloway et al. 2000, Rao et al. 2010, " Approaching automatic recognition of emotion from voice", ISCA workshop on speech and emotion.

[18] Gish, H., Krasner, M., Russell, W., and Wolf, J., "Methods and experiments for text-independent speaker recognition over telephone channels," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 11, pp. 865-868, Apr. 1986.

[19] Reynolds, D. A., and Rose, R. C., "Robust Text-Independent Speaker Identification using Gaussian Mixture Models'' IEEE-Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83, Jan-1995.

[20] "Emotion Recognition using Prosody Features", Anne KR Kuchibotla, 2015

[21] Diana Torres-Boza, Meshia Cedric Oveneke, et al. "Hierarchical Sparse Coding Framework for Speech Emotion Recognition" Speech Communication (2018),doi:10.1016/j.specom.018.01.006.

[22] Jainath Yadav, ,Md. Shah Fahad et al "Epoch detection from emotional speech signal using zero time windowing" Speech Communication 96 (2018) 142–149.

[23] Zhen-Tao Liu, et al. "Speech emotion recognition based on feature selection and

extreme learning machine decision tree", Neurocomputing (2017) ,doi: 0.1016/j.neucom.2017.07.050

[24] ShaolingJing, XiaMao and LijiangChen. "Prominence features: Effective emotional features for speech emotion recognition". Digital Signal Processing 2017, doi.org/10.1016/j.dsp.2017.10.06.

[25] M. Srikanth, D. Pravena, and D. Govind. "Tamil Speech Emotion Recognition Using Deep Belief Network (DBN)" Advances in Signal Processing and Intelligent Recognition Systems, Advances in Intelligent Systems and Computing 678,
2018. DOI 10.1007/978-3-319-67934-1 29 [26] W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion Recognition and AffectiveComputing on Vocal Social Media,"Inf. Manag., Feb. 2015.

[27] H. Cao, R. Verma, and A. Nenkova, "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[28] Peeters, G., "Chroma-based Estimation of Musical Key from Audio-Signal Analysis", In Proceedings of the 7th International Conference on Music Information Retrieval", Victoria (BC), Canada, 2006.

[29] Chin Kim On Paulraj M. Pandiyan Sazali Yaacob Azali Saudi, "Mel-Frequency Cepstral Coefficient Analysis in Speech Recognition", in proceedings of International Conference on Computing & Informatics, pp. 1 - 5, 2006.

[30] Mannepalli, K., Sastry, P.N. & Suman, M. Int J Speech Technol (2016) 19: 87. https://doi.org/10.1007/s10772-015-9328-y