ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645
Indian Journal of Science and Technology, *Vol 9(38), DOI: 10.17485/ijst/2016/v9i38/101938, October 2016*

Efficient Feature Extraction for Fear State Analysis from Human Voice

Palo Hemanta Kumar and Mihir N. Mohanty

Siksha 'O' Anusandhan University, Near PNB Bank Jagmohan Nagar, Khandagiri, Bhubaneswar - 751030, Odisha, India;
hemantapalo@soauniversity.ac.in, mihir.n.mohanty@gmail.com

Abstract

Background/Objectives: Analysis of human speech emotion has been continued since long. As the study and recognition helps the society in many respects, we intend to analyze the similar type of emotions. Methods/Statistical Analysis: 'Fear' and 'Nervousness' are being analyzed in comparison with normal voice. The correlation between these two emotions found to be very close. These voices belong to Oriya language. The popular features of speech, Mel-frequency cepstral coefficients (MFCCs) are used. As the fundamental frequency is unique from voice to voice, it is a suitable feature in case of similar voice signals. Findings: The combination of these two features outperformed the single feature based classification. In addition, the performance has been measured using log-likelihood ratio parameter. For recognition purpose, Gaussian mixture model (GMM) has been selected, and tested for these features. Novelty/Improvement: The individual MFCCs show 81.33%, whereas the combined features show 86.01% of accuracy. It is clearly evidenced in the result section.

Keywords: Correlation Coefficient, Feature Extraction, Fear State, Gaussian Mixture Model, Human Voice, Log-likelihood Ratio, Mel-frequncy Cepstral Coefficient

1. Introduction

Recent studies show, effective human emotion recognition systems can assist scientists investigate human state of mind from gestures, spoken language interaction, everyday activities, facial expressions, psychophysical information including personal possessions[1-5]. A short tree structure of human emotions including few similar, secondary and territory emotional attributes can be quite informative in this context[6]. There are emotional states often overlapped and create confusion for the recognition system[7]. Investigation on mostly exhibited primary speech emotions has been attempted during last few decades[7-15]. However, very little work on recognition of similar and overlapped speech emotional states has been addressed. Arguably, this has motivated the authors to explore one such issue. We have taken a step to distinguish two similar and often misleading emotional speech states such as fear and nervousness rarely investigated in literature. Few human states such as panic, Anxiety, tenseness, uneasiness, apprehension, distress, dread etc. can lead to fear. The authors claimed 70% accuracy with prosodic features using Gaussian mixture model (GMM) classifier for fear speech emotions. This kind of research can benefit security organization, public safety, medical therapists, and similar investigating agencies. For instance, online availability of acoustic changes and measurement can help these organizations to assess the level of fear through

*Author for correspondence

voices[16]. Nervousness can lead to certain fear or phobia. Nevertheless, it is nearly confused with the fear state; hence, is analyzed along with fear to validate the rating mechanism proposed in this work. These two states are compared with normal speech state by using the later as a reference model.

Very little database on fear and nervous speech states can be accessible. Few available database such as EMO-DB, SAFE (situation analysis in a fictional and emotional corpus), SAVEE (Surrey Audio-Visual Expressed Emotion) etc. are in German, English or such similar languages. No database in Indian regional Oriya language database for fear and nervousness speech states could be found. A novel collection of one such database is made for this experiment.

It is however, essential to select discriminant features that can demarcate closely related emotional states with a rating mechanism. F0 has been the most influential acoustic features for analyzing stressed speech. Several studies on actors, pilots, and tax induced pilot survey indicate a variation of pitch with different human states[17-19]. Hence, this feature has been considered as a primary feature in our experiment. Spectral features such as linear prediction coefficients (LPCs), linear prediction cepstral coefficients (LPCCs), Perceptual linear prediction coefficients (PLPs), Mel-frequency cepstral coefficients (MFCCs), and other variants have been quite effective for recognition of basic speech emotions[20]. LP models such as LPC and LPCC have simpler algorithm. However, sensitive to harmonic structure of the excitation source makes them unreliable. Another problem associated with LP analysis is the false formant identification during its tracking. In reported confusion between first and second formant and with first formant and pitch in classification of anger and happiness using LP analysis. Further these are all-pole models and approximate the spectrum of the signal equally well at the analyzing band. Alternatively, MFCC and PLP coefficients describe the speech signal both linearly and logarithmically to suit the human hearing and auditory mechanism. Hence, these have a certain edge over the conventional LPCs and LPCCs. PLPs differ from MFCCs in terms of the frequency conversion scale used. A major drawback of PLPs is that these features suffer from the limitation of LP analysis. For emotional speech recognition, MFCCs outperformed PLPs and more effective for recognition of non-telephonic speech[21]. As the utterances were collected in real world environment, MFCC has been chosen as another baseline feature in this work apart from F0.

A wide variety of classifiers as neural networks (NNs), Gaussian mixture model (GMM), Hidden Markov model (HMM), Support vector machine (SVM), random forest and Fuzzy logic etc. has been quite effectively used for human speech and emotion recognition[22-23]. Comparison among GMM, K-NN, and HMMs has gone in favor of GMM for emotional speech classification. GMMs are computationally complex similar to HMM but faster than the later. However, if time is not the constraint it can even outperform NNs for modeling speech emotions involving large feature sets. As frame-level features have been used in this piece of work, GMM has been opted as our compatible classifier.

2. Similarity Evaluation

2.1 Correlation Measure

Similarity between consecutive samples or observations can be measured using correlation analysis. It indicates the lag in time between adjacent samples and hence can indicate how closely the emotional states change as they progress with time. We have tested both the fear and nervous speech states using correlation. From the correlation analysis, it is found that both the emotions are close to each other as explained in the results.

2.2 Log-likelihood Measure

In this work, a statistical test has been carried out to determine and compare the goodness of fit of fear and nervous state of human speech using likelihood ratio parameter. The test will provide us optimized values of these features. In turn, these maximized values will indicate how far the features are under the influence of one emotional model than the other. The logarithms of these optimized values have been used to account for the Mel-scales of the feature sets. The critical optimized logarithmic values so obtained can help us to find the correlation between these states used here with respect to the neutral reference model.

As a parameterized family of pmf (probability mass functions) or pdf (probability density function), likelihood function can be a possible alternative for determining human emotional states hence used here. Defining a notional parameter β, we modeled these emotional states using a null and alternative hypothesis test that specifies these states completely. For convenience, we have taken fixed values of β for this hypothetical test, i.e. β

$$S_0 : \beta_0$$
$$S_1 : \beta_1 \quad (1)$$

Where, S_0 and S_1 are the null and alternative hypothesis respectively. The features analysis can be assumed to have certain specified distributions under either of these hypotheses. In other word, we need not to estimate any unknown parameters. Then, the likelihood ratio Λ , with likelihood function $L\left(\beta/p\right)$ and supremum function denoted by "sup" can be represented[24-25].

$$\Lambda(p) = \frac{L\left(\beta_0/p\right)}{sup\left\{L\left(\beta/p\right) : \beta \in \{\beta_0, \beta_1\}\right\}} \quad (2)$$

Where, p depends on the relation between the feature values. The numerator represents the extent of maximum likelihood that any designated outcome comes under S_0 , and the variation of the observed feature over the whole feature space is described by the denominator. The decision rule for the likelihood function is given as follows:

If $\Lambda > x, \beta_0$ is not discarded, else if $\Lambda < x, \beta_0$, then β_0 is rejected and else if $\Lambda = \mathbf{x}, \beta_0$ is rejected with probability 'r'.

Where, $0 \leq x \leq 1$. The value of x and *r,* are chosen based on some specified threshold or significance level, λ based on the relation given by

$$\mathbf{r} \cdot \mathbf{P}\left(\Lambda = \mathbf{x} | \beta_0\right) + P(\Lambda < x | \beta_0) = \lambda \quad (3)$$

So, a null hypothesis is rejected if the value of likelihood ratio statistic is very small. It is based on the value of λ , i.e. the extent of probability of type I error (rejection of a null hypothesis that is true) considered to be tolerable.

3. Feature Extraction Technique

Following features has been extracted for the classification of our chosen states of emotions. A block diagram representation of the technique is shown in Figure 1.

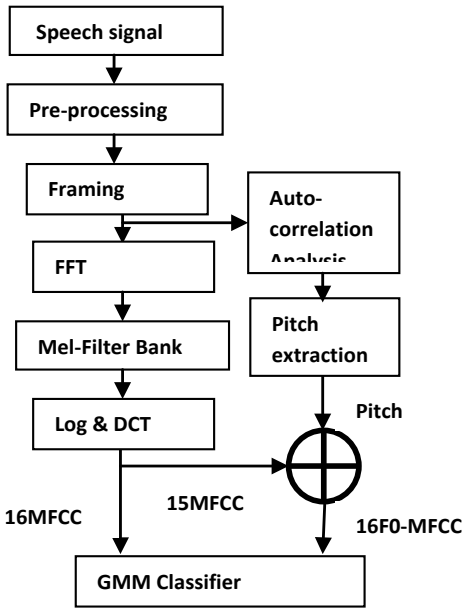

Figure 1. Feature extraction technique.

3.1 Fundamental Frequency

Fundamental frequency (F0) has been widely appreciated and acceptable feature for describing both explicit and implicit information of speech signal. It varies with speaker type, gender, age and among emotions. In this work, this has been extracted using auto correlation analysis due to following[26]: Firstly, the method is very reliable, simple, and more robust for F0 detection. Secondly, implementation technique of this method is straightforward. Thirdly, the coefficients can be computed directly from the waveform unlike other spectral based methods. Further, the computation is phase insensitive. Finally, hardware implementation using this scheme is less ambiguous, due to involvement of a single multiplier and an accumulator. Auto-correlation analysis can provide the periodicity information of any emotional speech signal. s*(n)*. For a time lag τ , the coefficients can be computed using the relation as

$$R(\tau) = \frac{1}{N}\sum_{n=0}^{N-1} s(n)s(n+\tau) \quad (4)$$

For condition given by $s(n) = s(n+\tau)$, $R(\tau)$ attains the largest value. Let, T is the period of s*(n)*, then $R(\tau)$ will have peaks at $\tau = IT$, where *I* is an integer. Generally, $R(\mathbf{0})$ provides the highest value among $R(\tau)$. Further with increase in τ , $R(\tau)$attains lower peaks. Hence, we can compute F0 at $\tau = T$ from the location of the peak.

3.2 MFCC

MFCC is often discussed, and mostly used spectral technique for speech emotion recognition. Approximation of both human auditory and hearing mechanism by using a Mel-scale makes this feature very effective. The scale compresses the higher frequency using a logarithmic scale 1 KHz and linear scale below this frequency. The generalized Mel-scalef_M is given as

$$f_M = 2595 * \log\left(1/700\right) \quad (5)$$

The block diagram for computation of MFCCs has been shown in Figure 1. It is observed that, most of the speaker or emotional information can be obtained from first few coefficients. Hence, out of *N* number of MFCCs computed from each frame, only 16MFCCs are retained for further processing in this experiment.

3.3 F0-MFCC

To represent paralinguistic information in speech signals both time and frequency domain approaches are quite effective. In time-domain approach, the desired attributes are extracted directly from the speech signal. Contrary to this, the desired parameters in frequency-domain approaches rely on spectrum analysis. To contain both the information a F0 pattern analysis has been performed with MFCC features in this work. The steps of combining F0 with MFCCs are explained:

- Initially 16MFCCs are extracted as per the block diagram in Figure 1.
- Out of this 15MFCCs feature vectors are retained.
- Sixteenth element of MFCC feature vector is replaced by the F0 values of each sentence. The new feature vector consists of 15MFCCs and one F0 vector of each utterance.

4. Method of Detection

The database of fear, nervous and normal utterances of Oriya language in Indian region has been collected from various sources. Around two hundred twenty five utterances of all the categories are collected. All the utterances belong to subjects in the age group of 18 to 45 years. Twenty-five utterances of each emotional state have been selected for this experiment based on a listening test by ten evaluators. The utterances are re-sampled at a rate of 16 KHz with 16bits. Database was digitized during conversion into .wav format using format factory software.

F0 and MFCCs have been used as basic features in this work. To eliminate the DC offset due to microphone and other recording background emotional speech signal is made cleaner by mean subtraction. In this average of the signal is subtracted from the original signal. In order to get rid of the silence part in the raw signal,

we applied the threshold. After usual normalization, the signal is pre-emphasized using a fast order pre-emphasis filter. Normalization was done to keep the recording and speaker variability to the minimum. The output of this filter provided more spectrally flattened signal. It also helped to make the signal less succumbed to any possible finite precision effects during further processing. The signal is framed into 25ms with 8ms overlapping between frames. Hamming window has been used for windowing the signal to remove edge effects and avoid loss of signal information.

Gaussian mixture model has been used to classify the chosen emotions in this work. It provides a smooth approximation to the feature distribution of each emotional class in terms of Gaussian mixture density (GMD)[27]. For *N* component densities, the GMDs can be computed as a weighted sum and is represented by

$$p(\vec{s}|\lambda) = \sum_{j=1}^{N} p_j b_j(\vec{s}) \quad (6)$$

Here, $\vec{s}$ is the feature vector dimension with independent component density given by $b_j(\vec{s})$ and $\lambda = \left\{p_j, \vec{\mu}, \sum_j \right\} j = 1, \ldots, N.$ an emotional model, These component densities representing features of an emotion, are parameterized by mean vectors (μ_j), state covariance matrices $\sum_j$ and a mixture weight satisfying the constraint $\sum_{j=1}^{N} p_j = 1$ as shown in Figure 2.

During training of the GMM classifier, features of twenty utterances in each emotional state have been used. These features for fear, nervousness, and neutral utterances were stored for generating appropriate GMM model. Three GMM models based on frame-level features are generated. They are used as reference models for the corresponding emotional states during testing.

During testing, frame-level, MFCCs and F0-MFCCs for 5 utterances of each emotional category were extracted similarly as explained before. A feature set is formed using feature vectors of first 90frames for each emotional state. Thus, 90 x 3= 270 feature vectors were tested with our three reference models. The decision on the classified emotional state is taken by computing the maximum log-likelihood measure of each tested feature vector with the reference GMM model. The classifier is tested using a closed-set emotional classification i.e. the tested samples formed a part of the training features. The procedure adapted in this experiment is shown in Figure 3.

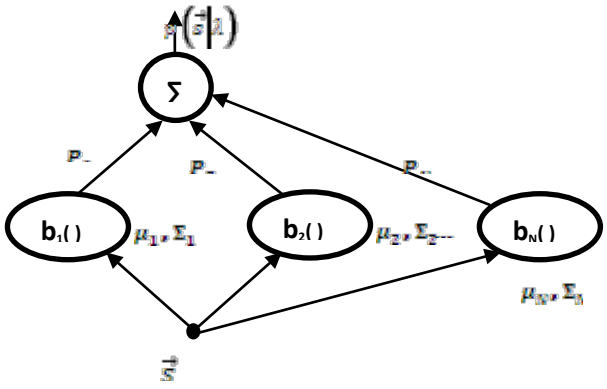

Figure 2. GMM model.

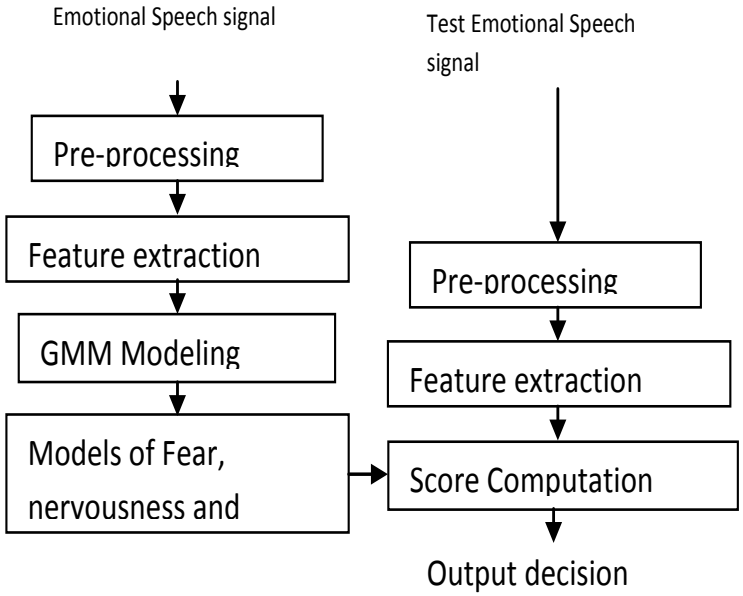

Figure 3. Steps of detection.

4. Results and Discussion

Variation of fundamental frequency for fear and nervous states against neutral state has been shown in Figure 4. It can be concluded that, nervousness is closely associated to fear as the trend suggest. However, fear possesses higher arousal level. The F0 values of mostly fear utterances lay between 300 to 500 and that of nervousness lies between 200 to 300. These two emotional states can be distinguished separately with these values. However, mostly values below 200 for neutral emotion can separate them from the neutral state.

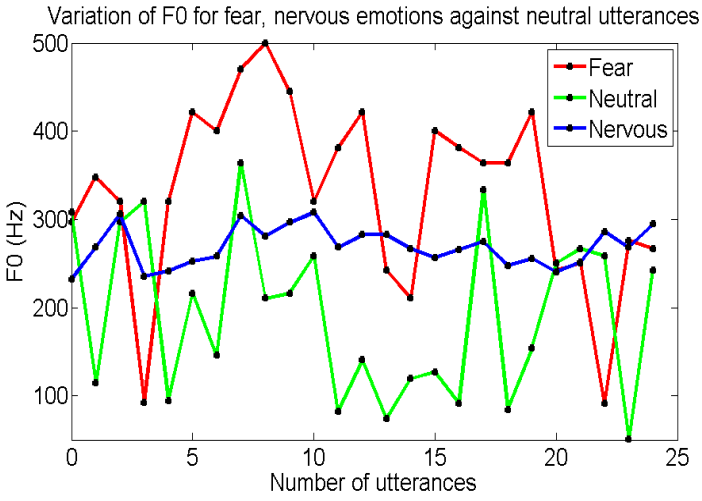

Figure 4. Variation of F0 for fear, nervous against neutral emotional states.

The likelihood model of fear and nervous against the neutral states has been graphically analyzed in Figure 5 and Figure 6. It is obvious that we cannot change the feature values as they are attributed to corresponding feature extraction technique describing the emotional states. Our aim is to generate a likelihood model for each of the emotional states by estimating possible optimized values for the features. The corresponding likelihood functional sets should represent the features more likely by maximizing their probability of distribution. If the tested result has low probability of occurrence under the null hypothesis as against the alternative, then the likelihood values are low. In this case, the null hypothesis cannot be discarded. Lower value thus indicates a better fit of the model. The likelihood values depends on the ratio given in equation (2). The value of the ratio lies between 0 and 1 as the numerator is less than the that of the denominator. However, the features analyzed here are MFCCs and F0-MFCCs that use a Mel-scale instead of the linear scale. Hence, we have computed the logarithms of likelihood functions for effective modeling. Further, it is simpler to work with this type of model. The log of the likelihood functions tends to be negative always. More closure of likelihood values towards zero or more negative the log-likelihood values are, better the fitting model.

Fear and nervous features are more negative than neutral features as shown in Figure 5. It indicates a better fitting model for these states than neutral. Further investigation of the figure reveals more similarity between fear and nervousness due to closure log-likelihood parameters as compared to neutral vs. fear.

Similar results have also been observed using this model with F0-MFCC features Figure 6. However, comparing Figure 5 and Figure 6 further details can be found. The model values using F0-MFCCs are more negative than that due to MFCCs. Therefore, F0-MFCCs have provided better goodness of fit model than MFCCs. Thus, use of F0 with MFCC enhanced the feature robustness. The log likelihood parameters are less negative (nearer to -26) for fear state, (nearer to -25) for nervous state and (nearer to -23) for neutral state using MFCCs. Comparable parameters for F0-MFCCs are between -36 to -35 and -34.1 to -34.9 for fear and nervous respectively and nearer to -33.6 to 33.9 for neutral states.

The comparison of log-likelihood parameters at the first iteration for chosen features is shown in Table 1, but

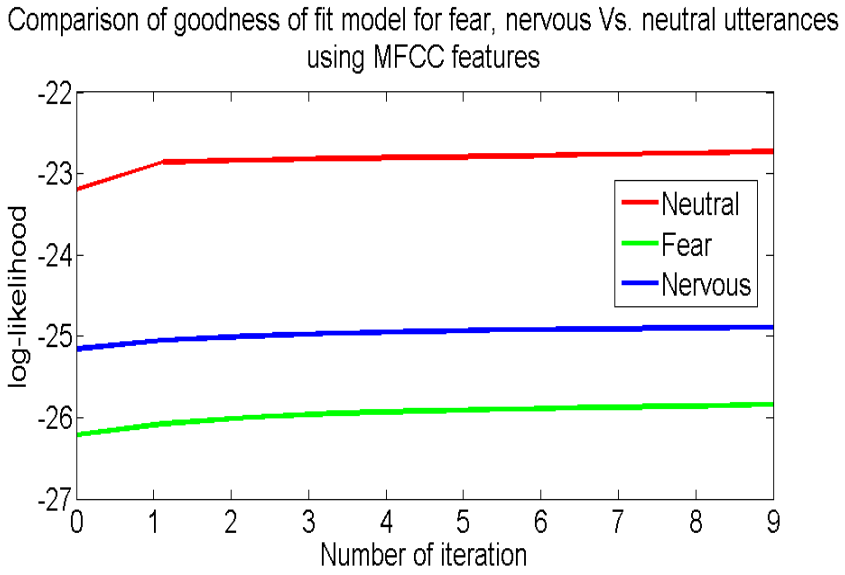

Figure 5. Comparison of goodness of fit model based on log likelihood ratio for fear, nervous and neutral speech emotional states with MFCC features.

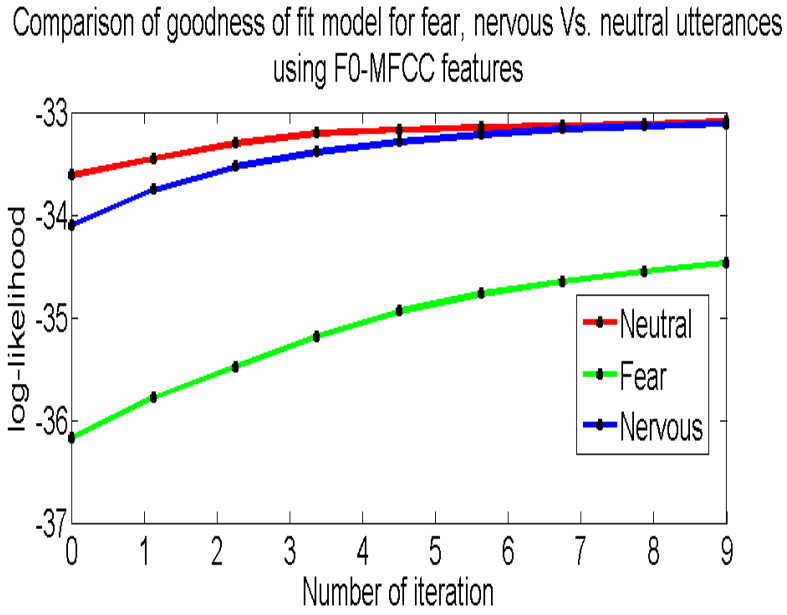

Figure 6. Comparison of goodness of fit model based on log likelihood ratio for fear, Nervous and neutral speech emotional states with F0-MFCC features.

Table 1. Comparison of log-likelihood ratio at first iteration

Emotions	MFCC	F0 based MFCC
Fear	-1291.506	-1745.751
Neutral	-1083.143	-1601.620
Nervous	-1125.688	-1622.923

not mentioned in Figure 5 and Figure 6 has clarified our claim.

On testing both emotions using correlation analysis, it is found that both fear and emotions are very close to each other. Most of correlation values lie in the range of 0.89 to 0.9 for fear and nervousness states. However, for normal voice these values are below 0.5. The range of correlation coefficient is 0.6529 for fear, (0.6412) for nervousness and 0.40 for neutral, indicates the variation of these arousal states against the neutral state of emotion.

The real time comparison of fear and nervous against the neutral utterances using the GMM classifiers is shown in Table 2. It also provides the classification accuracy for fear and nervousness emotional states against the neutral states. A Comparison of Fear and nervous vs. neutral utterances using GMM classifier is shown in Table 3. Size

Table 2. Real time comparison of Fear and nervous vs. neutral utterances using GMM classifier

Features	MFCC	F0-MFCC
Number of utterances/emotion	25	25
Feature extraction	27.66s	55.98s
Classification time	5.63s	5.81s
Total time	33.29s	61.79s
Recognition accuracy	81.33%	86.01%
Training data (%)	90%	90%
Testing data (%)	10%	10%

Table 3. Comparison of Fear and nervous vs. neutral utterances using GMM classifier

Features	**MFCC**	**F0-MFCC**
Feature dimension	N x16 MFCC	N x 16 (15MFCC+F0)
Frame size	25ms	25ms
Frame overlap	8ms	8ms
Training features	0.9 x N x16 MFCC	0.9N (15 MFCC + F0)
Testing features (T)	300 x 16	300 (15 + F0)

of feature vector involved during training and testing of the GMM classifier is also indicated in this table.

Both MFCC and proposed F0-MFCC features are compared. Use of F0 along with MFCC has resulted an enhanced accuracy than individual MFCCs. The recognition accuracy of 86.01% for F0-MFCC as compared to 81.33% for simple MFCC has proved the robustness of the former in classifing our chosen states of emotions.

6. Conclusion

In this work, an attempt is made to investigate two similar nature emotional states such as fear and nervousness. People become nervous under a tense condition that can lead to fear. Use of correlation coefficient and likelihood test statistics has been exploited to observe these states. We able to put a boundary between these two states.

However, closeness of the plots and graphs for fear and nervousness indicates their similarity. Fear has slightly more edge over nervousness in terms of arousal level. Further, use of fundamental frequency with MFCC features helped in enhancing the reliability of the combined feature for these emotional states. This has been validated with the classification accuracy improvement for F0-MFCC features compared to MFCC features using GMM classifier. Further, the robustness of the proposed feature is evidenced from the goodness of fit graph in modeling the chosen emotional states. Other statistical tests such as Z-test, chi-square test, F-test etc. can put further insight in this direction. Exploration of robust features to represent other similar nature emotion is an area need to be further addressed.

7. References

1. Wang JC, Chin YH, Chen B-W, Lin CH, Wu CH. Speech Emotion Verification Using Emotion Variance Modeling and Discriminant Scale-Frequency Maps. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2015 Oct; 23(10):1552–62.
2. McDuff D, Karlson A, Kapoor A, Roseway A, Czerwinski M. AffectAura: Emotional wellbeing reflection system. Proceedings 6th International conference Pervasive Computing Technolologies for Healthcare. 2012 May.
3. Prakash NS, Venkatram N. Establishing Efficient Security Scheme in Home IOT Devices through Biometric Finger Print Technique. Indian Journal of Science and Technology. 2016 May; 9(17):1-8.
4. Vaid S, Singh P, Kaur C. Classification of Human Emotions using Multiwavelet Transform based Features and Random Forest Technique. Indian Journal of Science and Technology. 2015 Oct; 8(28):1-7.
5. Rajkumar N, Ramalingam V. Cognitive Intelligent Tutoring System based on Affective State. Indian Journal of Science and Technology. 2015 Sep; 8(24):1–10.
6. Clavel C, Vasilescu I, Devillers L, Richard G, Ehrette T. Fear-type emotion recognition for future audio-based surveillance systems. Speech Communication. 2008 Jun; 50(6):487–503.
7. Palo HK, Mohanty MN. Classification of Emotions of Angry and Disgust. Smart Computing Review. 2015 Jun; 5(3):151-58.
8. Wang K, An N, Li BN, Zhang Y. Speech emotion recognition using Fourier parameters. IEEE Transaction on Affective Computing. 2015 Jan-Mar; 6(1):69-75.
9. Palo HK, Mohanty MN. Classification of Emotional Speech of Children Using Probabilistic Neural Network. International Journal of Electrical and Computer Engineering (IJECE). 2015 Apr; 5(2):311-17.
10. Ververidis D, Kotropoulos C. Emotional speech recognition: Resources, features and methods. Speech Communication. 2006 Sep; 48(9):1162-81.
11. Yildirim S, Bulut M, Lee CM, Kazemzadeh A, Busso C, Deng Z, Lee S, Narayanan S. An acoustic study of emotions expressed in speech. Proceedings of International conference on Spoken Language Processing (ICSLP '04). 2004 Jan; 1:2193-96.
12. Wu S, Tiago HF, Wai-Yip C. Automatic speech emotion recognition using modulation spectral features. Speech Communication. 2011 May-Jun; 53(5):768-85.
13. Lanjewar RB, Chaudhari DS. Comparative analysis of speech emotion recognition system using different classifiers on berlin emotional speech database. International Journal of Electrical and Electronics Engineering Research (IJEEER). 2013 Dec; 3(5):145-56.
14. Ayadi E, Kamal MS, Karray F. Survey on speech emotion recognition: features, classification schemes, and databases. Pattern Recognition. 2011 Sep; 44(3):572-87.
15. Palo HK, Mohanty MN, Chandra M. Efficient feature combination techniques for emotional speech classification. International Journal of Speech Technology. 2016 Mar; 19(1):135-50.
16. Detecting the emotion fear through voice. Date Accessed: 25/06/2008: Available from: http://mmi.tudelft.nl/~vrphobia/RA_IfaChaeron.pdf.
17. Devillers L, Vasilescu I, Vidrascu L. F0 and pause features analysis for Anger and Fear detection in real-life spoken dialogs. Speech Prosody 2004 Nara, Japan. 2004 Mar; p. 1-4.
18. Cairns DA, Hansen JHL. Nonlinear analysis and classification of speech under stressed conditions. Acoustical Society of America. 1994 Dec; 96(6):3392-400.
19. Mohanty MN, Jena B. Analysis of stressed human speech. International Journal of Computational Vision and Robotics. 2011 Sep; 2(2):180-87.
20. Kumar P, Chandra M. Pitch-based cepstral features for gender classification in noisy environments. International Journal Signal and Imaging Systems Engineering. 2013; 6(3):138-42.

21. Lu X, Dang J. Physiological feature extraction for text-independent speaker identification using non-uniform subband processing. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07. 2007 Apr; p. IV-461-IV-464.
22. Prabhu V, Gunasekaran G. Fuzzy Logic based NAM Speech Recognition for Tamil Syllables. Indian Journal of Science and Technology. 2016 Jan; 9(1):1-12.
23. Palo HK, Mohanty MN, Chandra M. New Features for Emotional Speech Recognition. IEEE Power, Communication and Information Technology Conference (PCITC). 2015 Oct; p. 424-29.
24. Casella G, Berger RL. Statistical Inference. Cengage Learning, Second edition. 2001 Jun.
25. Mood AM, Graybill FA. McGraw-Hill: Introduction to the Theory of Statistics. Second edition. 1963.
26. Rabiner LR. On the Use of Autocorrelation Analysis for Pitch Detection. IEEE transactions on acoustics, speech, and signal processing. 1977 Feb; 25(1):24-33.
27. Subhashree R, Rathna GN. Speech Emotion Recognition: Performance Analysis based on Fused Algorithms and GMM Modelling. Indian Journal of Science and Technology. 2016 Mar; 9(11):1-8.